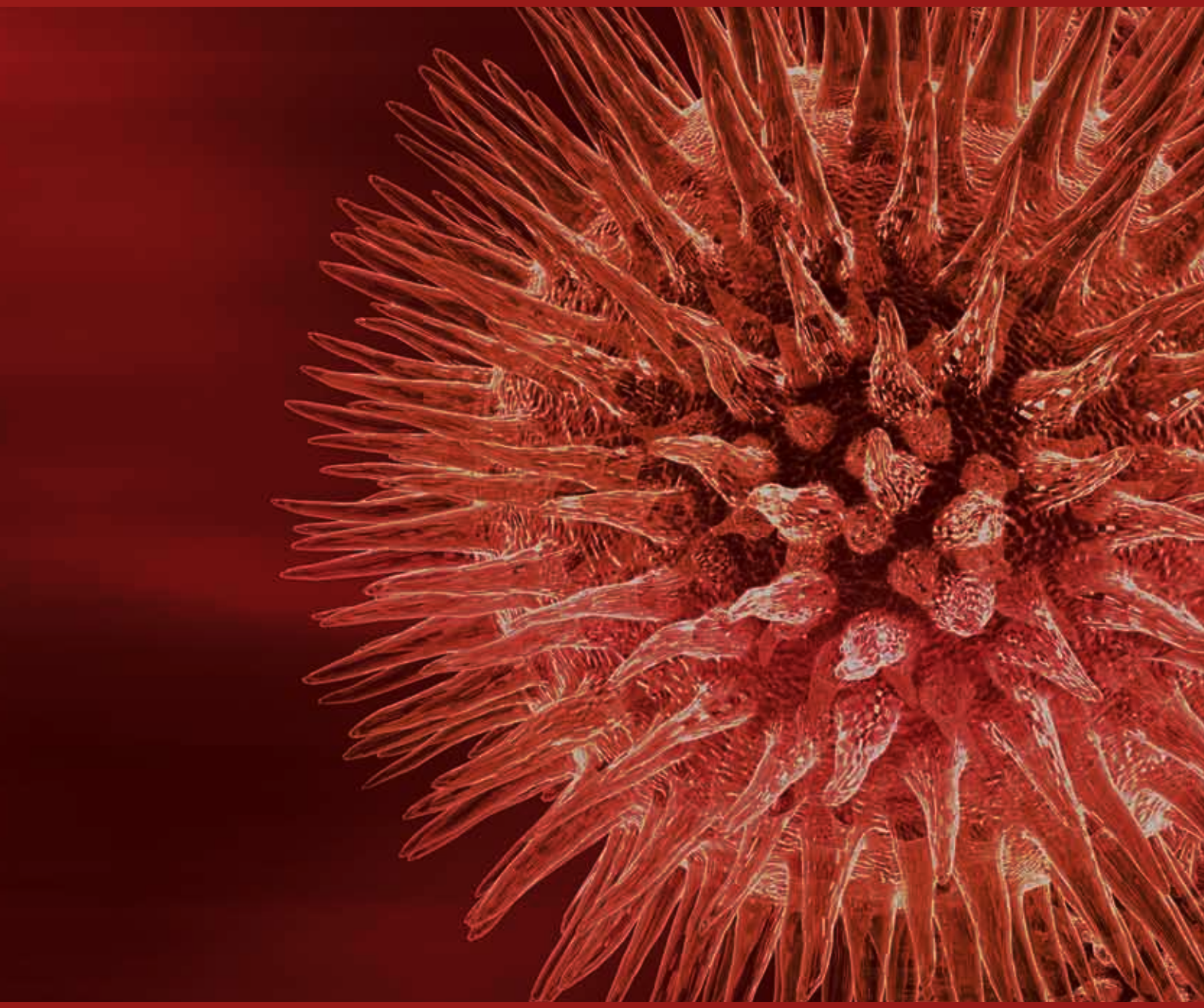
BioMed Research International

Application of Systems Biology and Bioinformatics Methods in Biochemistry and Biomedicine

Guest Editors: Yudong Cai, Tao Huang, Lei Chen, and Bin Niu



Application of Systems Biology and Bioinformatics Methods in Biochemistry and Biomedicine

BioMed Research International

Application of Systems Biology and Bioinformatics Methods in Biochemistry and Biomedicine

Guest Editors: Yudong Cai, Tao Huang, Lei Chen, and Bin Niu

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 651968, 2 pages
http://dx.doi.org/10.1155/2013/651968

Editorial

Application of Systems Biology and Bioinformatics Methods in Biochemistry and Biomedicine

Yudong Cai,[1] Tao Huang,[2] Lei Chen,[3] and Bin Niu[4]

[1] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[2] *Department of Genetics and Genomics Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA*
[3] *College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*
[4] *College of Life Science, Shanghai University, Shanghai 200444, China*

Correspondence should be addressed to Yudong Cai; cai_yud@126.com

Received 10 September 2013; Accepted 10 September 2013

With the explosively increasing high-throughput omics data, it is highly desired to develop effective computational methods and tools that can mine useful information to support the development of biochemistry, biomedicine, and drug design. Furthermore, in order to understand the protein-protein, protein-D/RNA, and other complex interactions, systems biology approaches are applied.

In this collection, diverse topics were covered and there are many novel methods and intriguing findings.

Y. Jiang et al. compared the gene expressions among the colorectal cancer patients in different stages and obtained the early and late stage biomarkers. Then, these two kinds of biomarkers were both mapped onto the protein interaction network, and the signal propagation path from the early stage biomarker to the late one was identified. Their findings may provide useful insights for revealing the mechanism of colorectal cancer progression at the cellular systems biology level.

L. N. Lili et al. investigated the process of stroma activation in human ovarian cancer by molecular analysis of matched sets of cancer and surrounding stroma tissues. They found that functionally significant variability exists among ovarian cancer patients in the ability of the microenvironment to modulate cancer development.

B. Yang et al. constructed a network-based inference framework for identifying cancer genes from gene expression data. Six identified genes (TSPYL5, CD55, CCNE2, DCK, BBC3, and MUC1) susceptible to breast cancer were verified through the literature mining, GO analysis, and pathway functional enrichment analysis.

Lung cancer is one of the most malignant cancers. B. Q. Li et al. identified 25 NSCLC and 38 SCLC genes with the shortest path approach in PPI networks. These candidate genes contained more cancer genes and more functional similarity with cancer genes than those identified from the gene expression profiles.

A. R. Iskandar et al. evaluated the perturbation of xenobiotic metabolism in response to cigarette smoke exposure in nasal and bronchial tissues. Their observation suggested that the effects of cigarette smoke exposure on the xenobiotic responses in the bronchial and nasal epithelium of smokers were similar to those observed in their respective organotypic models exposed to cigarette smoke, and nasal tissue could be a used as a reliable surrogate to measure the xenobiotic responses in the bronchial tissue.

E. G. Maiorov et al. identified interconnected markers for T-cell acute lymphoblastic leukemia (T-ALL). Their identified genes may serve as biomarkers, alternative to the traditional ones used for the diagnosis of T-ALL, and help understand the pathogenesis of the disease.

M. Kalita et al. used a multiplex gene expression profiling platform to investigate the perturbations of the innate pathways induced by TGF in a primary airway epithelial cell model of epithelial mesenchymal transition (EMT). Their results indicated that epigenetic changes produced by EMT induce dynamic state changes of the innate signaling pathway.

C. Lu et al. studied the functions of microRNAs related to the liver regeneration of the whitespotted bamboo shark, *Chiloscyllium plagiosum*. Their work deepened the understanding of mechanisms of liver regeneration and resulted in the addition of a significant number of novel miRNAs sequences to GenBank.

T. Alioto et al. presented a lightweight pipeline for first-pass gene prediction on newly sequenced genomes. The two main components are ASPic, a program that derives highly accurate, albeit not necessarily complete, EST-based transcript annotations from EST alignments. The other component is GeneID, a standard gene prediction program, which we have modified to take as evidence intron annotations. The pipeline was successfully tested on the entire C. elegans genome and the 44 ENCODE human pilot regions.

J. Zou et al. reviewed advanced systems biology methods in drug discovery and translational biomedicine. Their review provided a framework for addressing disease mechanism and approaching drug discovery.

L. Chen et al. proposed a computational method to predict the side effects of drugs, which integrated the information of chemical-chemical and protein-chemical interactions. Compared to most of the previous studies, the proposed method can provide the order information of the side effects for any query drug.

K. Wang et al. proposed an accurate method for protein-ligand binding site on protein surface using SVM and statistical depth function. The accuracy, sensitivity, and specificity on training set are 77.55%, 56.15%, and 87.96%, respectively, and on the independent test set the accuracy, sensitivity, and specificity are 80.36%, 53.53%, and 92.38%, respectively.

K. K. Tseng et al. presented a new system and novel approaches to classify different kinds of sperm images in order to assess their health. In their evaluation, the method reached accuracy of 87.5% and has better performance than the existing approaches to sperm classification.

A rapid method is required to mitigate complexity and computation challenges on high throughput protein identification. In Method for Rapid Protein Identification in a Large Database, an accelerated open method is presented by W. Zhang et al. to satisfy this requirement to some extent.

Q. Zou et al. proposed a novel method for distinguishing cytokine from other proteins. It is of vital importance of identifying cytokine in silicon. Ensemble classification strategy was employed for improving the prediction performance, and a friendly prediction web server was also developed.

Du and Yu introduced a novel method, SubMito-PSPCP, which embeds the PSSM into the pseudoamino acid compositions, to predict protein submitochondrial locations.

T. Gu et al. applied the Support Vector Regression and a two stage feature selection to developing the computational model which maps DPP-IV inhibitors to the activity. They also developed the online server.

Based on nonlinear mapping and Coulomb function, X. Liu et al. applied 3D kernel approach to predict the four protein tertiary structural classes and five membrane protein types with satisfactory results. It has not escaped our notice that kernel approaches may hold a high potential for predicting the other protein features.

T. H. Zhao et al. proposed a new method to predict protein disordered regions based on sequence features. The accuracy and MCC (Matthew's correlation coefficient) of their method are higher than three popular disordered region predictors: DISOPRED, DISOclust, and OnD-CRF.

M. S. M. Ali et al. studied the structure and function of LipA8 which is able to adapt to extreme temperatures. Simulations show that it is most stable at 0°C and 5°C. In extreme temperature, the catalytic domain (N-terminus) maintained its stability than the noncatalytic domain (C-terminus), but the noncatalytic domain showed higher flexibility than the catalytic domain.

A Boolean network (BN) is widely used as a model of gene regulatory networks. K. Kobayashi et al. proposed a BN model with two types of the control inputs and an optimal control method with duration of drug effectiveness. The optimal control problem is reduced to an integer programming problem.

J. Zhang et al. studied the microRNA-mediated regulation in biological systems with oscillatory behavior. They started with two specific microRNA-mediated regulatory circuits which show their fine-tuning roles in the modulation of periodic behavior and then applied these results to study the effects of miR369-3 regulation of cell cycle.

B. Yan et al. developed a mathematical model to study the mechanisms underlying the size checkpoint in fission yeast. They found that when the spatiotemporal regulation is coupled to the positive feedback loops, the mitosis-promoting factor (MPF) exhibits a bistable steady-state relationship with the cell size. The switch-like response from the positive feedback loops naturally generates the cell size checkpoint.

Detection of potential siRNA off-targets is crucial for High Content Screening (HCS) using small interfering RNAs (siRNAs). S. Das et al. performed a detailed off-target analysis of three most commonly used kinome siRNA libraries based on latest RefSeq version and created SeedSeq database, a new unique format to store off-target information.

L. Zhu et al. systematically investigated the characteristics and evolutionary pattern of actin gene family in primates. Phylogenetic analysis of 233 actin genes in human, chimpanzee, gorilla, orangutan, gibbon, rhesus monkey, and marmoset genomes showed that actin genes in the seven species could be divided into two major types of clades: orthologous group versus complex group. Codon usages and gene expression patterns of actin gene copies were highly consistent among the groups because of basic functions needed by the organisms but much diverged within species due to functional diversification.

J. Ping et al. performed long time-scale molecular dynamics simulations on both open and closed states of *Escherichia coli* adenylate kinase (ADK); based on which a conformational selection mechanism was proposed to explain the large scale domain motion of this enzyme.

Yudong Cai
Tao Huang
Lei Chen
Bin Niu

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 502827, 11 pages
http://dx.doi.org/10.1155/2013/502827

Research Article

ASPic-GeneID: A Lightweight Pipeline for Gene Prediction and Alternative Isoforms Detection

Tyler Alioto,[1] Ernesto Picardi,[2,3] Roderic Guigó,[4,5] and Graziano Pesole[2,3,6]

[1] *Centre Nacional d'Anàlisi Genòmica (CNAG), Parc Científic de Barcelona, 08028 Barcelona, Spain*
[2] *Dipartimento di Bioscienze, Biotecnologie e Biofarmaceutica, Università degli Studi di Bari, 70126 Bari, Italy*
[3] *Istituto di Biomembrane e Bioenergetica del Consiglio Nazionale delle Ricerche (CNR), 70126 Bari, Italy*
[4] *Centre de Regulació Genòmica (CRG), 08003 Barcelona, Spain*
[5] *Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain*
[6] *Centro di Eccellenza in Genomica Comparata, Università degli Studi di Bari, 70126 Bari, Italy*

Correspondence should be addressed to Graziano Pesole; graziano.pesole@uniba.it

Received 16 June 2013; Revised 1 August 2013; Accepted 4 August 2013

Academic Editor: Tao Huang

New genomes are being sequenced at an increasingly rapid rate, far outpacing the rate at which manual gene annotation can be performed. Automated genome annotation is thus necessitated by this growth in genome projects; however, full-fledged annotation systems are usually home-grown and customized to a particular genome. There is thus a renewed need for accurate *ab initio* gene prediction methods. However, it is apparent that fully *ab initio* methods fall short of the required level of sensitivity and specificity for a quality annotation. Evidence in the form of expressed sequences gives the single biggest improvement in accuracy when used to inform gene predictions. Here, we present a lightweight pipeline for first-pass gene prediction on newly sequenced genomes. The two main components are ASPic, a program that derives highly accurate, albeit not necessarily complete, EST-based transcript annotations from EST alignments, and GeneID, a standard gene prediction program, which we have modified to take as evidence intron annotations. The introns output by ASPic CDS predictions is given to GeneID to constrain the exon-chaining process and produce predictions consistent with the underlying EST alignments. The pipeline was successfully tested on the entire *C. elegans* genome and the 44 ENCODE human pilot regions.

1. Introduction

Despite great efforts over the last ten years in computational gene prediction, translating a genome to a set of exon-intron structures and the proteins they encode is still a challenging task. The falling costs of traditional DNA sequencing and the development of next-generation sequencing technologies is leading to an accelerated number of complete genome sequences [1]. The sheer number of genomes sequenced argues for a real and continued need to design accurate computational tools for gene finding, the basic requirement being a first-pass set of reliable protein coding gene models [2].

Once the genomic sequence of a given organism has been completed, a common approach for annotating genes encoding proteins involves using *ab initio* or *de novo* gene prediction programs [2, 3]. *Ab initio* gene finders in fact allow quick and cost-effective analyses—a genome-wide set of vertebrate genes can be determined in only a few hours, for instance [4]. Many such programs are based on hidden Markov models (HMMs) and need to be trained before their application [2–7]. Capturing all gene features of an organism in a reduced training set is not a feasible task and thus, the accuracy of *ab initio* gene finders is mainly limited to the quality and size of the training set. Nonetheless, it is almost always that gene predictions obtained using *ab initio* systems represent the starting point for annotating newly sequenced genomes [2, 3].

Given the limited nature and accuracy of *ab initio* gene finders [8], new computational tools have been developed which take into account external evidence [2, 3, 9]. Methods based on comparative genomics have proven to be more

accurate than previous systems even though they require that informant genomes be spaced at evolutionarily appropriate distances [10–12]. Newly sequenced genomes, however, do not always have an appropriately closely related genome available, reducing the global performances of such comparative methods. The recently sequenced grape genome [13], for instance, is not as strictly related to other available dicot plant genomes (such as *Arabidopsis thaliana* [14] or *Populus trichocarpa* [15] or *Lycopersicon esculentum* [16]) as those of human and mouse are to each other. However, during the last few years, methods using multiple genomes, which specifically take into account their evolutionary relationships, have been developed and are only beginning to show improvements over dual-genome prediction methods [11].

As emerged from the ENCODE Genome Annotation Assessment Project (EGASP) [8, 17], a community experiment to access the state of the art in genome annotation within the human ENCODE regions [18–20], programs relying on extrinsic evidence such as expressed sequence tags (ESTs) or mRNA sequences were found to be the most accurate in reproducing the manually curated annotations [8]. ESTs are in fact an invaluable source of evidence for the detection of exon-intron boundaries and likely alternative splicing variants [21]. Current methods for predicting genes using ESTs generally work by first performing an alignment of expressed sequence tags onto a target genomic region and then combining the alignment results with *ab initio* gene predictions [22]. However, the inclusion of EST alignments into HMM-based systems is not a simple task due to the requirement that alignments must be incorporated into the model in a probabilistic way, often leading to only negligible performance gains. A new version of HMMGene using EST evidence, for example, reported no improvement in predictions for *Drosophila melanogaster* [23]. Only methods combining EST alignments and comparative genomics such as TWINSCAN_EST or N_SCAN_EST and the recent Pairagon-N_SCAN_EST proved to be the most accurate in predicting exact exon-intron structures [8, 11, 24]. However, apart the from availability of one or more informant genomes, their approach to integrate information from EST alignments needs a training step. Also in this case, the quality and size of the training set may reduce the benefit due to ESTs, especially when they are used to predict genes in novel genomes with a limited amount of expression data. These limitations have been partially avoided by methods that use EST alignments to simulate the manual annotation. Exogean, for instance, is a program appropriately designed to employ EST or mRNA alignments as biological objects in a directed acyclic colored multigraphs (DACMs) [25]. Although Exogean has been indicated as one of the most accurate programs in predicting correct coding genes in EGASP project [8], it is subjected to strong limitations. A reduced number of available ESTs in fact may preclude gene prediction in genomic regions not covered by expression data.

In light of what has been previously discussed, we report here a simple and accurate method called ASPic-GeneID to improve gene prediction while maximizing the information gained from expressed sequence tags. Alignments of EST sequences to the genome are particularly good at pinpointing the location of splice sites and intronic sequence. Such introns can be easily used as evidence to improve the chaining of *ab initio* predicted exons, thus making gene models more accurate. Our procedure does not require complex probabilistic models and it is completely independent of EST training sets. Intronic sequences are directly inferred from expression data by means of the program ASPic [26, 27], whereas both the *ab initio* exon predictions and the gene assembly are performed using the GeneID software [28, 29]. Since ASPic is also able to detect the most likely transcript variants for a gene, we propose here two simple extensions to ASPic-GeneID that allow the prediction of alternative splicing transcripts.

We have tested ASPic-GeneID on the entire *C. elegans* genome (WS147) and the 44 human ENCODE regions and compared the results to those of programs representing the state of the art in nematode and human gene prediction. Our results suggest that ASPic-GeneID is a real and practical alternative to very complex pipelines that currently require all available evidence to obtain the same values of specificity and sensitivity.

In the next section, we explain in detail the methodology behind ASPic-GeneID and its implementation. Finally, we focus on ASPic-GeneID predictions on the human ENCODE regions and on the *C. elegans* genome.

2. Materials and Methods

2.1. The ASPic-GeneID Pipeline. ASPic-GeneID represents the integration of two complementary methods for predicting gene structures in a target genome: the *ab initio* gene predictor GeneID [28], and the alternative splicing prediction program ASPic [26, 27]. Given intronic locations deduced from alignment of expressed sequence tags (ESTs) to the genome, ASPic-GeneID attempts to predict complete gene structures in the target genome sequence.

The first component of the pipeline is ASPic, a method to predict alternative splicing isoforms expressed by a gene and their exon-intron organization at the genomic level through the information provided by available expression data, mostly EST sequences [26, 27]. In contrast with the majority of other tools for the analysis of alternative splicing, ASPic performs a multiple alignment of transcript data to the genomic sequence and refines exon-intron boundary alignments through dynamic programming [27]. Such techniques improve the quality of the splice site predictions by minimizing the number of false positives. ASPic also provides the minimal set of nonmergeable transcript variants compatible with the detected splicing events [27].

The other component of the pipeline is GeneID, a well-known *ab initio* gene finder that predicts and scores all potential coding exons along a query genomic sequence [28]. From the set of predicted exons, GeneID assembles the gene structure maximizing the sum of the scores of the assembled exons using a dynamic programming chaining algorithm [28]. The hierarchical structure of the program separates the problem of exon assembly from the prediction of coding exons along a given query genomic sequence. Simple rules

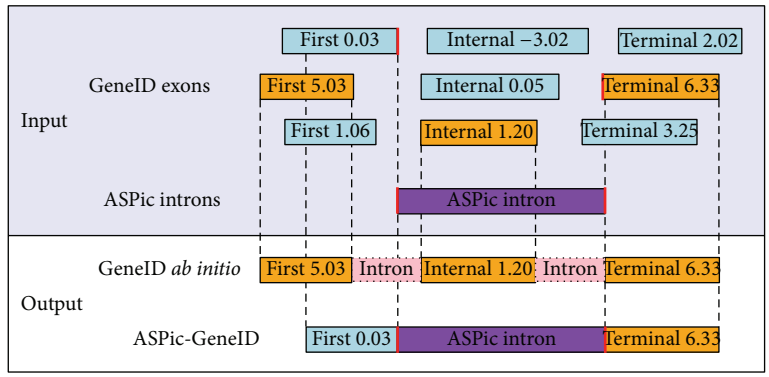


Figure 1: Graphical overview of ASPic-GeneID. In the absence of ASPic introns, the dynamic programming algorithm implemented in GeneID (called Genamic) assembles the most likely gene structure according to frame-compatible exons with the highest combined score. When ASPic introns are provided, they act as anchors in the chaining of exons so that exons with intron-compatible splice sites are always joined together if they conform to a valid gene model. In this hypothetical example, the "First" exon with the highest score (in orange) is replaced by the one with the lowest score (in blue), but which possesses an ASPic intron-compatible splice site (in red).

describing the relationships among initial, internal, and terminal exons as well as other gene signals (poly-adenylation, etc.) have to be imposed in an appropriate and organism-specific external parameter file in order to infer the most likely gene structures [28].

Current *ab initio* gene finders, including GeneID, suffer from both low specificity (they tend to predict too many genes and exons) and less than perfect sensitivity (correct exons may be predicted with low scores and consequently excluded from the final gene structures), leading to inaccurate predictions.

To improve both sensitivity and specificity of exon/transcript prediction, our novel procedure implemented in ASPic-GeneID is designed to improve the chaining of inferred GeneID exons by introducing constraints during gene assembly. We surmised that a good constraint candidate would be introns with reliable splice sites such as those predicted by ASPic. ASPic introns, in fact, are directly deduced from expression data and, thus, they constitute an invaluable source of evidence. We have introduced changes in GeneID so that the optimal path through the dynamic programming matrix is one which maximizes the number ASPic-inferred introns incorporated into the predicted transcript models. In other words, given a set of evidence introns, GeneID tries to join potential exons that have splice sites compatible with these introns (Figure 1).

2.2. Running ASPic-GeneID. Given a query genomic sequence (whether it is a single gene or a chromosome or a complete genome) and a collection of EST and/or mRNA sequences belonging to the same organism, we map all expressed sequences to the query sequence using GMAP [30], a computational tool specifically designed to reliably align a large number of ESTs and mRNAs to a genomic sequence. It has been shown that GMAP outperforms BLAT, which is another program widely used for the same purpose [30, 31]. The corresponding software has been downloaded from the website of the author (http://research-pub.gene.com/gmap/) and run with default parameters. Results of GMAP are then parsed to obtain clusters of ESTs and/or mRNAs related to specific regions of the query sequence. During the parsing only alignments with a minimum identity of 95% (98% in human) and EST coverage greater than 90% are retained. Each EST cluster should correspond to a specific gene. However, we may expect that different genes, depending on their peculiar expression profile are represented by clusters of different sizes or are not represented at all.

To construct clusters, we first collect overlapping ESTs and/or mRNAs according to GMAP coordinates on the genomic sequence and then we address compatible ESTs to the same cluster. Two ESTs or mRNAs are assumed to be compatible if they have at least one splice site in common, allowing a minimal mismatch around exon-intron boundaries in order to overcome potential GMAP misalignments or EST sequencing errors. In the case of unspliced ESTs, they are added to the relevant cluster according to mapping coordinates and strand. Each EST cluster and the mapping genomic sequence, form the input used by ASPic, run with default parameters, to predict introns. Depending on the coverage of the gene region by EST sequences, ASPic also provides a more or less reliable prediction of potential alternative transcripts.

After each ASPic run, we parse the corresponding output in order to collect all predicted introns in the general feature format [32]. The intron evidence is then given to GeneID which then predicts the most likely gene structures given this evidence and its statistical models for signals and coding sequence. The source code of GeneID has been updated in order to incorporate GeneID into the ASPic-GeneID context, in particular to accommodate the use of introns as evidence. A small adjustment has also been made to the parameter file in order to add introns to the gene model for the dynamic programming module implemented in GeneID.

We have written simple Python and Perl scripts to perform all the components of the ASPic-GeneID analysis transparently: the parsing of GMAP results, the generation of ESTs clusters, the ASPic intron predictions, and finally the GeneID predictions.

ASPic-GeneID has essentially no limits with respect to the length of the input sequence or the number of related ESTs. It can handle chromosomes as well as complete genomes.

2.3. Implementing the Alternative Splicing Prediction. The running of ASPic-GeneID as previously described allows the prediction of only one transcript per gene locus. Although this limitation should not reduce the gene prediction accuracy of our system in genomes with a low prevalence of alternative splicing, it is expected to affect the global performance in the case of genomes from organisms in which alternative splicing is a widespread phenomenon.

To overcome this limitation, we have implemented two extensions of ASPic-GeneID which allow for the prediction of alternative transcripts. In the first procedure, which we call ASPic-GeneID_AS1, we map ESTs to a query sequence using GMAP and then build EST clusters related to specific gene regions using exactly the same methodology as above. To each cluster and gene region we apply ASPic and collect in two separate files in GFF format all inferred introns and transcripts. Introns are used as evidence to run ASPic-GeneID as previously described and to obtain an initial gene set without alternative splicing. After that, all predicted ASPic-GeneID genes that overlap transcripts deduced by ASPic are removed. The remaining genes are then combined with transcripts inferred by ASPic to produce the final gene set. In this way, we employ ASPic alternative transcript predictions for all genomic regions fully covered by ESTs and/or mRNAs and ASPic-GeneID predictions for the remaining genomic regions partially covered or not covered by expression data.

In the second procedure, which we call ASPic-GeneID_AS2, we again map expression data to the query sequence using GMAP and run ASPic on each EST/mRNA cluster to collect deduced introns and full-length transcripts in GFF format. From each ASPic transcript we extract the longest open reading frame. Overlapping ASPic CDSs are then assigned to separate bins, where the number of bins used corresponds to the number transcripts belonging to the locus with the highest number of alternative transcripts. In order to maximally cover the genome, for loci with fewer transcripts than there are bins, ASPic CDS spans are reassigned to empty bins. We then run GeneID on each bin using both ASPic CDS and unassociated introns as evidence. Finally, we remove redundant identical transcripts from the combined predictions of each run of GeneID to produce a final gene set.

Relationships between ASPic and GeneID are shown in Figure 2. When GeneID uses only ASPic introns we have ASPic-GeneID predictions without alternative splicing. In contrast, when ASPic transcripts are used in combination with ASPic-GeneID we have ASPic-GeneID_AS1 and ASPic-GeneID_AS2 predictions with alternative splicing (Figure 2).

2.4. Sequence and Prediction Sets. The *C. elegans* genome sequence version WS147 was downloaded from the WormBase website (http://www.wormbase.org/). Gene predictions from several other *ab initio* gene finders such as Genefinder

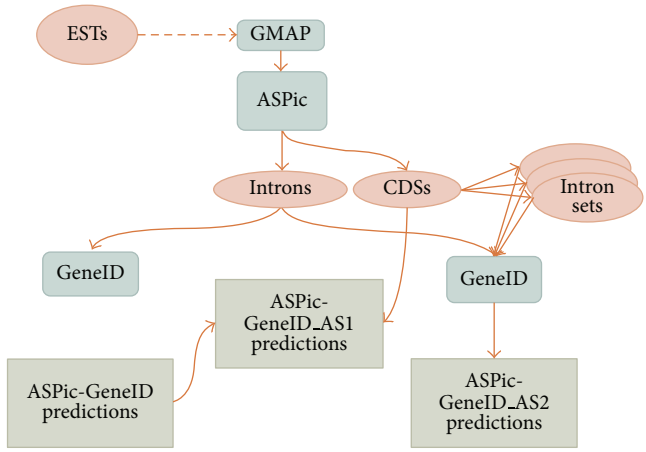


Figure 2: Relationship between ASPIC and GeneID predictions. ASPIC predicts both introns and full-length transcripts. When ASPIC introns are given to GeneID, we obtain ASPIC-GeneID predictions without alternative splicing. In contrast, when ASPIC transcripts are used, we can predict alternative variants in two ways. The first combining ASPIC transcripts and nonoverlapping ASPIC-GeneID predictions and the second giving ASPIC transcripts to GeneID as evidence and then removing redundant predictions.

(release 980504; P. Green, unpublished), FGENESH, and SNAP were downloaded from the Sanger Centre (http://www.sanger.ac.uk/Software/analysis/genomix/). TWINSCAN and TWINSCAN_EST predictions were downloaded from http://mblab.wustl.edu/. All available *C. elegans* ESTs and mRNAs have been retrieved from the Unigene database. We use Unigene sequences instead of dbEST sequences because Unigene sequences are filtered to avoid redundant and erroneous ESTs.

All 44 human ENCODE regions were downloaded from the UCSC genome browser (http://genome.ucsc.edu/) according to the human genome assembly hg17. Predictions from diverse gene finding programs belonging to different EGASP categories (*ab initio*, ESTs, mRNAs, and proteins based, all evidence based) were downloaded from the official EGASP repository (http://genome.imim.es/datasets/egasp2005/). The complete list of programs used in the evaluation is available in Table 2.

All human ESTs and mRNAs related to the 44 ENCODE regions were downloaded from GenBank according to their accession numbers retrieved from the Otter database [33].

2.5. Evaluation. Annotated *C. elegans* CDSs (WS147) were downloaded from WormBase. ASPic-GeneID predictions (including those from ASPic-GeneID_AS1 and ASPic-GeneID_AS2) as well as other predictions from different gene finding systems were evaluated against the annotation using an evaluation program written in Perl (Eduardo Eyras, personal communication), which takes into account alternative transcripts.

Briefly, the evaluation.pl program compares predictions and annotations in two ways: on a per gene basis and on a per best transcript pair (BTP) basis. For both methods, a gene is defined as a cluster of transcripts according to exon-overlap.

Table 1: Accuracy of gene finding programs on the complete *C. elegans* genome.

Evaluation at gene level																		
Program	SNg	SPg	SSg	WG	MG	SNe	SPe	SSe	WE	ME	SNi	SPi	SSi	WI	MI	SNn	SPn	SSn
GeneID	0.97	0.83	0.90	0.19	0.03	0.67	0.69	0.68	0.16	0.15	0.70	0.74	0.72	0.26	0.30	0.87	0.88	0.87
ASPic	0.34	0.96	0.65	0.03	0.66	0.29	**0.91**	0.60	0.02	0.69	0.30	**0.97**	0.63	0.03	0.70	0.31	**0.98**	0.64
ASPic-GeneID	**0.99**	0.93	0.96	0.09	0.01	0.85	0.81	0.83	0.08	0.03	0.93	0.88	0.90	0.12	0.07	0.95	0.95	**0.95**
ASPic-GeneID_AS1	**0.99**	0.75	0.87	0.21	0.01	**0.87**	0.78	0.82	0.12	0.02	0.93	0.86	0.89	0.14	0.07	**0.96**	0.91	0.93
ASPic-GeneID_AS2	0.98	**0.98**	**0.98**	0.07	0.02	0.86	0.83	**0.84**	0.07	0.03	**0.94**	0.89	**0.91**	0.11	0.06	**0.96**	0.93	0.94
TWINSCAN	0.95	0.87	0.91	0.12	0.05	0.76	0.77	0.76	0.11	0.11	0.81	0.83	0.82	0.17	0.19	0.90	0.91	0.90
TWINSCAN_EST	0.95	0.88	0.91	0.09	0.05	0.79	0.81	0.80	0.09	0.09	0.84	0.87	0.85	0.13	0.16	0.91	0.92	0.91
FGENESH	0.97	0.88	0.92	0.10	0.03	0.76	0.74	0.75	0.13	0.09	0.80	0.79	0.79	0.21	0.20	0.93	0.89	0.91
Genefinder	0.95	0.97	0.96	0.05	0.05	0.77	0.74	0.75	0.13	0.08	0.83	0.78	0.80	0.22	0.17	0.93	0.89	0.91
SNAP	0.96	0.69	0.82	0.22	0.04	0.70	0.66	0.68	0.18	0.12	0.74	0.73	0.73	0.27	0.26	0.90	0.86	0.88
Evaluation at transcript level																		
Program	SNt	SPt	SSt	WT	MT	SNet	SPet	SSet	WEt	MEt	SNit	SPit	SSit	WIt	TMIt	SNnt	SPnt	SSnt
GeneID	0.23	0.23	0.23	0.05	0.19	0.68	0.70	0.69	0.17	0.18	0.70	0.72	0.71	0.28	0.30	0.84	0.87	0.85
ASPic	0.26	**0.71**	**0.48**	0.01	0.29	0.81	**0.95**	**0.88**	0.01	0.15	0.82	**0.98**	**0.90**	0.02	0.18	0.81	**0.99**	0.90
ASPic-GeneID	0.44	0.47	0.45	0.03	0.17	0.86	0.81	0.83	0.10	0.05	0.92	0.85	0.88	0.15	0.08	0.93	0.93	**0.93**
ASPic-GeneID_AS1	**0.53**	0.44	**0.48**	0.08	0.11	0.87	0.85	0.86	0.08	0.06	0.91	0.88	0.89	0.12	0.09	0.92	0.94	**0.93**
ASPic-GeneID_AS2	0.46	0.50	**0.48**	0.03	0.17	**0.88**	0.81	0.84	0.11	0.04	**0.94**	0.85	0.89	0.15	0.06	**0.95**	0.90	0.92
TWINSCAN	0.35	0.36	0.35	0.04	0.15	0.77	0.78	0.77	0.11	0.12	0.81	0.82	0.81	0.18	0.19	0.88	0.91	0.89
TWINSCAN_EST	0.43	0.45	0.44	0.04	0.13	0.80	0.83	0.81	0.08	0.11	0.84	0.87	0.85	0.13	0.16	0.89	0.93	0.91
FGENESH	0.32	0.33	0.32	0.06	0.16	0.75	0.76	0.75	0.13	0.13	0.78	0.79	0.78	0.21	0.22	0.88	0.89	0.88
Genefinder	0.30	0.35	0.32	0.05	0.18	0.79	0.73	0.76	0.16	0.10	0.83	0.75	0.79	0.25	0.17	0.91	0.86	0.88
SNAP	0.27	0.22	0.24	0.09	0.11	0.67	0.73	0.70	0.11	0.19	0.70	0.77	0.73	0.23	0.30	0.82	0.92	0.87

The highest values are shown in bold. SN indicates sensitivity. SP indicates specificity. SS indicates the average between SN and SP. Gene (g), transcript (t), exon (e), intron (i), and nucleotide (n) were assessed.

For evaluation on the basis of a gene, the program performs a projection of all transcripts to the genome and then calculates for exons, introns, and nucleotides the sensitivity (SN), the specificity [13], the wrong cases (W), and the missing cases (M). All accuracy measures follow the definitions of Burset and Guigó [34]. Briefly, for each level (nucleotide, exon, and gene) the sensitivity is SN = TP/(TP + FN) and the specificity is SP = TN/(TN + FP), where TP are true positives, TN are true negatives, FN are false negatives, and FP are false positives [34].

Calculation of statistics on a BTP basis is performed as follows. For each transcript cluster, the evaluation program establishes a one-to-one (and one-to-many in the case of split/joined transcripts) mapping between predicted and annotated transcripts. It then produces similar measures as above but only for best transcript pairs. These measures give a better estimate of the accuracy of connectivity of the predicted transcripts.

Summary statistics for both methods given in the results are derived from total feature counts for the entire evaluation set.

The accuracy of ASPic-GeneID has been evaluated on two different data sets, the entire *C. elegans* genome (version WS147), and the 44 human ENCODE pilot regions, using WormBase and Gencode annotations, respectively, as the reference annotations. Such data sets have been appropriately chosen to better assess the performances of ASPic-GeneID in two organisms differentially subjected to alternative splicing. Moreover, *C. elegans* and human ENCODE regions differ in the amount of EST coverage, which in turn can affect the quality of ASPic predictions. In the ENCODE pilot regions, the EST coverage is nearly complete: 91.5% of all introns (98.6% of introns in coding sequence) are covered by ESTs at a specificity of 85%. In contrast, ESTs cover only about 60% of the *C. elegans* genome which often results in EST clusters with incomplete exon-intron structures.

In the case of the *C. elegans* genome, we have compared ASPic-GeneID predictions with those from programs representing the state of the art in nematode gene finding, including *ab initio*, comparative, and EST-based methods (Table 1). Likewise, all predictions on the 44 ENCODE regions have been compared to those from a number of established gene finding programs involved in the human ENCODE genome annotation assessment project (EGASP) (Table 2).

2.6. Availability. GeneID source code (version 1.3) as well as *C. elegans* and human parameter files can be downloaded from http://genome.crg.es/software/geneid/. For large-scale jobs, we recommend to download the off-line version of ASPic from the following web page: http://150.145.82.212/aspic/aspicgeneid.tar.gz. In addition, Python and Perl scripts to automate ASPic are also provided (including all ESTs that could get a very big file. On the other hand, ESTs sequences can easily be downloaded from Unigene database).

Table 2: Accuracy of gene finding programs on the ENCODE pilot regions.

Evaluation at gene level																		
Program	SNg	SPg	SSg	WG	MG	SNe	SPe	SSe	WE	ME	SNi	SPi	SSi	WI	MI	SNn	SPn	SSn
DOGFISH	0,75	0,84	0,80	0,26	0,25	0,61	0,72	0,67	0,18	0,28	0,61	0,72	0,67	0,28	0,39	0,68	0,80	0,74
Ensembl	0,94	0,81	0,88	0,18	0,06	0,81	0,74	0,78	0,15	0,08	0,87	0,82	0,85	0,18	0,13	0,91	0,86	**0,89**
Exogean	0,83	**0,95**	**0,89**	0,12	0,17	0,87	0,78	0,83	0,12	0,09	**0,90**	0,82	0,86	0,18	0,10	0,85	**0,88**	0,87
Exonhunter	0,94	0,26	0,60	0,72	0,06	0,69	0,41	0,55	0,51	0,13	0,72	0,49	0,61	0,51	0,28	0,90	0,58	0,74
FGENESH++	0,93	0,55	0,74	0,42	0,07	0,78	0,63	0,71	0,31	0,13	0,78	0,67	0,73	0,33	0,22	0,88	0,70	0,79
GeneID	0,92	0,76	0,84	0,32	0,08	0,56	0,59	0,58	0,27	0,28	0,53	0,59	0,56	0,41	0,47	0,77	0,74	0,76
Genemark	0,94	0,43	0,69	0,50	0,06	0,53	0,46	0,50	0,41	0,28	0,51	0,50	0,51	0,50	0,49	0,77	0,61	0,69
PAIRAGON-any	0,91	0,76	0,84	0,23	0,09	0,85	**0,83**	**0,84**	0,14	0,10	0,85	0,86	0,86	0,14	0,15	0,88	0,86	0,87
PAIRAGON-Multiple	0,88	0,80	0,84	0,24	0,12	0,75	0,77	0,76	0,18	0,16	0,75	0,78	0,77	0,22	0,25	0,85	0,83	0,84
SGP2	0,92	0,39	0,66	0,54	0,08	0,65	0,48	0,57	0,40	0,16	0,66	0,55	0,61	0,45	0,34	0,84	0,66	0,75
TWINSCAN	0,89	0,74	0,82	0,30	0,11	0,71	0,58	0,65	0,30	0,17	0,73	0,62	0,68	0,38	0,27	0,86	0,71	0,79
AUGUSTUS-abinitio	0,90	0,57	0,74	0,33	0,10	0,60	0,61	0,61	0,28	0,26	0,57	0,62	0,60	0,38	0,43	0,80	0,71	0,76
AUGUSTUS -any	0,95	0,66	0,81	0,26	0,05	0,81	0,73	0,77	0,21	0,08	0,79	0,74	0,77	0,26	0,21	**0,93**	0,78	0,86
AUGUSTUS-dual	0,93	0,60	0,77	0,29	0,07	0,70	0,65	0,68	0,25	0,15	0,68	0,67	0,68	0,33	0,32	0,89	0,75	0,82
AUGUSTUS_EST	0,95	0,70	0,83	0,24	0,05	0,80	0,74	0,77	0,20	0,09	0,78	0,75	0,77	0,25	0,22	0,91	0,79	0,85
ASPic	0,84	0,66	0,75	0,29	0,16	0,87	0,75	0,81	0,11	0,09	0,89	**0,87**	**0,88**	0,13	0,11	0,81	0,85	0,83
ASPic-GeneID_AS1	0,89	0,54	0,72	0,42	0,11	0,88	0,71	0,80	0,16	0,08	0,89	0,85	0,87	0,15	0,11	0,85	0,72	0,79
ASPic-GeneID_AS2	**0,96**	0,46	0,71	0,50	0,04	**0,90**	0,62	0,76	0,21	0,06	**0,90**	0,76	0,83	0,24	0,10	**0,93**	0,60	0,77
ASPic-GeneID	**0,95**	0,56	0,76	0,46	0,05	0,81	0,65	0,73	0,28	0,05	0,86	0,72	0,79	0,28	0,14	0,89	0,64	0,77
Evaluation at transcript level																		
Program	SNt	SPt	SSt	WT	MT	SNet	SPet	SSet	WEt	MEt	SNit	SPit	SSit	WIt	TMIt	SNnt	SPnt	SSnt
DOGFISH	0,06	0,13	0,10	0,01	0,44	0,73	0,76	0,75	0,15	0,19	0,72	0,75	0,74	0,25	0,28	0,78	0,82	0,80
Ensembl	0,25	0,24	0,25	0,13	0,23	0,84	0,87	0,86	0,04	0,08	0,90	0,93	**0,92**	0,07	0,10	**0,93**	0,95	**0,94**
Exogean	0,51	0,43	0,47	0,26	0,10	**0,89**	0,89	0,89	0,08	0,07	0,90	0,90	0,90	0,10	0,10	0,91	0,91	0,91
Exonhunter	0,06	0,03	0,05	0,03	0,46	0,69	0,67	0,68	0,21	0,19	0,71	0,70	0,71	0,30	0,29	0,85	0,80	0,83
FGENESH++	0,43	0,38	0,41	0,07	0,30	0,80	0,85	0,83	0,08	0,14	0,80	0,86	0,83	0,14	0,20	0,87	0,92	0,90
GeneID	0,03	0,04	0,04	0,01	0,50	0,60	0,62	0,61	0,24	0,27	0,57	0,60	0,59	0,40	0,43	0,78	0,78	0,78
Genemark	0,05	0,04	0,05	0,05	0,41	0,46	0,61	0,54	0,22	0,41	0,45	0,62	0,54	0,38	0,55	0,65	0,79	0,72
PAIRAGON-any	0,51	**0,46**	0,49	0,19	0,21	**0,89**	0,92	**0,91**	0,05	0,08	0,89	0,93	0,91	0,07	0,11	0,90	0,94	0,92
PAIRAGON-Multiple	0,21	0,35	0,28	0,01	0,43	0,87	0,83	0,85	0,12	0,08	0,87	0,83	0,85	0,17	0,13	0,91	0,88	0,90
SGP2	0,05	0,04	0,05	0,06	0,41	0,63	0,66	0,65	0,19	0,23	0,65	0,69	0,67	0,31	0,35	0,76	0,85	0,81
TWINSCAN	0,10	0,08	0,09	0,23	0,22	0,74	0,68	0,71	0,22	0,16	0,76	0,69	0,73	0,31	0,24	0,84	0,78	0,81
AUGUSTUS-abinitio	0,13	0,16	0,15	0,04	0,38	0,59	0,73	0,66	0,15	0,31	0,56	0,71	0,64	0,29	0,44	0,74	0,86	0,80
AUGUSTUS -any	0,27	0,34	0,31	0,04	0,41	0,80	0,86	0,83	0,07	0,14	0,79	0,86	0,83	0,14	0,21	0,88	0,93	0,91
AUGUSTUS-dual	0,15	0,17	0,16	0,05	0,39	0,65	0,77	0,71	0,11	0,25	0,64	0,77	0,71	0,23	0,36	0,79	0,90	0,85
AUGUSTUS_EST	0,27	0,36	0,32	0,03	0,42	0,80	0,86	0,83	0,07	0,14	0,79	0,86	0,83	0,14	0,21	0,88	0,94	0,91
ASPic	0,63	0,37	**0,50**	0,37	0,06	0,84	**0,95**	0,90	0,02	0,13	0,85	**0,96**	0,91	0,04	0,15	0,86	**0,97**	0,92
ASPic-GeneID_AS1	**0,65**	0,33	0,49	0,33	0,06	0,84	0,94	0,89	0,02	0,13	0,85	**0,96**	0,91	0,04	0,15	0,86	**0,97**	0,92
ASPic-GeneID_AS2	0,64	0,25	0,45	0,41	0,05	0,84	0,93	0,89	0,03	0,12	0,85	0,94	0,90	0,06	0,15	0,88	0,96	0,92
ASPic-GeneID	0,21	0,22	0,22	0,02	0,48	0,86	0,77	0,82	0,15	0,06	**0,91**	0,81	0,86	0,19	0,09	0,91	0,85	0,88

The highest values are shown in bold. SN indicates sensitivity. SP indicates specificity. SS indicates the average between SN and SP. Gene (g), transcript (t), exon (e), intron (i), and nucleotide (n) were assessed.

Additional Python and Perl scripts to automate ASPic-GeneID (for Linux and Mac OS X) are available upon request.

3. Results and Discussion

3.1. ASPic Intron and Gene Prediction in C. elegans Genome and All 44 ENCODE Regions. The underlying principle of our system is that introns can guide *ab initio* gene assembly. This task, however, can only be addressed using reliably predicted introns. Available methods to align EST sequences to the genome are mainly based on BLAST [35] or BLAT [31] and sometimes lead to poor splice site predictions. In order to obtain a high-quality set of intron positions, we first mapped all available *C. elegans* ESTs onto the complete worm genome (version WS147) using GMAP as described in Section 2. Then, EST clusters related to potential gene regions were

exploited to run ASPic. In contrast to other EST to genome alignment programs, ASPic employs a novel and efficient algorithm to minimize the number of exon predictions and hence of alignment inferred splice sites. ASPic is also able to infer alternative splicing variants of a gene given a related collection of ESTs.

When applied to each *C. elegans* EST cluster, ASPic can predict intron sequences and also full-length splicing variants whenever ESTs completely cover specific gene regions.

In all six *C. elegans* chromosomes, ASPic proves to be extremely specific. Out of 100723 predicted introns, 96.1% exactly match (with the same exact splice sites) annotated introns. However, overall sensitivity is low—likely due to the fact that coverage of the genome by ESTs is only about 60%.

Overall, ASPic's nucleotide and exon specificities are 98% and 91%, respectively. ASPic is also very specific when comparing only to the best overlapped transcripts where nucleotide and exon specificities increase to 99% and 95%, respectively.

Moreover, it is able to find exact transcript variants with a specificity of 71%, which is the highest reported up to now. Similar specificity values have been reported by Genomix, a new gene finder system working as a combiner [36]. However, Genomix specificities at exon and nucleotide levels of 87.3% and 91.9%, respectively, have been calculated only on a reduced subset of 1534 confirmed *C. elegans* genes (version WS147) [36].

Despite the high specificity, ASPic shows very low sensitivity values at all levels except when the comparison with the annotation is limited to only transcripts overlapped by a prediction (BTP level). In this case, the accuracy of ASPic, as given by (Sn + Sp)/2, at the nucleotide and exon levels are 90% and 88%, respectively.

It has been currently demonstrated that the expected distribution of spliceosomal intron lengths is correlated to the quality of the annotation [37]. Since introns are removed after transcription, intron lengths are not expected to respect coding frame. For this reason, the number of genomic introns that are multiple of three bases should be similar to the number of introns that are a multiple of three plus one or two bases [37]. In effect, ASPic predicted introns follow this behaviour. Of all *elegans* inferred introns, 33.5% are a multiple of three bases, whereas 33.6% and 33.0% are multiples of three plus one and two bases, respectively. These results strongly corroborate ASPic's ability to predict *bona fide* exon-intron boundaries.

The same approach used for the complete *C. elegans* genome has been applied to all 44 ENCODE human regions. In this case, however, single EST clusters related to gene regions have been generated using a subset of all available human expressed sequence tags downloaded from the Otter database in order to reduce potential pitfalls due to low quality ESTs or to aberrant mRNAs from pathological tissues.

ASPic is able to predict more introns than annotated in ENCODE. However, we focus only on annotated coding regions and it is well known that ENCODE contains many noncoding transcripts in addition to a number of introns located in UTR regions. Restricting, thus, the comparison to coding regions only, we found ASPic to be the most accurate system to predict introns in human ENCODE. This is derived mostly from its higher specificity; it is the most specific, with 87% of all predicted introns corresponding exactly to an annotated intron. This value increases to 96% when making the BTP comparison, demonstrating that the novel alignment algorithm behind ASPic is quite efficient and results can be comparable to those based on PAIRAGON, indicated as one of the best program to align mRNA sequences to genome [24, 38]. As shown in Table 2, ASPic outperforms PAIRAGON-any in predicting correct introns. Considering that PAIRAGON-any aligns only high quality sequences from the full ORF Mammalian Gene Collection (MGC) [39] and from the human RefSeq database, ASPic's performance which is based only on ESTs is even more remarkable.

ASPic is not highly specific at the transcript level where it is outperformed by Exogean and PAIRAGON-any. However, it is as specific as the combiner Fgenesh++ [40] and it is 12% more sensitive at the transcript level than Exogean and PAIRAGON-any. ASPic is also more sensitive than Ensembl [41], AUGUSTUS-any, and AUGUSTUS-EST at the exon level [22]. When comparing at the BPT level, it has the highest exon specificity (95%) (Table 2).

Like for the previous results described for the whole *C. elegans* genome, in the human ENCODE regions, ASPic predicted intron length distributions are not skewed. Of all ASPic introns, 33.0% are a multiple of three bases and 33.2% and 33.7% are multiples of three plus one and two bases, respectively.

3.2. ASPic-GeneID Accuracy without Alternative Splicing. Depending on the EST coverage of each gene region, ASPic can predict just introns or both introns and alternative splicing variants. For this reason, we can independently use two main sources of evidence from ESTs such as individual introns and full-length transcripts to improve GeneID *ab initio* predictions. When only introns are given as evidence to GeneID, the program is able to predict at most one transcript per locus. As outlined in Section 2, introns with correct splice sites can aid the correct assembly of *ab initio* predicted exons during the exon-chaining step. The dynamic programming procedure implemented in GeneID builds gene structures using exons with the highest scores respecting frame compatibility and gene model rules [28]. The introduction of ASPic introns to GeneID forces exons with compatible frames and splice sites to be joined. Since such evidence introns do not interfere with the main GeneID exon prediction process, it is expected that they are used only when compatible *ab initio* exons really exist. Our procedure to handle evidence introns as implemented in ASPic-GeneID is also expected to improve the accuracy at the transcript level.

When all ASPic predicted introns on the complete *C. elegans* genome are given as evidence to GeneID, our combined ASPic-GeneID system is found to be the most accurate in predicting exact nematode transcripts. The results show 21% improvement in sensitivity and 24% in specificity in predicting exact transcript structures compared to GeneID, which does not use ASPic introns (Table 1). ASPic-GeneID is, in turn, significantly more accurate than SNAP [4], FGENESH [42], and GENEFINDER [8], the most widely used

ab initio gene prediction program for nematodes. Moreover, ASPic-GeneID outperforms TWINSCAN [43], which uses the *C. briggsae* genome as an informant genome, at all sensitivity and specificity measures. Most interestingly, our gene prediction method is also more accurate than TWINSCAN_EST [9], a new system that combines EST alignments with TWINSCAN. In particular, ASPic-GeneID is 6% more sensitive at the exon level than TWINSCAN_EST. Taking the mean between sensitivity and specificity, ASPic-GeneID is also more accurate than TWINSCAN_EST in predicting exact transcript structures, 45.5% versus 44%.

The strength of ASPic-GeneID relies on the use of reliable intron sequences. Even when EST genome coverage is not high and, thus, the number of ASPic predicted introns is low, our system should predict genes with an accuracy better than GeneID alone. To verify the effect of the number of introns in improving *ab initio* GeneID predictions, we performed the following experiment. ASPic was run on the complete *C. elegans* genome using all available ESTs as described in Section 2. Then, from all the predicted introns we randomly selected increasing percentages of introns ranging from 0% to 100% and ran GeneID using each intron subset. The number of introns is undoubtedly related to the EST genome coverage and, thus, a low number of ESTs should yield a low number of introns. Results of this experiment are given in Figure 3 where the averages between sensitivity and specificity at gene [44], exon (SSe), and nucleotide levels (SSn) are reported as a function of growing intron percentages. The benefit due to introns increases linearly with the number of input introns and we can register a gene prediction improvement at all levels, even when the number of introns is very low (10%). These data indicate that ESTs and, thus, introns related to some genes can improve the accuracy of neighbouring genes. In practice, GeneID mistakes such as extension and inclusion of exons in neighbour genes become much less common because introns introduce real constraints in gene assembly.

The accuracy of ASPic-GeneID using only introns has also been evaluated on all 44 human ENCODE regions. Here, however, the situation is quite different because human genes are subjected to extensive alternative splicing and because human gene density is low. A system such as ASPic-GeneID which predicts only one transcript per locus is a disadvantage. Nonetheless, ASPic-GeneID outperforms all *ab initio* gene prediction programs such as Genemark [45] or AUGUSTUS_abinitio [5], currently one of the most accurate programs to find *ab initio* gene structures in mammals [8]. ASPic-GeneID is 18% more sensitive and specific than GeneID alone in predicting exact transcript structures. Moreover, ASPic-GeneID accuracy at the exon level is 73%, a value which is higher than the corresponding value obtained from other systems that use ESTs such as ExonHunter [46] or informant genomes such as TWINSCAN [43], SGP2 (an extension of GeneID) [10], DOGFISH [8], and AUGUSTUS_dual [22] or both evidence sources such as FEGENESH++ [40]. However, our system is less accurate at exon and nucleotide level than programs that use all available evidence for human (Ensembl, PAIRAGON-any, and AUGUSTUS-any) or programs that predict more than one transcript per locus (Exogean [25]). Nonetheless, in several measures, ASPic-GeneID shows high sensitivity. At the exon level, for instance, ASPic-GeneID sensitivity is 81%, 1% more than AUGUSTUS_EST, an improved version of AUGUSTUS that uses ESTs and proteins alignments, 6% more than PAIRAGON_multiple and 3% more than FGENESH++, a combiner that uses all available evidence. In the BTP comparison, ASPic-GeneID sensitivity at the exon level increases to 86%, 2% higher than Ensembl.

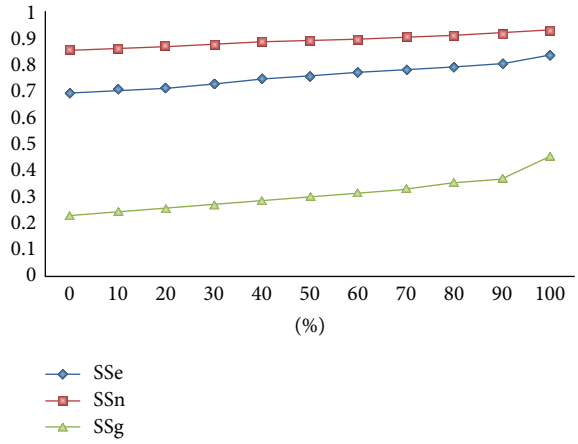


Figure 3: ASPic-GeneID performance on *C. elegans* according to number of introns. The accuracy [(Sn + Sp)/2] of ASPic-GeneID predictions (AG) is plotted according to the proportion of introns output by ASPic and provided to GeneID as evidence. SSe, SSn, and SSg indicate the accuracy at exon, nucleotide, and gene level, respectively.

On the whole, ASPic-GeneID remains one of most accurate systems to predict correct intronic sequences, attesting its sensitivity at 86% and specificity at 72%. These last values go up to 91% and 81%, respectively, when intron evaluation is assessed at the BTP level.

We noted, however, that ASPic CDS predictions alone are better than those of the combined ASPic-GeneID on the ENCODE regions (see Figure 4). We surmised that this must be due to EST coverage. Unlike in *C. elegans*, where EST coverage is somewhat low, the coverage of annotated human coding sequences by human ESTs is very high (85% of all introns and nearly 99% of introns in coding sequences). To determine at what level of EST coverage using our combined approach may be beneficial, we performed the following experiment. We selected random sets of ESTs corresponding to 10%, 20%, 30%, and so forth up to 100% of the ESTs available as input to ASPic. These EST sets had an intron coverage ranging from 27% to 85% of annotated introns. When using less than 35% of the available ESTs (corresponding to about 62% intron coverage) ASPic-GeneID performed better at the exon level than ASPic alone. At higher coverage, we found that ASPic CDS predictions are clearly more accurate. The performance of ASPic at the transcript level is quite good even at low EST coverage levels. This is perhaps due to the presence of a class of highly expressed transcripts that are well covered by ESTs. ASPic will predict them correctly, while ASPic-GeneID may try to extend the transcripts with additional predicted exons.

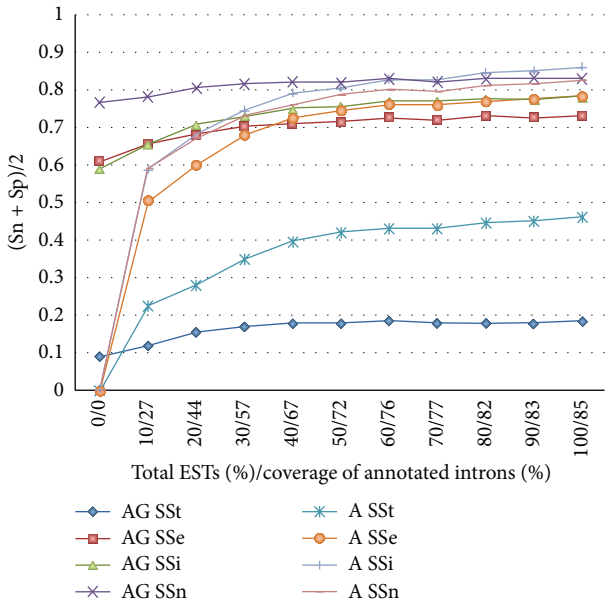


Figure 4: ASPic and ASPic-GeneID performance on ENCODE according to EST coverage. The accuracy [(Sn + Sp)/2] of ASPic CDS predictions (A) and ASPic-GeneID predictions (AG) is plotted according to the proportion of available ESTs given as input to ASPic. The percent coverage of annotated introns by the input ESTs is also given. SSn, SSe, SSi, and SSt indicate the accuracy at nucleotide, exon, intron, and transcript level, respectively.

3.3. ASPic-GeneID Accuracy with Alternative Splicing. In order to improve the performance of ASPic-GeneID, especially in mammalian gene finding, we implemented the possibility to predict alternative transcripts. In particular, we addressed the alternative splicing task with two independent procedures that should be considered simple extensions of ASPic-GeneID. In the first procedure, we predicted alternative variants by ASPic in genomic regions fully covered by ESTs and then we added all ASPic-GeneID predictions not overlapping ASPic transcripts. This procedure, called here ASPic-GeneID_AS1, combines in the simplest manner ASPic and ASPic-GeneID predictions, giving rise to ASPic predictions because they are directly deduced by expression data. The second procedure, called here ASPic-GeneID_AS2, is, instead, mainly dependent on GeneID. As for ASPic-GeneID_AS1, a complete pool of alternative transcripts was obtained by ASPic. Each predicted variant (represented by a set of exon-linked introns) was then given to GeneID as evidence. All predicted transcripts were finally combined and filtered in order to produce the final nonredundant set of transcript predictions (more details in Section 2).

Overall, results from ASPic-GeneID_AS1 and ASPic-GeneID_AS2 on the 44 ENCODE human regions are quite similar and overlapping (Table 2). Both procedures outperform ASPic-GeneID at all levels in the BTP comparison. Although ASPic-GeneID_AS1 is not more sensitive than ASPic-GeneID_AS2, it appears to be more specific (as it directly utilizes all ASPic transcripts). As shown in Table 2, all results can be compared to those from programs that currently use all available evidence or protein alignments to improve gene prediction in human.

In particular, when predictions are evaluated at gene level, ASPic-GeneID_AS2 is the most sensitive in finding genes (96%), exons (90%), introns (90%), and nucleotides (93%). In contrast, ASPic-GeneID_AS1 is the most specific at the BTP exon, intron and nucleotide levels. Focusing on methods using ESTs and mRNAs alignments but excluding proteins, ASPic-GeneID_AS1 is as accurate as PAIRAGON-any (49%) and 2% more accurate than Exogean (47%), indicated as the best gene finding program by the EGASP assessment, in predicting exact transcript structures. Moreover, ASPic-GeneID_AS1 has a transcript sensitivity of 65% which is the highest registered up to now. In the comparison with one of the most widely used pipelines as Ensembl, ASPic-GeneID_AS1 is 2% more accurate at both exon and intron levels. On the other hand, at transcript level both our systems are on average 22% and 3.5% more accurate than Ensembl in finding transcripts and exons, respectively.

On the whole, as shown in Table 2, ASPic-GeneID_AS1 and ASPic-GeneID_AS2 appear to outperform also many other well-established gene prediction tools at different measures. Although it is difficult to assess which program is really the best annotation system for human ENCODE regions, our simple methods that use only ESTs as main source of evidence prove highly competitive and comparable to very complex pipelines.

When we move to the *C. elegans* genome in which the impact of alternative splicing is low, the performances of ASPic-GeneID_AS1 and ASPic-GeneID_AS2 are slightly better than ASPic-GeneID at all levels. However, the possibility to predict alternative transcripts improves the global finding of exact transcripts and exons. At the gene level ASPic-GeneID_AS2 seems to be more accurate than ASPic-GeneID_AS1. In contrast, at transcript level, ASPic-GeneID_AS1 appears to be more efficient than ASPic-GeneID_AS2 since it directly uses ASPic inferred transcripts.

3.4. ASPic-GeneID and Deep Transcriptome Sequencing. ASPic-GeneID has been developed to handle long transcriptome sequences as main biological evidence to improve gene structures and detect potential alternative splicing transcripts.

Current high-throughput sequencing methodologies as RNA-Seq aim to provide global overview of entire transcriptomes. However, huge amount of short reads from Illumina or SOLiD platforms pose other challenges than classical Sanger ESTs and in many cases the detection of reliable transcripts is not optimal. Long-reads, therefore, as those from Sanger sequencing or the Roche 454 sequencer (Titanium chemistry with reads longer than 500 bases) represent the main source of evidence to reliably identify splice sites and alternative isoforms, other than simplify the deciphering of complex eukaryotic gene structures.

ASPic-GeneID is ready to analyse long EST-like reads from modern sequencer as Roche 454 and very long reads that are coming with the third generation sequencing platforms as PacBio. Although ESTs and ESTs-like sequences are

optimal for our pipeline, in principle it could work with Illumina reads. However, computational times are expected to be very onerous and no extensive tests have been performed to assess the biological quality of results.

4. Conclusions

Despite the advent of novel sequencing technologies [47, 48], the accurate genome annotation is yet a hard and challenging task. In order to improve such a process once a new genome sequence has been completed, we proposed here a simple computational strategy to accurately identify coding regions employing expressed sequences, mostly ESTs. Our framework, called Aspic-GeneID, uses EST based predictions by Aspic to improve *ab initio* gene structures by GeneID. Moreover, it can predict alternative transcripts providing a global view of the transcriptome. Aspic-GeneID is quite flexible depending on EST coverage. In organisms with a low impact of alternative splicing as *C. elegans*, it provides optimal predictions resulting in one of the most accurate gene finding programs. In contrast, when the impact of alternative splicing is high as in human, it can outperform existing gene finders at different levels. Moreover, the ability to predict multiple transcripts per gene locus makes Aspic-GeneID results comparable with those from very complicated pipelines like Ensembl, PAIRAGON-any, or AUGUSTUS-any that tend to use all available evidence.

Our strategy is based on expressed sequences as ESTs, but it can be easily applied to transcriptome sequences generated by next generation sequencing technologies. Indeed, recent tools as Cufflinks [49] can predict alternative transcripts and individual introns, making our methodology extremely recent and useful to improve genome annotations also in absence of canonical ESTs (generally produced by Sanger sequencing).

Authors' Contributions

Tyler Alioto and Ernesto Picardi contributed equally to this work.

Acknowledgments

The authors thank Francisco Camara for technical advice about the GeneID software. This work was supported by the Italian Ministero dell'Istruzione, Università e Ricerca (MIUR): PRIN 2009 and 2010; Consiglio Nazionale delle Ricerche: Flagship Project Epigen, Aging Program 2012–2014, and by the Italian Ministry for Foreign Affairs (Italy-Israel actions).

References

[1] J. Zhang, R. Chiodini, A. Badr, and G. Zhang, "The impact of next-generation sequencing on genomics," *Journal of Genetics and Genomics*, vol. 38, no. 3, pp. 95–109, 2011.

[2] E. Picardi and G. Pesole, "Computational methods for ab initio and comparative gene finding," *Methods in Molecular Biology*, vol. 609, pp. 269–284, 2010.

[3] T. Alioto, "Gene prediction," *Methods in Molecular Biology*, vol. 855, pp. 175–201, 2012.

[4] I. Korf, "Gene finding in novel genomes," *BMC Bioinformatics*, vol. 5, article 59, 2004.

[5] M. Stanke, R. Steinkamp, S. Waack, and B. Morgenstern, "AUGUSTUS: a web server for gene finding in eukaryotes," *Nucleic Acids Research*, vol. 32, pp. W309–W312, 2004.

[6] S. L. Cawley and L. Pachter, "HMM sampling and applications to gene finding and alternative splicing," *Bioinformatics*, vol. 19, supplement 2, pp. ii36–ii41, 2003.

[7] J. S. Pedersen and J. Hein, "Gene finding with a hidden Markov model of genome structure and evolution," *Bioinformatics*, vol. 19, no. 2, pp. 219–227, 2003.

[8] R. Guigó, P. Flicek, J. F. Abril et al., "EGASP: the human ENCODE Genome Annotation Assessment Project," *Genome Biology*, vol. 7, supplement 1, pp. S2.1–S2.31, 2006.

[9] C. Wei and M. R. Brent, "Using ESTs to improve the accuracy of de novo gene prediction," *BMC Bioinformatics*, vol. 7, article 327, 2006.

[10] G. Parra, P. Agarwal, J. F. Abril, T. Wiehe, J. W. Fickett, and R. Guigó, "Comparative gene prediction in human and mouse," *Genome Research*, vol. 13, no. 1, pp. 108–117, 2003.

[11] M. J. van Baren, B. C. Koebbe, and M. R. Brent, "Using N-SCAN or TWINSCAN to predict gene structures in genomic DNA sequences," in *Current Protocols in Bioinformatics*, A. D. Baxevanis, Ed., chapter 4, unit 4.8, 2007.

[12] S. S. Gross, C. B. Do, M. Sirota, and S. Batzoglou, "CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction," *Genome Biology*, vol. 8, no. 12, article R269, 2007.

[13] O. Jaillon, J.-M. Aury, B. Noel et al., "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla," *Nature*, vol. 449, no. 7161, pp. 463–467, 2007.

[14] "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana," *Nature*, vol. 408, no. 6814, pp. 796–815, 2000.

[15] G. A. Tuskan, S. DiFazio, S. Jansson et al., "The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)," *Science*, vol. 313, no. 5793, pp. 1596–1604, 2006.

[16] The Tomato Genome Consortium, "The tomato genome sequence provides insights into fleshy fruit evolution," *Nature*, vol. 485, no. 7400, pp. 635–641, 2012.

[17] R. Guigó and M. G. Reese, "EGASP: collaboration through competition to find human genes," *Nature Methods*, vol. 2, no. 8, pp. 575–577, 2005.

[18] N. de Souza, "The ENCODE project," *Nature Methods*, vol. 9, no. 11, article 1046, 2012.

[19] L. L. Elnitski, P. Shah, R. T. Moreland, L. Umayam, T. G. Wolfsberg, and A. D. Baxevanis, "The ENCODEdb portal: simplified access to ENCODE consortium data," *Genome Research*, vol. 17, no. 6, pp. 954–959, 2007.

[20] J. Harrow, F. Denoeud, A. Frankish et al., "GENCODE: producing a reference annotation for ENCODE," *Genome Biology*, vol. 7, supplement 1, pp. S4.1–S4.9, 2006.

[21] S. H. Nagaraj, R. B. Gasser, and S. Ranganathan, "A hitchhiker's guide to Expressed Sequence Tag (EST) analysis," *Briefings in Bioinformatics*, vol. 8, no. 1, pp. 6–21, 2007.

[22] M. Stanke, A. Tzvetkova, and B. Morgenstern, "AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome," *Genome Biology*, vol. 7, supplement 1, pp. S11.1–S11.8, 2006.

[23] A. Krogh, "Using database matches with HMMGene for automated gene detection in Drosophila," *Genome Research*, vol. 10, no. 4, pp. 523–528, 2000.

[24] M. Arumugam, C. Wei, R. H. Brown, and M. R. Brent, "Pairagon+N-SCAN_EST: a model-based gene annotation pipeline," *Genome Biology*, vol. 7, supplement 1, pp. S5.1–S5.10, 2006.

[25] S. Djebali, F. Delaplace, and H. R. Crollius, "Exogean: a framework for annotating protein-coding genes in eukaryotic genomic DNA," *Genome Biology*, vol. 7, supplement 1, pp. S7.1–S7.10, 2006.

[26] T. Castrignanò, R. Rizzi, I. G. Talamo et al., "ASPIC: a web resource for alternative splicing prediction and transcript isoforms characterization," *Nucleic Acids Research*, vol. 34, pp. W440–W443, 2006.

[27] P. Bonizzoni, R. Rizzi, and G. Pesole, "ASPIC: a novel method to predict the exon-intron structure of a gene that is optimally compatible to a set of transcript sequences," *BMC Bioinformatics*, vol. 6, article 244, 2005.

[28] E. Blanco, G. Parra, and R. Guigo, "Using geneid to identify genes," in *Current Protocols in Bioinformatics*, A. D. Baxevanis, Ed., chapter 4, unit 4.3, 2007.

[29] G. Parra, E. Blanco, and R. Guigó, "GeneId in *Drosophila*," *Genome Research*, vol. 10, no. 4, pp. 511–515, 2000.

[30] T. D. Wu and C. K. Watanabe, "GMAP: a genomic mapping and alignment program for mRNA and EST sequences," *Bioinformatics*, vol. 21, no. 9, pp. 1859–1875, 2005.

[31] W. J. Kent, "BLAT—the BLAST-like alignment tool," *Genome Research*, vol. 12, no. 4, pp. 656–664, 2002.

[32] G. format, http://www.sanger.ac.uk/Software/formats/GFF/.

[33] S. M. J. Searle, J. Gilbert, V. Iyer, and M. Clamp, "The Otter annotation system," *Genome Research*, vol. 14, no. 5, pp. 963–970, 2004.

[34] M. Burset and R. Guigó, "Evaluation of gene structure prediction programs," *Genomics*, vol. 34, no. 3, pp. 353–367, 1996.

[35] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[36] A. Coghlan and R. Durbin, "Genomix: a method for combining gene-finders' predictions, which uses evolutionary conservation of sequence and intron-exon structure," *Bioinformatics*, vol. 23, no. 12, pp. 1468–1475, 2007.

[37] S. W. Roy and D. Penny, "Intron length distributions and gene prediction," *Nucleic Acids Research*, vol. 35, no. 14, pp. 4737–4742, 2007.

[38] D. V. Lu, R. H. Brown, M. Arumugam, and M. R. Brent, "Pairagon: a highly accurate, HMM-based cDNA-to-genome aligner," *Bioinformatics*, vol. 25, no. 13, pp. 1587–1593, 2009.

[39] D. S. Gerhard, "The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC)," *Genome Research*, vol. 14, no. 10, pp. 2121–2127, 2004.

[40] V. Solovyev, P. Kosarev, I. Seledsov, and D. Vorobyev, "Automatic annotation of eukaryotic genes, pseudogenes and promoters," *Genome Biology*, vol. 7, pp. S10.11–S10.12, 2006.

[41] V. Curwen, E. Eyras, T. D. Andrews et al., "The Ensembl automatic gene annotation system," *Genome Research*, vol. 14, no. 5, pp. 942–950, 2004.

[42] A. A. Salamov and V. V. Solovyev, "Ab initio gene finding in Drosophila genomic DNA," *Genome Research*, vol. 10, no. 4, pp. 516–522, 2000.

[43] J. Q. Wu, D. Shteynberg, M. Arumugam, R. A. Gibbs, and M. R. Bren, "Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing," *Genome Research*, vol. 14, no. 4, pp. 665–671, 2004.

[44] D. Weissglas-Volkov, C. L. Plaisier, A. Huertas-Vazquez et al., "Identification of two common variants contributing to serum apolipoprotein B levels in mexicans," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 30, no. 2, pp. 353–359, 2010.

[45] J. Besemer and M. Borodovsky, "GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses," *Nucleic Acids Research*, vol. 33, no. 2, pp. W451–W454, 2005.

[46] B. Brejová, D. G. Brown, M. Li, and T. Vinař, "ExonHunter: a comprehensive approach to gene finding," *Bioinformatics*, vol. 21, supplement 1, pp. i57–i65, 2005.

[47] M. L. Metzker, "Sequencing technologies the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[48] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends in Genetics*, vol. 24, no. 3, pp. 133–141, 2008.

[49] C. Trapnell, A. Roberts, L. Goff et al., "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," *Nature Protocols*, vol. 7, no. 3, pp. 562–578, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 505864, 17 pages
http://dx.doi.org/10.1155/2013/505864

Research Article

Systems Approaches to Modeling Chronic Mucosal Inflammation

Mridul Kalita,[1] Bing Tian,[2] Boning Gao,[3] Sanjeev Choudhary,[1,2,4] Thomas G. Wood,[1,4,5] Joseph R. Carmical,[5] Istvan Boldogh,[1,6] Sankar Mitra,[1,5] John D. Minna,[3] and Allan R. Brasier[1,2,4]

[1] *Sealy Center for Molecular Medicine, The University of Texas Medical Branch, 301 University Boulevard, Galveston, TX 77555, USA*
[2] *Department of Internal Medicine, The University of Texas Medical Branch, 301 University Boulevard, Galveston, TX 77555, USA*
[3] *Hamon Center for Therapeutic Oncology Research, Department of Internal Medicine Pharmacology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA*
[4] *Institute for Translational Sciences, The University of Texas Medical Branch, 301 University Boulevard, Galveston, TX 77555, USA*
[5] *Departments of Biochemistry and Molecular Biology, The University of Texas Medical Branch, 301 University Boulevard, Galveston, TX 77555, USA*
[6] *Microbiology and Immunology, The University of Texas Medical Branch, 301 University Boulevard, Galveston, TX 77555, USA*

Correspondence should be addressed to Allan R. Brasier; arbrasie@utmb.edu

Received 12 June 2013; Revised 8 August 2013; Accepted 9 August 2013

Academic Editor: Tao Huang

The respiratory mucosa is a major coordinator of the inflammatory response in chronic airway diseases, including asthma and chronic obstructive pulmonary disease (COPD). Signals produced by the chronic inflammatory process induce epithelial mesenchymal transition (EMT) that dramatically alters the epithelial cell phenotype. The effects of EMT on epigenetic reprogramming and the activation of transcriptional networks are known, its effects on the innate inflammatory response are underexplored. We used a multiplex gene expression profiling platform to investigate the perturbations of the innate pathways induced by TGFβ in a primary airway epithelial cell model of EMT. EMT had dramatic effects on the induction of the innate pathway and the coupling interval of the canonical and noncanonical NF-κB pathways. Simulation experiments demonstrate that rapid, coordinated cap-independent translation of TRAF-1 and NF-κB2 is required to reduce the noncanonical pathway coupling interval. Experiments using amantadine confirmed the prediction that TRAF-1 and NF-κB2/p100 production is mediated by an IRES-dependent mechanism. These data indicate that the epigenetic changes produced by EMT induce dynamic state changes of the innate signaling pathway. Further applications of systems approaches will provide understanding of this complex phenotype through deterministic modeling and multidimensional (genomic and proteomic) profiling.

1. Introduction

Respiratory epithelial cells provide the principal barrier of the airways, facilitating gas exchange and mucociliary particulate clearance and are the major source of protective airway lining fluid [1, 2]. In the presence of injury, inflammation, and airway remodeling, signal transduction pathways induce global epigenetic reprogramming events to induce type II epithelial-mesenchymal transition (EMT) within a so-called epithelial mesenchymal trophic unit [3–7]. In the setting of cellular transformation, EMT is also implicated in cancer metastasis. During EMT, epithelial cells lose apicobasal polarity and decrease the expression of intercellular tight junctions (TJs), adherens junctions (AJs), and desmosomes [8, 9]. These changes lead to disruption of adhesion of the basal epithelial layer and allow cellular penetration into an extracellular matrix (ECM), promoting enhanced ECM production and fibrosis. As a result, type II EMT plays a central role in normal tissue response to injury and tissue remodeling and repair, whereas type III EMT is linked to cancer dissemination [10].

Despite the key role of EMT in normal tissue development and repair, dysregulated EMT has been proposed

to be responsible for extracellular matrix (e.g., collagen) overproduction and fibrosis occurring in chronic respiratory diseases such as idiopathic pulmonary fibrosis [11, 12], asthma [13], the chronic obstructive pulmonary disease (COPD) [14], and obliterative bronchiolitis [4]. Genetic lineage studies in rodent models of airway fibrosis have shown that alveolar epithelial cells are a major source of progenitor cells for repair of the injured airway [15]. Alveolar cells undergoing EMT become dedifferentiated, capable of both self-renewal and formation of fibroblastic foci, with the latter constituting sites of active fibrogenesis. EMT-activated alveolar epithelial cells synthesize a variety of fibrogenic cytokines, including TGFβ, TNFα, endothelin-1 and connective tissue growth factor (CTGF) [16], ECM components (Col 1A), and matrix metalloproteases (MMPs), all contributing to the pathological extracellular matrix remodeling. The process of dysregulated EMT has also been implicated in cancer progression. Here, EMT transition promotes growth factor independence, resistance to chemotherapeutic agents, and acquisition of stem-cell-like phenotype. These latter events are partly responsible for the maintenance of cancer and metastatic behavior.

Mechanistic studies have shown that EMT is initiated by a wide variety of agents linked to chronic inflammation including oxidative stress (ROS) [16], growth factors (TGF, EGF, and IGF) [17], and cytokines (TNFα). In the airways, these factors are produced by injured epithelial cells, fibroblasts, and eosinophils [18]. The prototypical inducer of type II EMT is TGFβ1, a cytokine master regulator that induces EMT via the canonical TGFβR1-Smad3-dependent signaling pathway [13]. The TGFβ-Smad3 pathway activates downstream Wnt, notch, and NF-κB signaling pathways to coordinate the complex genetic changes underlying EMT [19]. The core transcriptional regulators of the EMT program include the transcriptional regulators SNAIL (SNAI)1/2, Zebra (ZEB), Twist, and FOXC2 proteins. These factors coordinate acquisition of the mesenchymal phenotype through (1) downregulation of epithelial cadherin (E-Cad) involved in maintenance of cell polarity, (2) induction of intermediate filament proteins such as vimentin, (3) activation of small GTPases (to induce motility), and (4) expression of matrix metalloproteinases (MMPs) and collagen to induce fibrogenesis. Although EMT can be reversible, stable EMT is maintained through global epigenetic reprogramming, including the reorganization (and increase) of the transcriptionally active histone (H3) Lys (K) 36 trimethyl marks within large organized heterochromatin domains [20].

In addition to their central role in airway repair, epithelial cells also function as sentinel cells to trigger innate host response to microbial and nonmicrobial challenges/invasion [21]. These cells are activated by a plethora of processes including pathogen-associated molecular patterns (PAMPs) [22]. The presence of PAMPS triggers the innate intracellular signaling pathways, converging on the NF-κB and IRF3 signaling pathways [23]. Of these, the NF-κB signaling pathway is composed of two coupled pathways known as the canonical and noncanonical pathways, distinguished by the regulatory kinase and different cytoplasmic reservoir from which NF-κB is activated [23–25]. In the rapidly activated canonical pathway, induced within minutes of stimulation, sequestered RelA·50 kDa NF-κB1 heterodimers in the cytoplasm are liberated and enter the nucleus to activate numerous genes including proinflammatory and antiapoptotic ones [26–28]. By contrast, the slower noncanonical NF-κB pathway, induced within hours of stimulation, is mediated by MAP3K14/NF-κB-inducing kinase (NIK)·IKKα complex that processes the 100 kDa precursor into 52 kDa NF-κB2 (p52), liberating the RelA·p52 heterodimer for nuclear import. We have recently shown that the slower noncanonical NF-κB pathway is linked to the canonical pathway by a feed-forward module consisting of the TNF receptor (TRAF)-1 factor that complexes with, stabilizes, and activates the NIK·IKKα complex [24]. The noncanonical pathway induces expression of a temporally and biologically distinct group of genes [27].

The complex phenotype produced by EMT is beginning to be examined using an unbiased profiling coupled with computational inference [29]. Although the mechanism of inducing EMT has been extensively explored, much less is known about the effect of EMT on the network of innate inflammatory signaling pathways. By interrogating a primary human airway epithelial cell model of type II EMT using a multiplex gene expression profiling platform, we describe here how EMT perturbs the innate response. Here, we observe that the EMT phenotype dramatically affects the kinetics and patterns of the innate response. Remarkably, alterations in the NF-κB pathway are mediated by transcriptional elongation mediated by enhanced phospho-Ser^2 carboxy terminal domain (CTD) of RNA Pol II binding to innate gene promoters. Dynamic deterministic modeling shows that the accelerated expression of TRAF1 and NF-κB2 in a coordinated IRES-dependent manner produces a shorter canonical-noncanonical NF-κB coupling constant. These studies suggest that a global analysis of the EMT phenotype can be further examined using computational deterministic modeling, multidimensional profiling, and analysis of large-scale chromatin structure.

2. Materials and Methods

2.1. RNA Extraction and Quantitation. Total cellular RNA was extracted using either RNAqueous phenol-free total RNA isolation kits (Life Technologies, CA) or Quick-RNA MiniPrep kits (ZYMO Research) according to the manufacture's recommendations. RNA was quantitated spectrophotometrically using a NanoDrop ND-1000 (NanoDrop Technologies, DE). The quality of the purified RNA was assessed by the visualization of 18S and 28S RNA bands using an Agilent BioAnalyzer 2100 (Agilent Technologies, CA). The resulting electropherograms were used in the calculation of the 28S/18S ratio and the RNA integrity number [30].

2.2. Assay Design and Validation. Real-time quantitative RT-PCR (QPCR) assays were designed from the coding sequence (CDS) of the gene of interest (NCBI), and exon-exon junctions were mapped via BLAT [31]. Whenever possible, at least one of the two PCR primers was designed to transcend an exon-intron junction in order to reduce

the potential impact of genomic DNA contamination in the surveyed RNA samples. Primers were designed using Primer Express 2.0 (Applied Biosystems) with default settings (Primer T_m = 58°C–60°C, GC content = 30–80%, Length = 9–40 nucleotides, and Amplicon Length = 90–150 nucleotides). Primers were synthesized (IDT) and reconstituted to a final concentration of 100 uM (master stock) and a working stock of 5 uM. Specificity of each assay was confirmed via PCR under the reaction conditions listed below and analyzed by PAGE. A reverse transcriptase minus control was included to determine the existence of signal contribution from genomic DNA. Assays were formatted using a 96-well platform.

2.3. Real-Time QPCR Analysis of Gene Expression. Reverse transcription was performed on 1 ug of total RNA with random primers, utilizing TaqMan reverse transcription reagents (Applied Biosystems) under conditions recommended by the manufacturer. Although the mass of input RNA should not be utilized for normalization purposes, the amounts of input RNA to be assayed were equivalent. The reverse transcription product was used as a template for the subsequent PCR reaction, consisting of SYBR Green PCR Master Mix, template cDNA, and assay primers (Table 1) in a total reaction volume of 25 uL. Thermal cycling (50°C, 2 min; 95°C, 10 min; and 40 cycles at 95°C, 15 S; 60°C, 1 min) was performed using an ABI prism 7500 sequence detection system (Life Technologies, CA). Threshold cycle numbers (C_t) were defined as fluorescence values, generated by SYBR Green binding to double stranded PCR products, exceeding baseline. Relative transcript levels were quantified as a comparison of measured C_t values for each reaction, normalized using a reference assay for human polymerase beta (NM_002690; Fwd: 5′ACAATCAATGAGTACACCATCCGT3′; Rev: 5′TCCTGCAACTCCAGTGACTCC3′) and compared to those of a "control sample" (SAEC TGFβ minus) [32].

2.4. Cell Culture and EMT Transformation. The primary (nonimmortalized) human small airway epithelial cell (HSAEC) was purchased from Lonza. An immortalized cell line was established by infecting primary HSAECs with human telomerase (hTERT) and cyclin dependent kinase (CDK) 4 expressing retrovirus constructs and selecting under 250 ng/mL puromycin and 30 μg/mL G418 as described in [33]. The immortalized HSAECs were grown in small airway epithelial cell growth medium (SAGM; Lonza, Walkersville, MD) in a humidified atmosphere of 5% CO_2. The immortalized HSAEC shows characteristics of normal cells such as contact inhibition of growth and failure to form soft agar colonies or form tumors in immune compromised mice. Immortalized human bronchial epithelial cells (HBEC) were grown in Keratinocyte serum-free medium (K-SFM, Life Technologies. Grand Island, NY) as described previously [33]. A549 cells were grown in RPMI supplemented with 5% fetal bovine serum. For induction of EMT, cells were TGFβ treated for 5 d (10 ng/mL, R&D systems, Minneapolis, MN) in the growth medium.

2.5. Dual Cross-Link Chromatin Immunoprecipitation (XChIP). XChIP was performed as described previously [34, 35]. A549 cells (~6 × 10^6 per 100-mm dish) were washed twice with PBS. Protein-protein cross-linking was first performed with disuccinimidyl glutarate (2 mM, Pierce), followed by protein-DNA cross-linking with formaldehyde. Equal amounts of sheared chromatin were immunoprecipitated overnight at 4°C with 4 μg of the indicated Ab in ChIP dilution buffer. Immunoprecipitates were collected with 40 μL protein A magnetic beads (Dynal Inc.), washed, and eluted in 250 μL elution buffer for 15 min at room temperature. Samples were de-cross-linked in 0.2 M NaCl at 65°C for 2 h. The precipitated DNA was phenol-chloroform extracted, precipitated with 100% ethanol, and dried.

Gene enrichment in XChIP was determined by quantitative real-time genomic PCR (Q-gPCR) as previously described [34] using region-specific PCR primers (Table 1). Standard curves were generated using a dilution series of genomic DNA (from 1 ng to 100 ng) for each primer pair. The fold change of DNA in each immunoprecipitate was determined by normalizing the absolute amount to the input DNA reference and calculating the fold change relative to that amount in unstimulated cells.

2.6. Deterministic Mathematical Modeling and Simulations. The deterministic mathematical model for both canonical and non-canonical arms of NF-κB pathway is recently published where we have shown that TRAF1·NIK acts as a central, rate-limiting feed-forward signaling complex to activate noncanonical pathway [24]. The complete model consists of 28 ordinary differential equations (ODEs) and 58 parameters. The expression of genes in the canonical pathway is RelA-dependent, whereas, the noncanonical pathway genes are p52-dependent. We observed that there is a time delay in TRAF1 translation by nearly 120 min. Likewise, a delay of 90 min between NFκB2 mRNA expression and translation has been reported earlier [36]. To examine the functional significance of these time delays in noncanonical pathway coupling, we performed different sets of simulations. In each simulation, the time delay function of TRAF1 and NFκB2 translation was altered by the addition or subtraction of 15 min on either side of nominal value, while keeping the translation rate the same in each simulation. These simulations covered the time delay span of 0–10 h for both TRAF1 and NFκB2 to study the effects of increasing and decreasing translation times on noncanonical pathway gene expression. In one set of simulations, we perturbed the system by varying the time-delay function of either TRAF1 or NFκB2 one at a time and keeping all other rate constants at nominal values. In another set, we perturbed the system by varying these time-delay functions of both TRAF1 and NF-κB2 simultaneously while keeping all other rate constants at nominal values. While the first set generated 42 such simulations for each protein, the second set generated more than 1600 such simulations. The latter will determine the synergistic effect of TRAF1 and NF-κB2 on the gene expression of the noncanonical pathway.

Table 1: Primer sequences in Q-RT-PCR assay.

Pathway	Gene	Accession #	Amplicon (bp)	Forward (5′-3′)	Reverse (5′-3′)
Canonical	NFKB1A	NM_020529	116	CCGCAGGAGGTGCCG	ATCACTTCCATGGTCAGTGCC
	CXCL2/GroB	NM_002089	100	ATTCACCTCAAGAACATCCAAAGTG	GCCCATTCTTGAGTGTGGCTAT
	TNFA1P3/A20	NM_001270508	111	GAAGCACCATGTTTGAAGGATACTG	CTCTGCGCTGGCTCGATC
	IL-6	NM_000600	90	CTGGATTCAATGAGGAGACTTGC	TCAAATCTGTTCTGGAGGTACTCTAGG
	IL-8	NM_000584	92	AAGACATACTCCAAACCTTTCCACC	CAATAATTTCTGTGTTGGCGCA
	TSLP	NM_033035	91	TCTTGTAGCAATCGGCCACA	ACATTTCTTTGGCGAGCGA
	TRAF1	NM_005658	91	TGGAAGATCACCAATGTCACCA	ATACTTGGCAGTGTAGAAGGCTGG
	TRAF3	NM_145752	91	GAAGGCGTGTAAATACCGGG	ACAGTCGGTGTCTTCGTGTTTCT
	NFKB2	NM_001077494	96	ACATGACTGCCCAATTTAACAACC	GGAGCCGCTGCCTCTGA
	IL-25	NM_022789	91	CACCCAGAGTCCTGTAGGGC	GGTTCAAGTCTCTGTCCAACTCATATC
	IL-33	NM_033439	51	AACACCCCTCAAATGAATCAGGT	TTGGCATGCAACCAGAAGTCT
Noncanonical EMT	TNIP1	NM_001252385	92	ATCCAGTGGCACCTCCTCTG	CCAGCGCCATCGCATT
	SMA/ACTA2	NM_001141945	91	TGTAAGGCCGGCTTTGCT	TTCCCACCATCACCCCCT
	Col1A1	NM_000088	95	CCAGAAGAACTGGTACATCAGCA	CGCCATACTCGAACTGGAATC
	Vimentin	NM_003380	91	GCTCAATGTTAAGATGGCCCTT	TGGAAGAGGCAGAGAAATCCTG
	Desmin	NM_001927	91	GGAGAGGAGAGCCGGATCA	GGACCTCAGAACCCCTTTGC
	Twist1	NM_000474	101	TCTCGGTCTGGAGGATGGA	CAATGACATCTAGGTCTCCG
	Twist2	NM_001271893	119	ACGAGCGCCTCAGCTACG	CGCGACGGACAGCCCTG
	SLUG	AF084243	91	TGTGTGGACTACCGCTGCTC	ACTCACTCGCCCCAAAGATG
	SNAI1	NM_005985	95	GCGCTCTTTCCTCGTCAGG	GGGCTGCTGGAAGGTAAACTCT
	ITGA2	NM_002203	91	AGCCGAAGTACCAACAGGAGTTATA	GCCGAGCTTCCATAAAATTGC
	FSP1/S100A4	NM_002961	91	AGGGTGACAAGTTCAAGCTCAAC	GCTTCATCTGTCCTTTTCCCC
	SMA/ACTA2	NM_001141945	91	TGTAAGGCCGGCTTTGCT	TTCCCACCATCACCCCCT
IFN	IFNb	NM_002176	95	GCAGTTCCAGAAGGAGGACG	TCCAGCCAGTGCTAGATGAATC
	IFNa6	NM_021002	91	GTGGTGCTCAGCTGCAAGTC	CCAGGAGCATCATGGTCCTC
	IFNa21	NM_002175	91	TGATCTGCCTCAGACCCACA	CTTCAGGCAGGAGAAAGGAGAG
	CCL5/RANTES	NM_002985	96	TCTACACCAGTGGCAAGTGCTC	CCCGAACCCATTTCTTCTCTG
	DDX58/RIG1	NM_014314	132	CCACTTAAACCCAGAGACAATAACAA	TTGCCACGTCCAGTCAATATG
	LMP2	NM_002800	91	TTCACCACAGACGCTATTGCTC	CCACACCGGCAGCTGTAAT
	TAP1	NM_000593	91	GTTTTTCCAACAGAACCAGACAGG	GCTCAGATTCTCACTCAGAGAATCACT
	STAT1	NM_007315	104	TCCTGCTGCGGTTCAGTG	GGGTTCAACCGCATGGAAG
	STAT2	NM_005419	118	CCTGAAACACAGGCTCATTGTG	TGGCACCAGCCCTAGTTCC
	IRF1	NM_002198	91	AGCAAGGCCAAGAGGAAGTCA	TGCTGTGGTCATCAGGCAGA
	IRF9	NM_006084	91	AGCCACAGGAAGTTACAGACACAA	GCAGTGAGTAGTCTGGCTCTGGA
	SOCS3	NM_003955	91	CTTTCTGATCCGCGACAGCT	ACACTGGATGCGCAGGTTC
	IF127	NM_001130080	91	CAGTCACTGGGAGCAACTGGA	GCCCAGGATGAACTTGGTCA
	MX1	NM_001144925	106	GAACCACCCATATTTCAGGGATC	ATGTGTGATGAGCTCGCTGGTA
	IFNL1 (IL-29)	NM_172140	103	AAGCCCACCACAACTGGG	ACTCTTCCAAGGCGTCCC
	IFNL2 (IL-28A)	NM_172138	101	ACATGACTGGGGACTGCAC	ATCCGGGAGAGCCCCGT
	IFNL3 (IL-28B)	NM_172139	99	ATGACCGGGGACTGCATG	ATCCGGGAGAGCCCCGC
Growth factor	Fos	NM_005252	91	GGGCAAGGTGGAACAGTTATCT	GTTGCGGCATTTGGCTG
	c-Jun	NM_002228	91	GTCCCAGGAGCGGATCAAG	GGCGATTCTCTCCAGCTTCC
	SOD2	NM_000636	90	TGCTTGTCCAAATCAGGATCC	TGAAGGTAGTAAGCGTGCTCCC
	TM4SF1	NM_014220	91	GAGGTGGCCTGCTGATGCT	CCACAGTTTTCATGGCCACAG

Table 1: Continued.

Pathway	Gene	Accession #	Amplicon (bp)	Forward (5′-3′)	Reverse (5′-3′)
ROS/DNA	NOX1	NM_007052	91	GGGCATCCCCCTGAGTCT	TCTGCTGGGAGCGGTAAAAC
	NOX4	NM_016931	91	ACTCAACACCCTGTTGGATGACT	CCAACGGAAGGACTGGATATCT
	HOXB9	NM_024017	91	GGCCGGATCAAACCAACC	TCCAGCGTCTGGTATTTGGTG
	ATM	NM_000051	127	GCTTCTCAGGATAATCCGCAAG	CCAAGCAGCTTCCAACAGC
	STK39 (SRT39)	NM_013233	91	GCCCAAAGAGCCAAAAAGGT	CGTCGTCACTCCACTCCCA
Cell cycle	PCNA	NM_002592	91	TGCGCCGGCAATGAA	CTTTCTCCTGGTTTGGTGCTTC
	CDKN1A	NM_000389	92	AGCAGGCTGAAGGGTCCC	GCGTTTGGAGTGGTAGAAATCTGT

2.7. Clustering Approaches. A clustered image map (CIM) of a normalized matrix was created that correlates gene expression pattern to different time points in HSAECs cells upon stimulation by TNFα and TGFβ. For each gene, mean and standard deviation were calculated from their expression fold changes (see Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2013/505864) across the time course and were normalized to unstimulated HSAECs [26–28]. Z-score transformation was calculated for each of the 50 genes (Table S2) by subtracting each fold change value by the row mean and dividing by the row standard deviation [37]. In essence, the Z-score gives an estimation of the deviation of the measurement from the row mean in standard deviation units. Hierarchical clustering was performed using an average-linkage clustering algorithm across six different time points in the absence or presence of stimulants. The cluster tree of genes is represented on the y-axis, and time-points and stimulants are shown on the x-axis. Each block of red or green represents a high positive or negative correlation between the gene expression and the stimulant under a specific time point.

3. Results and Discussion

3.1. Induction of Type II EMT. To establish a model of type II EMT, a continuously replicating line of human small airway epithelial cells was generated by immortalization using human telomerase (hTERT) and CDK4 expression [33]. These cells show a stable epithelial morphology and differentiated cytokeratin isoforms after over 100 population doublings, express the stem cell marker p63 and high levels of p16INK4a, and have an intact p53 checkpoint pathway [33].

To characterize type II EMT, HSAECs were incubated in the absence or presence of TGFβ (5 ng/mL) for 10 d. Transformed type II alveolar epithelial cells (A549) and immortalized bronchial epithelial cells (HBECs) were used as reference. Cells were fixed, stained with FITC-conjugated phalloidin (for distribution of F actin) and DAPI (a nuclear DNA stain), and examined by confocal microscopy. In the absence of TGFβ stimulation, HSAECs assumed a normal cuboidal morphology with perinuclear cytoplasmic distribution of F-actin (Figure 1(a)). By contrast, TGFβ-treated HSAECs showed an elongated shape with markedly induced F-actin staining (Figure 1(a)). This morphological change of enhanced front-rear polarity and cytoskeletal actin rearrangement are similar to those observed in TGFβ-treated A549 and HBECs; all are characteristic morphological changes of EMT [8].

To further confirm the induction of EMT, expression of ECM genes and EMT-associated transcription factors were assessed by Q-RT-PCR. TGFβ-treated HSAECs showed marked upregulation of extracellular matrix (ECM; Col1A), mesenchymal intermediate filament protein (vimentin), and TGFβ-induced transcription factors (SNAI1, Twist1/2) mediating the EMT genetic program (Figure 1(b)). This gene expression pattern is similar to those observed in other primary epithelial cells [13]. Together, these data suggest that TGFβ induces morphological and gene signatures of stable type II EMT in HSAECs.

3.2. Systems Profiling. Our goal was to establish a platform for the systematic perturbation of the signaling phenotype induced by stable type II EMT using reiterated rounds of stimulus perturbations and profiling measurements to inform the development of predictive models of complex behavior (schematically illustrated in Figure 2). For these experiments we developed a quantitative plate-based PCR assay to monitor the signature gene expression of the innate pathway, including the interferon, canonical and noncanonical NF-κB dependent pathways, the DNA damage response pathways, the EMT programs, JAK/STAT pathway, and growth factor pathways (Table 1).

3.3. Perturbations of Signaling Dynamics in Response to EMT. The time series experiments of the gene expression driven by the canonical NF-κB pathway were measured by using plate-based Q-RT-PCR in control and EMT-transformed HSAECs in response to the prototypical activating cytokine, TNFα. Relative changes in mRNA expression were normalized to DNA polymerase β as a housekeeping gene, and the data were Z-score-transformed to show deviations of expression in standard deviation units [26, 37]. The gene expression of the pathway was then visualized using hierarchical clustering (Figure 3, and Supplemental Tables). The HSAECs after EMT showed unique quantitative and qualitative expression patterns. For example, the top cluster contains members of the interferon-gamma (IFN) pathway such as ITGA2, myxovirus resistance 1 (MX1), signal transducer and activator of transcription (STAT1), proteasomal components low molecular weight proteins (LMP2), and transporters associated with antigen processing (TAP2); these genes are markedly downregulated as a function of EMT (Figure 3). By contrast, genes within the NF-κB pathway, growth factor response, and EMT signatures were markedly upregulated (Figure 3).

3.4. EMT Effects on the Canonical NF-κB Signaling Pathway. Because of the marked effects of the EMT state on the NF-κB pathway, we sought to further investigate this mechanism. Hierarchical clustering of the NF-κB-dependent gene pathway is shown in Figure 4. Genes that selectively respond to the canonical pathway include NFKBIA/IκBa and TNFAIP3/A20; we have shown that the expression of these genes depends on IκBα proteolysis and that they have direct NF-κB/RelA binding sites in their promoters [26, 38]. Greater upregulation of these genes was observed at the earliest time point measured, with a second wave of expression at 12–24 h after TNFα stimulation. This mimics the oscillatory behavior of the NF-κB signaling pathway observed in cancer cells [39].

3.5. EMT Affects the Coupling of the Noncanonical to the Canonical NF-κB Signaling Pathways. EMT effects on the time-dependent expression of the TNIP1/Naf1 gene, a hallmark of the noncanonical NF-κB pathway [24, 27], were also evident in the hierarchical clustering (Figures 3 and 4).

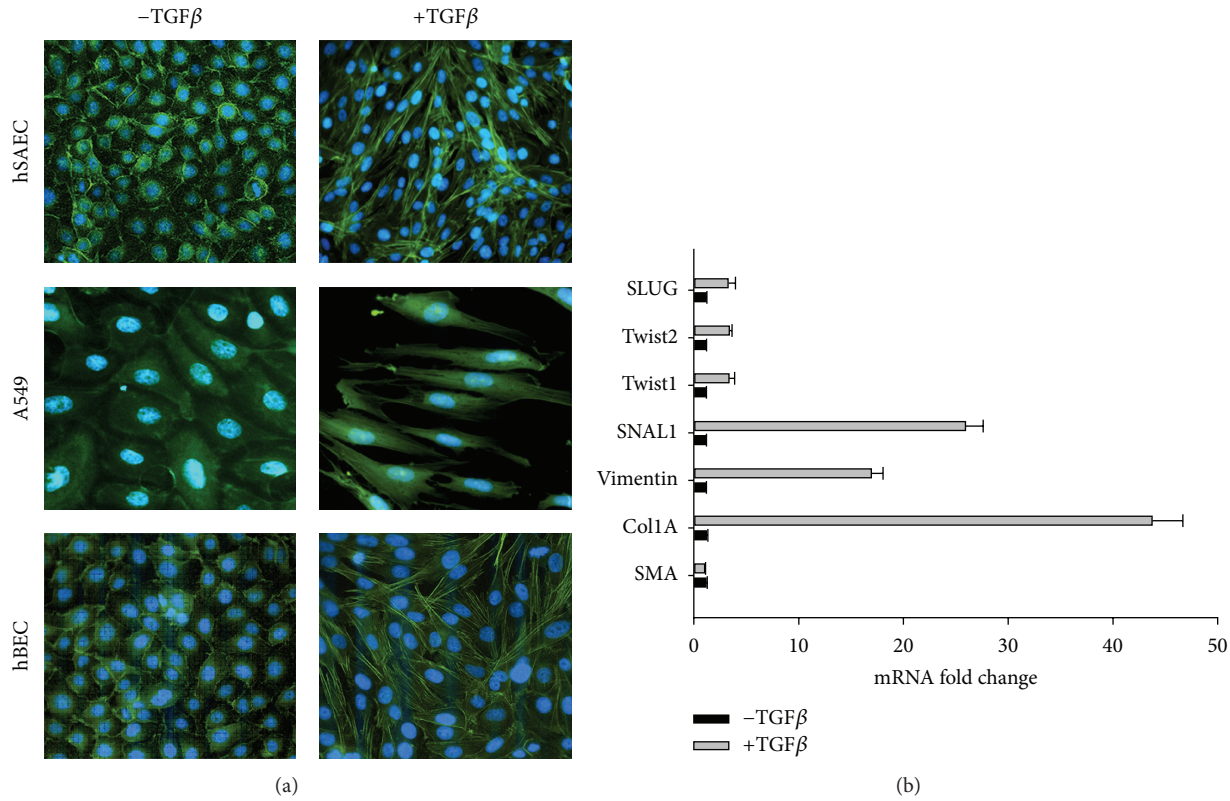


Figure 1: TGFβ induces EMT transition in HSAECs. (a) F-actin staining. Shown in green is a confocal microscopic image of FITC-phalloidin staining merged with DAPI staining (blue) of HSAECs, A549 cells, and HBECs in the absence or presence of TGFβ stimulation as indicated. (b) Expression of EMT program. Shown is normalized mRNA expression in HSAECs in the presence or absence of EMT after 10 d of TGFβ treatment. For each gene, mRNA expression was normalized to β-pol as a housekeeping gene and expressed as the fold change relative to its expression in the absence of TGFβ stimulation. Shown is mean and SE of replicate measurements. Each point is the mean of a duplicate biological experiment, measured with three technical replicates. Abbreviations: SLUG, Snail 2; SNAL1, Snail 1; col1A, collagen type 1a.

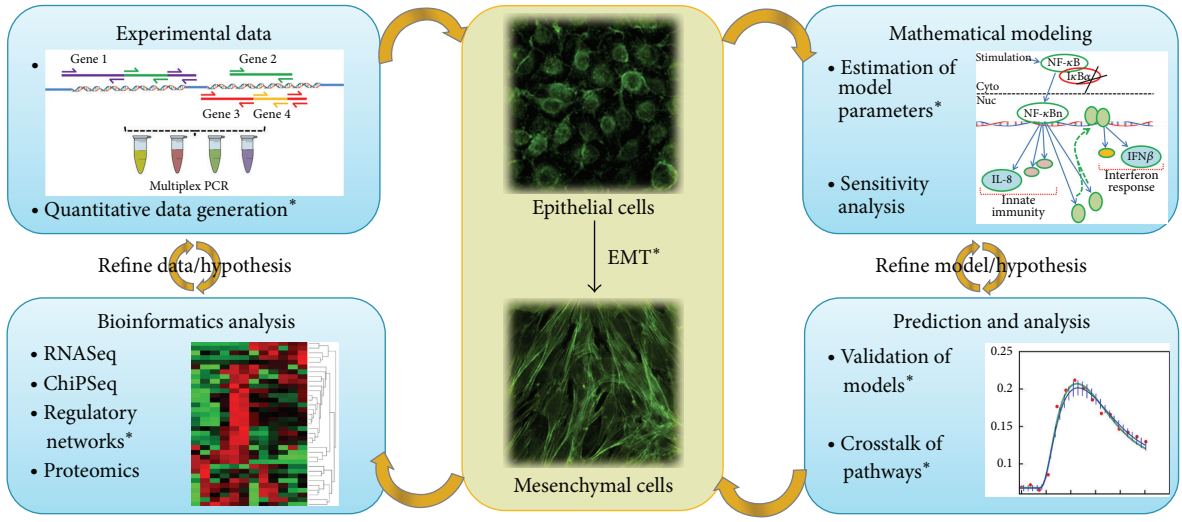


Figure 2: Systems approach to understanding EMT phenotypes. Schematic diagram of approach to understanding EMT phenotype using profiling and bioinformatics inferences. *indicates all the techniques/concepts that have been used in the current study, whereas others will be conducted in future studies.

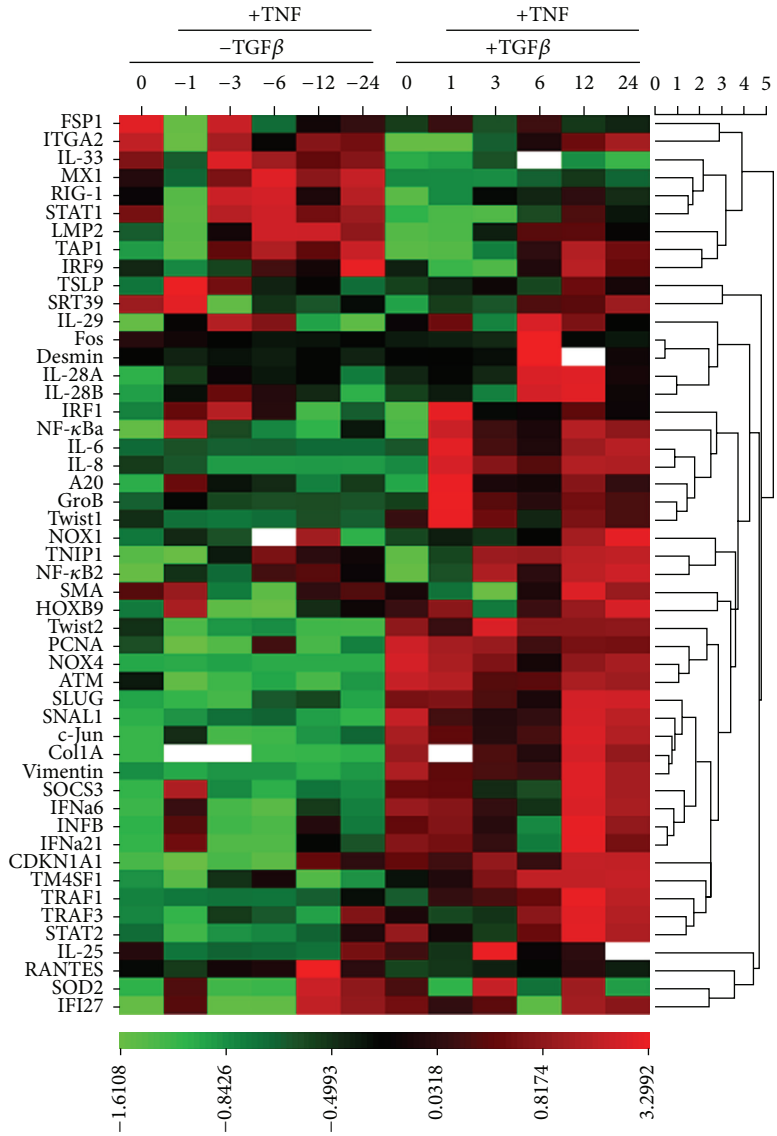


FIGURE 3: Effect of EMT on signaling pathways. (a) Heat map of gene expression. A time series of the TNFα stimulation of the HSAECs stimulated in the absence or presence of TGFβ-induced type II EMT. Data are Z-score transformed and relative to unstimulated normal HSAECs and expressed as standard deviation (SD) units from the row mean. Scale for SD deviation is shown at the bottom. Gene abbreviations are shown in Table 1. The white rectangles represent the missing data where the fold changes and Z-scores could not be calculated.

As discussed above, the noncanonical pathway is coupled to the canonical pathway through the expression of TNF receptor associated factor-1 (TRAF-1). TRAF-1 is unique for the TRAF isoforms that complexes with, stabilizes, and activates the NIK·IKKα complex to trigger the noncanonical pathway [24]. Analysis of relative fold change of expression showed that TRAF-1 expression was strongly upregulated within 1 h of TNFα stimulation and peaked at 12 h in EMT-HSAECs versus HSAECs, whereas, TRAF-1 was upregulated by 20-fold at 12 h after TNFα stimulation in EMT-HSEACs versus 3-fold in HSEACs (Figure 5(a)).

We especially noted that the noncanonical NF-κB pathway-dependent TNIP1/Naf1 gene showed a leftward shift in the temporal expression profile, where, in the presence of EMT, TNIP1/Naf1 expression was more rapidly induced, reaching the plateau within 3 h of stimulation in the EMT-HSAECs versus the 6 h required to reach plateau seen in HSAECs (Figure 5(a)). Together, these data indicate that EMT induces more robust expression of TRAF1 and produced a rapid coupling of the canonical to the noncanonical NF-κB pathway.

3.6. EMT Effects on Chromatin Modification and Transcriptional Elongation. Transcription of protein-coding genes by RNA Pol II is a highly regulated process involving preinitiation (assembly of basal transcription factors and co-activator recruitment), leading to initiation, elongation, and termination phases of RNA synthesis [40]. Previous work by our group and others has shown that transcriptional elongation is

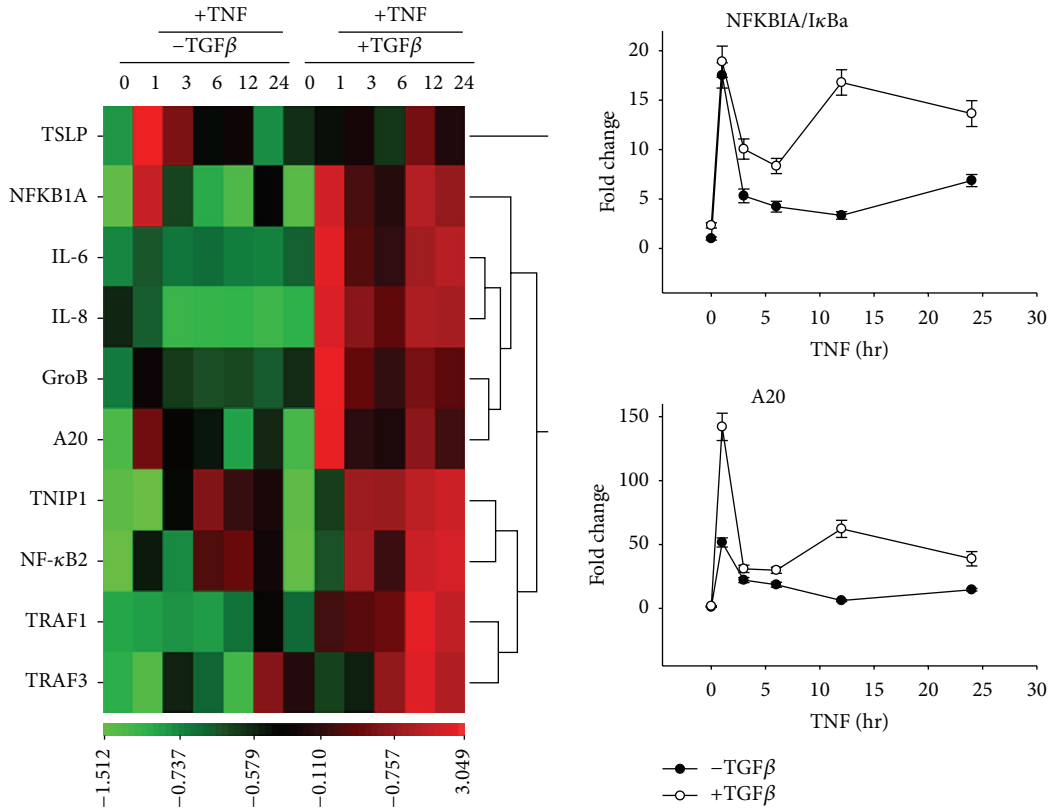


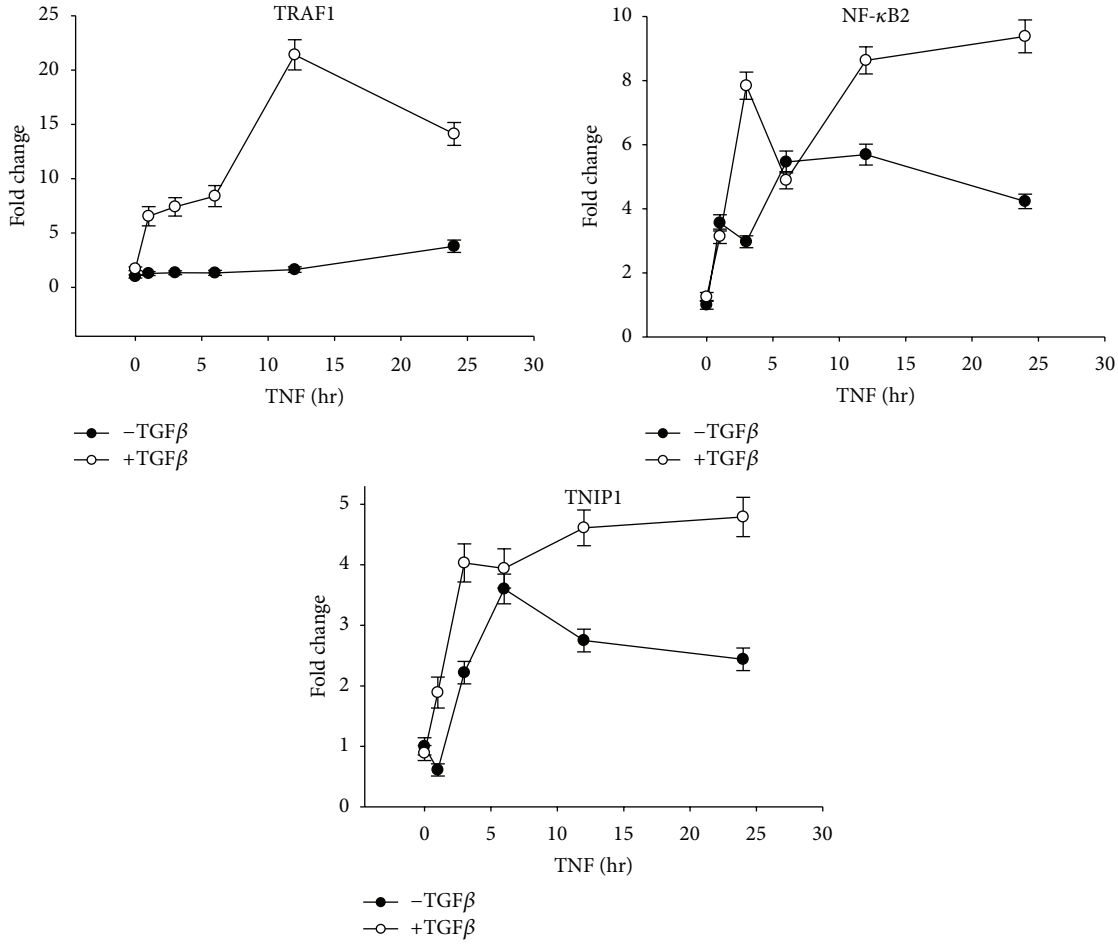
Figure 4: Dysregulation of canonical NF-κB signaling in EMT. Left panel, heat map of NF-κB selective reporter genes. Data are *Z*-score-transformed and relative to unstimulated normal HSAECs and expressed as standard deviation (SD) units from the row mean. Scale for SD deviation is shown at the bottom. Right panel, quantitation of NFKBIA/IκBa and TNFAIP3/A20 gene expressions in HSAECs and EMT-HSAECs as indicated. Shown is fold change mRNA relative to unstimulated normal HSAECs. Each point is the mean of a duplicate biological experiment, measured with three technical replicates.

a major regulated event controlling the innate response [38, 41, 42]. Because genome-wide analysis of EMT has revealed that TGFβ induces global reorganization of transcriptionally active marks (histone (H3) Lys (K) 36 trimethyl) in large organized heterochromatin domains [20], we examined whether the NF-κB-dependent genes were affected by chromatin remodeling or via changes in transcriptional elongation.

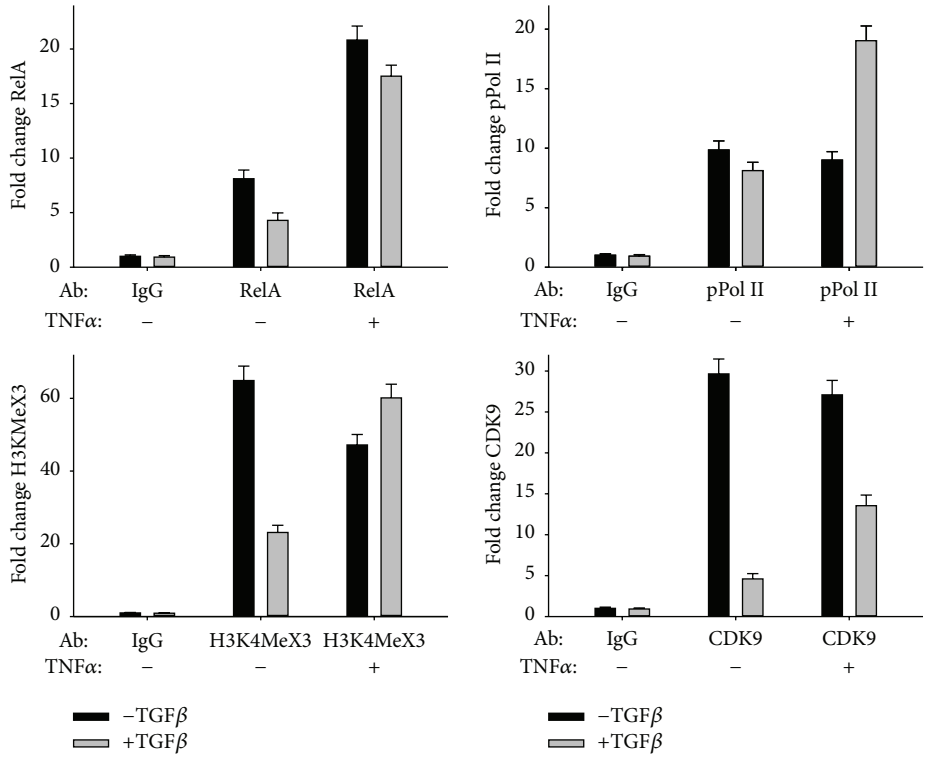
We therefore measured the binding of the transactivator (NFκB/RelA), transcriptional elongation complex (CDK9), activated transcriptional elongation polymerase (phospho-Ser^2 RNA Pol II), and euchromatin marks (H3K4Me3) on the TRAF-1 promoter using a highly quantitative two-step chromatin immunoprecipitation (XChIP) assay [34, 43]. Here, TNFα induced a similar level of increase in NF-κB/RelA binding to promoter in both EMT-HSAECs and normal HSAECs. This induction of NF-κB/RelA binding was not affected by EMT (Figure 5(b)). TNFα stimulation induced a significant increase in phospho-Ser^2 RNA Pol II binding in the EMT-HSAECs, despite reduced amounts of CDK9 binding in both basal and TNFα stimulated conditions (Figure 5(b)). EMT induced the level of H3K4Me3 binding to the TRAF1 promoter, even in the absence of TNFα stimulation. Together, these findings indicate that EMT promotes remodeling of the chromatin environment of the TRAF1 gene, poising it for more rapid induction via transcriptional elongation.

3.7. Computational Simulations of the Effect of EMT on Canonical-Noncanonical NF-κB Coupling. We previously reported a deterministic mathematical model linking the canonical and noncanonical arms of NF-κB pathway, where we have shown that the formation of the TRAF1·NIK complex acts as a rate-limiting feed-forward signaling complex necessary to activate the noncanonical pathway [24].

Based on the time delay between the expression of TRAF1 mRNA and TRAF1 protein and the presence of an atypical internal ribosomal binding site (IRES) and translational initiation site at an internal site in TRAF1 [44], we predicted that a delay in TRAF1 protein translation via a cap-independent mechanism controls the rate of noncanonical NF-κB pathway activation in response to TNFα. Interestingly, our expression profiling data here suggest that pathway coupling is affected by EMT in the HSAECs. Through the enhanced formation of transcriptional elongation-competent RNA Pol II, TRAF1 and NF-κB2 expression is enhanced, resulting in a shift of noncanonical pathway activation (TNIP1/Naf1) to earlier times.



(a)



(b)

Figure 5: Continued.

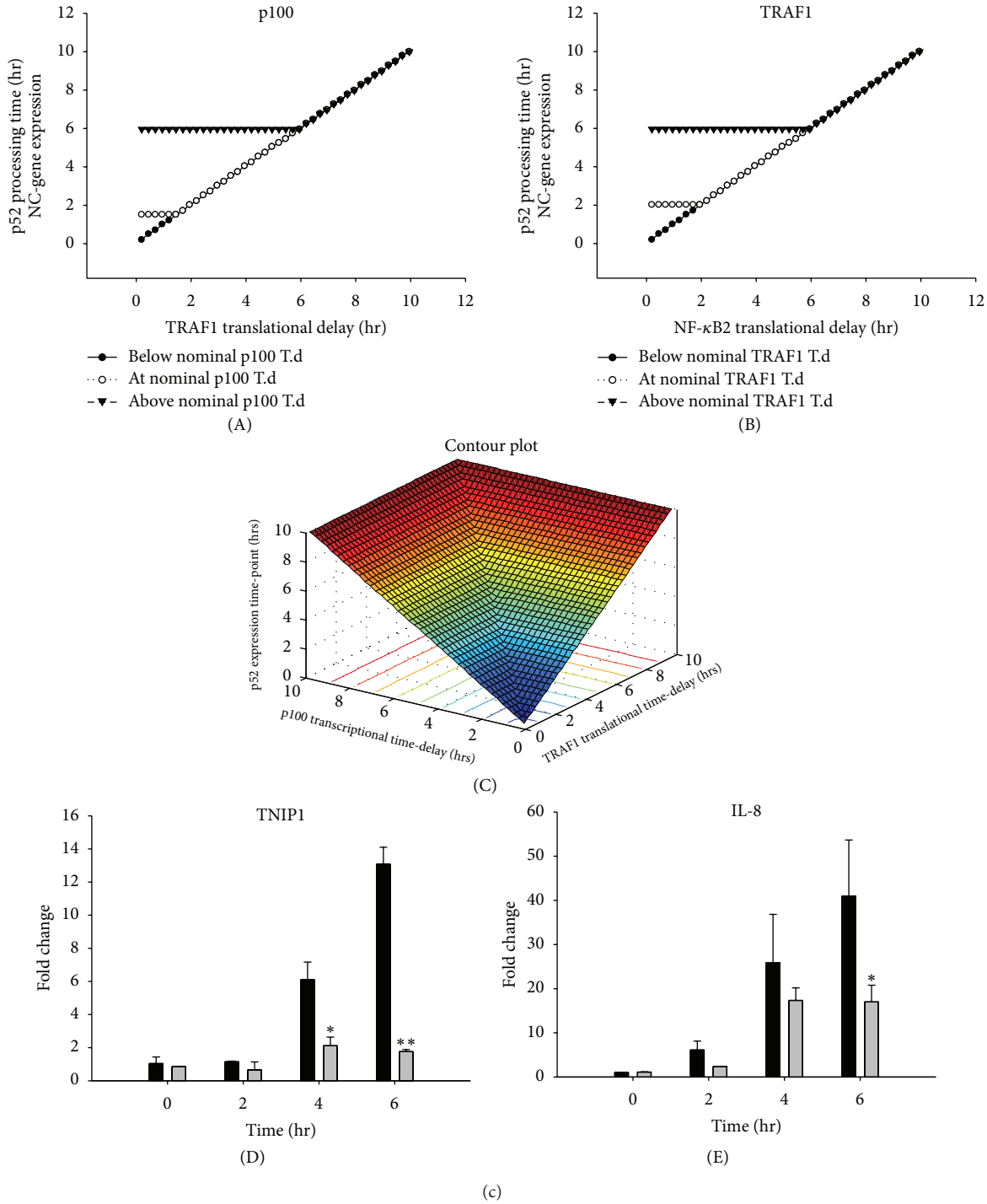


Figure 5: Effect of EMT on canonical-noncanonical NF-κB pathway coupling. (a) Relative changes in TRAF1, NF-κB2, and TNIP1 mRNA expressed as fold-change measurements in the absence or presence of TGFβ-induced EMT as indicated. Each point is the mean of a duplicate biological experiment, measured with three technical replicates. (b) XChIP experiments of HSAECs in the presence (grey bars) or absence (black bars) of TGFβ-induced EMT. Shown is fold change in the TRAF1 promoter quantified by Q-gPCR relative to unstimulated HSAEC signal in duplicate experiments. (c) Computational simulations of p52 processing as a function of translational delay for TRAF1 and NFκB2. Abbreviations; T.d., translational delay. (A) shows the effect of increasing TRAF1 translational delay on p52 processing time while keeping the translational delay of NFκB2 either at nominal rate (90′) or higher than nominal rate or lower than nominal rate. (B) shows similar effect but for increasing NFκB2 translational delay (x-axis). (C) shows the contour plot of all simulations (D, E) amantadine-treated A549 cells (200 μg/mL) were stimulated with poly I:C and TNIP1/Naf1 (D), and IL8 (E) expression was measured by Q-RT-PCR. Data expressed as fold change as compared to untreated cells after normalizing to internal controls, GAPDH. Data analyzed by a 2-way ANOVA with multiple comparisons. Significantly different from amantadine untreated samples: $^{*}P < 0.05$ and $^{**}P < 0.001$. Amantadine-treated cells (light bars) showed higher level of noncanonical pathway inhibition compared to untreated cells (dark bars).

We, therefore, performed simulations of the noncanonical pathway coupling by systematically varying the translation delay function of TRAF1. Interestingly, the coupling interval, that is, time lag between the activation of canonical NF-κB pathway and the formation of p52, monotonically reduced as TRAF-1 translation was increased until it reached a critical threshold of 90 min (Figure 5(c): (A)). Further reductions in the TRAF-1 translation delay function alone, even to zero, did not reduce the coupling interval for p52 formation and TNIP/Naf1 expression.

After examination of other proteins in the noncanonical pathway, we noted that NF-κB2 translation delay function produced a similar effect, where the coupling interval reached a plateau below which p52 processing was not further increased (Figure 5(c): (B)). In these experiments, a total of 42 simulations were conducted for each protein covering the time points from zero to 10 h.

The finding that the coupling interval was inert to reducing the translational delay of each protein individually prompted us to perform a simulation by changing the translational delay parameters for both TRAF1 and NF-κB2 simultaneously. The biological rationale for this is that TRAF-1 is translated by a cap-independent mechanism [44] and our previous studies show that both of these mRNAs are translated in a coordinated, delayed manner distinct from those mRNAs undergoing cap-dependent translation. A total of 1681 combinatorial simulations of the translational delay were conducted covering the time points from 0 to 10 h. Interestingly, we found that reducing the IRES-dependent translational delays for both proteins together resulted in shorter coupling interval (e.g., an earlier onset of p52 processing and noncanonical gene expression) to values as quickly as 15 min relative to TNFα stimulation (versus the nominal time of 6 h) (Figure 5(c): (A), (B)). In both figures, reducing the translational delay of both proteins to zero resulted in the early processing of p100 to p52. Figure 5(c): (C) shows the cumulative effect of all combinations of simulations for these two proteins. A conceptual representation is shown in Table 2. This model thus predicts that altering the translation rate of these two proteins produces a leftward shift in noncanonical gene expression kinetics relative to the canonical pathway.

Our computational model predicted that TRAF1 and p100 translation is under IRES-dependent control. Our earlier studies have shown that the activation kinetics of noncanonical pathway is coincident with that of the canonical pathway in response to viral infection and dsRNA stimulation [45]. We noted that the effects of dsRNA are similar to that of EMT by reducing the coupling constant between the canonical and noncanonical pathways. To confirm whether the noncanonical pathway was coupled to NF-κB/RelA through an IRES dependent mechanism, we perturbed an IRES mediated translation using amantadine, a known inhibitor of IRES-dependent translation [46]. A549 cells were activated using poly I:C treatment in the absence or presence of amantadine, and canonical and noncanonical pathway kinetics were measured. We observed that the amantadine-treated cells showed significant inhibition of TNIP1/Naf1 expression compared to control cells (no amantadine treatment) after 4 and 6 h of treatment (Figure 5(c): (D)). At this dose, amantadine seems to have some inhibitory effect on IL8 expression at 6 hr (Figure 5(c): (E)) although this effect was significantly lower than its effect on TNIP1/Naf1 (7-fold versus 2-fold) (Figure 5(c): (D), (E)). These results suggest that EMT results in chromatin remodeling and transcriptional elongation of TRAF-1/NFκB2 expression as well as influencing the rate of translation of IRES-dependent genes, resulting in shortening the coupling interval between the canonical and noncanonical pathways (Figure 6).

Table 2: Conceptual representation of combinatorial simulations of translational delays of TRAF1 and NF-κB2.

Time-delay combinations (min)	NF-κB2/p100 processing kinetics
(1) TRAF1 (<120) + P100 (<90)	Fast
(2) TRAF1 (≤120) + P100 (=90) OR TRAF1 (=120) + P100 (≤90)	Near nominal or nominal
(3) TRAF1 (≥120) + P100 (>90) OR TRAF1 (>120) + P100 (≤90) OR TRAF1 (≤120) + P100 (>90)	Delayed

4. Discussion

In this study, we applied systems level approaches to interrogate the dynamic state of EMT in a model of chronic mucosal inflammation. Although EMT processes (type I EMT) are essential during histogenesis and organogenesis, type II EMT is involved in organ remodeling and chronic diseases, and type III EMT is involved in progression of transformed epithelial cells to a metastatic phenotype. These newly formed mesenchymal cells transiently express distinct markers acquiring "front-rear" polarity, become invasive, and induce extracellular matrix remodeling. Using immortalized human small airway epithelial cells, we demonstrate that TGFβ induces stable EMT transition morphologically through the restructuring of actin cytoskeleton, induction of intermediate mesenchymal cytoskeletal proteins, and expression of the SNALI/Twist transcription factors. Although the signals conferring EMT have been extensively investigated, the effect of EMT on signal transduction pathways has not been systematically explored. Here, we investigated the effect of EMT on innate signaling and discovered that EMT dramatically affects two coupled arms of the NF-κB signaling pathway. This coupling is mediated through the induction of transcriptionally active euchromatin marks and enhancement of transcriptional elongation. Our computational simulations further predict that translational rate of these two key rate-limiting coupling proteins must also be affected. As a result, the innate pathway is hyperresponsive in type II EMT, with a more rapid coupling between the canonical and noncanonical arms of the NF-κB pathway. Because EMT has been observed in chronic lung fibrotic disease in humans [47] and epithelial cells isolated from human asthmatics undergo

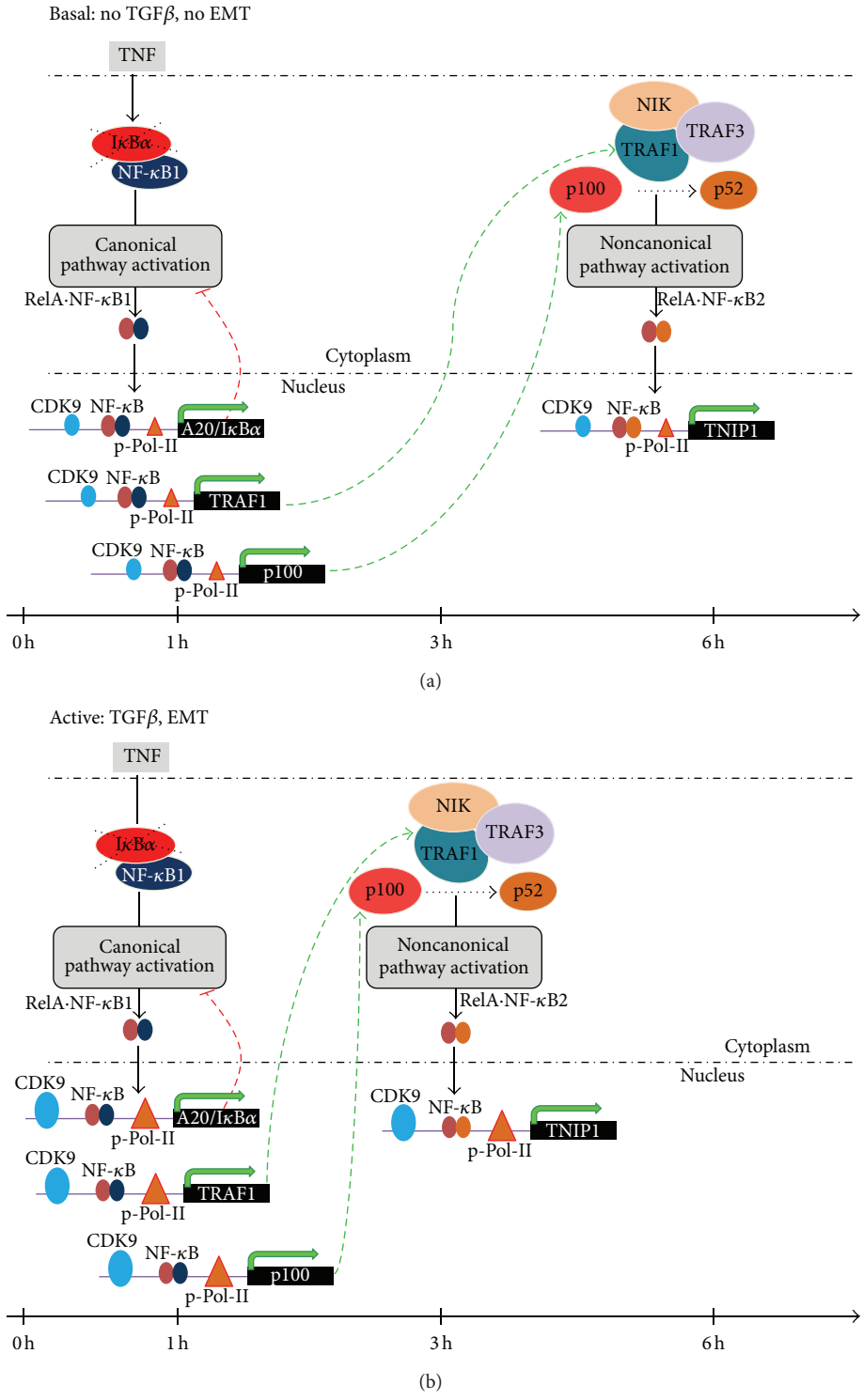


FIGURE 6: Effect of EMT on canonical-noncanonical NF-κB coupling. Schematic diagram of EMT effects on NF-κB signaling pathway. EMT induces changes in euchromatin marks of NF-κB-dependent genes, transcriptional elongation, and translational delay to reduce the coupling constant between the NF-κB canonical and noncanonical pathways. A timeline is shown for both states of EMT-Basal (no EMT) and TGFβ-mediated EMT (0 to 6 h).

a greater degree of EMT than normal subjects [13], these studies have special relevance to airway pathophysiology.

Type II EMT may play a critical cellular role in the progression of human pulmonary fibrotic diseases. Although the molecular basis of airway remodeling and loss of epithelial integrity in asthma is still undefined, several lines of evidence implicate EMT in this process [48]. For example, the airway epithelium is recognized as an important contributor to intrapulmonary fibroblast accumulation after injury, including idiopathic pulmonary fibrosis [12, 49], asthma, and COPD [50]. The bronchial epithelium in asthma is highly abnormal with structural changes involving the separation of columnar cells from their basal attachment leading to the disruption of epithelial barrier and functional changes including increased expression and release of profibrotic factors [2]. Subepithelial fibrosis is influenced by epithelial cell-derived VEGF, IL-33, IL-25, and thymic stromal lymphopoietin (TSLP), as well as leukocyte-derived TGFβ. Epithelial cells from individuals with asthma produce high amounts of periostin, a cytokine that stimulates TGF-β production and modifies myofibroblast collagen synthesis [51]. The increase in myofibroblasts causes thickening and increased density of the subepithelial basement membrane, contributing to a progressive decline in lung function [48, 52].

The relationship between epithelial fibrosis and innate inflammation is a complex and interrelated phenomenon. Specifically, the TNFα-NF-κB pathway has a number of modulatory actions on epithelial EMT. For example, activated NF-κB suppresses E-Cad and induces the expression of mesenchymal vimentin, two hallmarks of EMT [53]. Studies of insulin breast cancer epithelial cells have shown that like growth factor (IGF)1 receptor-induced EMT is partially mediated by NF-κB-increased expression of SNAL1 [54]. SNAI1 upregulation is mediated at multiple levels, including direct binding to the SNAI1 promoter leading to an increased mRNA expression at the level of transcription [55] and secondly by induction of CSN2, a protein that disrupts binding GSK-3β and β-Trcp binding to SNAI1 resulting in its posttranslational stabilization [56]. Our observations suggest that EMT-HSAECs express high levels of SNAI1, and that upregulation of NF-κB is partially responsible for mediating type II EMT. More work will be required to determine the requirement of NF-κB signaling in EMT of HSAECs. We note that other studies have shown that NF-κB activates ZEB in mammary epithelial cells through cis-regulatory elements in the ZEB promoter [53], perhaps suggesting that NF-κB is a master regulator of EMT programs.

An exciting and novel finding in our studies is that EMT sensitized the innate pathway to be hyperresponsive to its canonical activating signals. Specifically, we demonstrate that EMT dramatically upregulates innate genes and accelerates the coordinated induction of TRAF1 and NF-κB2 expression by modulating euchromatin marks and inducing transcriptional elongation. Our XChIP experiments indicate that genes downstream of the NF-κB pathway are regulated by a mechanism involving enhanced formation of phospho-Ser^2 Pol II. The NF-κB pathway is a major arm of the innate immune response, whose rapid activation generates effectors that restrict pathogen spread [23, 57]. For innate genes bounded within closed chromatin, preinitiation complex formation is an essential first step in inducible expression [40]. In this process, the histone acetyltransferases, such as p300/CBP, are recruited to destabilize repressive histone, resulting in p300/CBP dissociation the and binding of general transcription factors, including TFIID. By contrast, for immediate early genes located within open chromatin, Pol II is preengaged in a hypophosphorylated state, producing short ~30–50 nt transcripts [58]. Promoter proximal pausing is reversed by the activated positive transcription elongation factor (P-TEFb), a multiprotein complex containing CDK9, BRD4 and cyclin T1 or T2 subunits [41, 42]. Upon PTEF-b binding, phospho-Ser^2 Pol II is formed, and enters elongation mode to produce full-length, spliced transcripts [41, 42, 59]. Our XChIP analysis demonstrates that innate genes in EMT-HSAECs are located within an euchromatin environment, associated with transcriptionally active histone H3 marks. This finding is consistent with RelA binding preferring open chromatin domains, inferred by our earlier ChIP-Seq studies [43]. The effect of EMT on enhanced formation of phospho-Ser^2 RNA Pol II has not been previously observed and will require further investigation into the effects of EMT on the PTEF-b protein interaction network.

Our data further indicate that EMT affects the feed-forward link between the canonical and noncanonical NF-κB pathways. Earlier, we showed that TRAF1 and NF-κB2 are canonical genes whose protein synthesis is under a translational delay [24]. Newly synthesized TRAF1 binds and stabilizes NIK by disrupting its interaction with TRAF2-cIAP2. Activated NIK is the rate limiting step responsible for the co-translational processing of NF-κB2 precursor into the active 50 kDa DNA binding form [24]. Our simulations predict that EMT controls the rate of IRES-dependent translation of both TRAF1 and NF-κB2 to explain this effect. The amantadine inhibition experiments validate these predictions; however, more research will be required to understand the effect of EMT on IRES-dependent translational control.

We contend that the systematic examination of type II EMT of airway epithelial cells has important implications for human disease. Recent studies suggest that the structural components of the lung may first respond to environmental risk factors for inflammatory/fibrotic lung diseases, such as asthma [60], COPD, and cancer. In asthma, EMT remodeling of the epithelial basement membrane promotes sensitization to inhaled allergen by causing persistent dendritic cell activation and migration [61]. In COPD, the reticular basement membrane is fragmented with "clefts" of cells staining for MMP-9 and S100A4, hallmarks of EMT [13]. We note that airways infection with respiratory syncytial virus (RSV), the most common viral respiratory pathogen in small children which is widely considered as a risk factor for the development of allergic asthma later in life is characterized by an increased expression of SNAI1, MMP-2, and TGFβ1 [62, 63]. Finally, EMT plays a critical role in switching a primary tumor to a malignant cancer with metastatic phenotype. These data suggest that induction of EMT is seen in a number of lung pathologies.

In summary, using RNA profiling of an EMT cell model, we demonstrate that this transition has dramatic effects on the induction of the innate pathway through a mechanism involving chromatin remodeling of the TRAF1 locus, enabling enhanced phospho-Pol II loading and transcriptional elongation. Computational simulation experiments suggest that rapid, coordinate, TRAF1/NFκB2 expression is coupled to their IRES-dependent translation. These rapidly expressed proteins produce a feed-forward mechanism that reduces the coupling interval between the canonical and noncanonical signaling arms of the NF-κB pathways. Further applications of systems approaches will promote a comprehensive understanding of this complex phenotype through reiterated rounds of gene perturbation, multidimensional profiling, and deterministic modeling.

Acknowledgments

This work was supported, in part, by the Sealy Center for Molecular Medicine, National Institute of Health Grants PO1 AI062885 (ARB, IB), NHLBI contract HHSN268201000037C (ARB, IB, KP) UL1TR00007, Clinical and Translational Science Award (ARB), GM086885(ARB), and NCI P50CA70907 (SPORE in Lung Cancer to JDM). The authors thank Rolf Konig, PhD, and Konrad Pazdrak, MD, PhD, for helpful discussions and input.

References

[1] A. J. Polito and D. Proud, "Epithelial cells as regulators of airway inflammation," *Journal of Allergy and Clinical Immunology*, vol. 102, no. 5, pp. 714–718, 1998.

[2] B. N. Lambrecht and H. Hammad, "The airway epithelium in asthma," *Nature Medicine*, vol. 18, pp. 684–692, 2012.

[3] B. C. Willis, J. M. Liebler, K. Luby-Phelps et al., "Induction of epithelial-mesenchymal transition in alveolar epithelial cells by transforming growth factor-β1: potential role in idiopathic pulmonary fibrosis," *The American Journal of Pathology*, vol. 166, no. 5, pp. 1321–1332, 2005.

[4] C. Ward, I. A. Forrest, D. M. Murphy et al., "Phenotype of airway epithelial cells suggests epithelial to mesenchymal cell transition in clinically stable lung transplant recipients," *Thorax*, vol. 60, no. 10, pp. 865–871, 2005.

[5] S. G. Royce, L. Tan, A. A. Koek, and M. L. K. Tang, "Effect of extracellular matrix composition on airway epithelial cell and fibroblast structure: implications for airway remodeling in asthma," *Annals of Allergy, Asthma and Immunology*, vol. 102, no. 3, pp. 238–246, 2009.

[6] D. E. Davies and S. T. Holgate, "Asthma: the importance of epithelial mesenchymal communication in pathogenesis: inflammation and the airway epithelium in asthma," *International Journal of Biochemistry and Cell Biology*, vol. 34, no. 12, pp. 1520–1526, 2002.

[7] S. T. Holgate, J. Holloway, S. Wilson, F. Bucchieri, S. Puddicombe, and D. E. Davies, "Epithelial-mesenchymal communication in the pathogenesis of chronic asthma," *Proceedings of the American Thoracic Society*, vol. 1, no. 2, pp. 93–98, 2004.

[8] R. Y. Huang, P. Guilford, and J. P. Thiery, "Early events in cell adhesion and polarity during epithelial-mesenchymal transition," *Journal of Cell Science*, vol. 125, pp. 4417–4422, 2012.

[9] J. Lim and J. P. Thiery, "Epithelial-mesenchymal transitions: insights from development," *Development*, vol. 139, pp. 3471–3486, 2012.

[10] B. C. Willis and Z. Borok, "TGF-β-induced EMT: mechanisms and implications for fibrotic lung disease," *American Journal of Physiology*, vol. 293, no. 3, pp. L525–L534, 2007.

[11] R. Jain, P. W. Shaul, Z. Borok, and B. C. Willis, "Endothelin-1 induces alveolar epithelial-mesenchymal transition through endothelin type A receptor-mediated production of TGF-β1," *American Journal of Respiratory Cell and Molecular Biology*, vol. 37, no. 1, pp. 38–47, 2007.

[12] B. C. Willis, R. M. DuBois, and Z. Borok, "Epithelial origin of myofibroblasts during fibrosis in the lung," *Proceedings of the American Thoracic Society*, vol. 3, no. 4, pp. 377–382, 2006.

[13] T.-L. Hackett, S. M. Warner, D. Stefanowicz et al., "Induction of epithelial-mesenchymal transition in primary airway epithelial cells from patients with asthma by transforming growth factor-β1," *American Journal of Respiratory and Critical Care Medicine*, vol. 180, no. 2, pp. 122–133, 2009.

[14] S. S. Sohal, D. Reid, A. Soltani et al., "Evaluation of epithelial mesenchymal transition in patients with chronic obstructive pulmonary disease," *Respiratory Research*, vol. 12, p. 130, 2011.

[15] I. Y. Adamson, L. Young, and D. H. Bowden, "Relationship of alveolar epithelial injury and repair to the induction of pulmonary fibrosis," *The American Journal of Pathology*, vol. 130, no. 2, pp. 377–383, 1988.

[16] S. Cannito, E. Novo, A. Compagnone et al., "Redox mechanisms switch on hypoxia-dependent epithelial-mesenchymal transition in cancer cells," *Carcinogenesis*, vol. 29, no. 12, pp. 2267–2278, 2008.

[17] T. Vincent, E. P. Neve, J. R. Johnson et al., "A SNAIL1-SMAD3/4 transcriptional repressor complex promotes TGF-β mediated epithelial-mesenchymal transition," *Nature Cell Biology*, vol. 11, no. 8, pp. 943–950, 2009.

[18] E. M. Minshall, D. Y. M. Leung, R. J. Martin et al., "Eosinophil-associated TGF-β1 mRNA Expression and airways fibrosis in bronchial asthma," *American Journal of Respiratory Cell and Molecular Biology*, vol. 17, no. 3, pp. 326–333, 1997.

[19] J. Yang and R. A. Weinberg, "Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis," *Developmental Cell*, vol. 14, no. 6, pp. 818–829, 2008.

[20] O. G. McDonald, H. Wu, W. Timp, A. Doi, and A. P. Feinberg, "Genome-scale epigenetic reprogramming during epithelial-to-mesenchymal transition," *Nature Structural and Molecular Biology*, vol. 18, no. 8, pp. 867–874, 2011.

[21] R. P. Garofalo and H. Haeberle, "Epithelial regulation of innate immunity to respiratory syncytial virus," *American Journal of Respiratory Cell and Molecular Biology*, vol. 23, no. 5, pp. 581–585, 2000.

[22] S. Akira, S. Uematsu, and O. Takeuchi, "Pathogen recognition and innate immunity," *Cell*, vol. 124, no. 4, pp. 783–801, 2006.

[23] A. R. Brasier, "The NF-k B signaling network: insights from systems approaches," in *Cellular Signaling and Innate Immune Responses to RNA Virus Infections*, A. R. Brasier, S. M. Lemon, and A. Garcia-Sastre, Eds., pp. 119–135, American Society for Microbiology, 2008.

[24] S. Choudhary, M. Kalita, L. Fang et al., "Inducible TNF receptor associated factor 1 expression couples the canonical to the non-canonical NF-kB pathway in TNF stimulation," *The Journal of Biological Chemistry*, vol. 288, pp. 14612–14623, 2013.

[25] A. R. Brasier, "The NF-kappaB regulatory network," *Cardiovascular Toxicology*, vol. 6, no. 2, pp. 111–130, 2006.

[26] B. Tian, D. E. Nowak, M. Jamaluddin, S. Wang, and A. R. Brasier, "Identification of direct genomic targets downstream of the nuclear factor-κB transcription factor mediating tumor necrosis factor signaling," *The Journal of Biological Chemistry*, vol. 280, no. 17, pp. 17435–17448, 2005.

[27] B. Tian, D. E. Nowak, and A. R. Brasier, "A TNF-induced gene expression program under oscillatory NF-κB control," *BMC Genomics*, vol. 6, article 137, 2005.

[28] B. Tian and A. R. Brasier, "Identification of a nuclear factor KappaB-dependent gene network," *Recent Progress in Hormone Research*, vol. 58, pp. 95–130, 2003.

[29] S. Thomson, F. Petti, I. Sujka-Kwok et al., "A systems view of epithelial-mesenchymal transition signaling states," *Clinical and Experimental Metastasis*, vol. 28, no. 2, pp. 137–155, 2011.

[30] A. Schroeder, O. Mueller, S. Stocker et al., "The RIN: an RNA integrity number for assigning integrity values to RNA measurements," *BMC Molecular Biology*, vol. 7, article 3, 2006.

[31] W. J. Kent, "BLAT—the BLAST-like alignment tool," *Genome Research*, vol. 12, no. 4, pp. 656–664, 2002.

[32] K. J. Livak and T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method," *Methods*, vol. 25, no. 4, pp. 402–408, 2001.

[33] R. D. Ramirez, S. Sheridan, L. Girard et al., "Immortalization of human bronchial epithelial cells in the absence of viral oncoproteins," *Cancer Research*, vol. 64, no. 24, pp. 9027–9034, 2004.

[34] B. Tian, J. Yang, and A. R. Brasier, "Two-step cross-linking for analysis of protein-chromatin interactions," *Methods in Molecular Biology*, vol. 809, pp. 105–120, 2012.

[35] D. E. Nowak, B. Tian, and A. R. Brasier, "Two-step cross-linking method for identification of NF-κB gene network by chromatin immunoprecipitation," *BioTechniques*, vol. 39, no. 5, pp. 715–724, 2005.

[36] S. Basak, H. Kim, J. D. Kearns et al., "A fourth IκB protein within the NF-κB signaling module," *Cell*, vol. 128, no. 2, pp. 369–381, 2007.

[37] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker, "Analysis of microarray data using Z score transformation," *Journal of Molecular Diagnostics*, vol. 5, no. 2, pp. 73–81, 2003.

[38] D. E. Nowak, B. Tian, M. Jamaluddin et al., "RelA Ser276 phosphorylation is required for activation of a subset of NF-κB-dependent genes by recruiting cyclin-dependent kinase 9/cyclin t1 complexes," *Molecular and Cellular Biology*, vol. 28, no. 11, pp. 3623–3638, 2008.

[39] D. E. Nelson, A. E. C. Ihekwaba, M. Elliott et al., "Oscillations in NF-κB signaling control the dynamics of gene expression," *Science*, vol. 306, no. 5696, pp. 704–708, 2004.

[40] K. Chiba, J. Yamamoto, Y. Yamaguchi, and H. Handa, "Promoter-proximal pausing and its release: molecular mechanisms and physiological functions," *Experimental Cell Research*, vol. 316, no. 17, pp. 2723–2730, 2010.

[41] A. R. Brasier, B. Tian, M. Jamaluddin, M. K. Kalita, R. P. Garofalo, and M. Lu, "RelA Ser276 phosphorylation-coupled Lys310 acetylation controls transcriptional elongation of inflammatory cytokines in respiratory syncytial virus infection," *Journal of Virology*, vol. 85, no. 22, pp. 11752–11769, 2011.

[42] B. Tian, Y. Zhao, M. Kalita et al., "CDK9-dependent transcriptional elongation in the innate ISG response to RSV infection in airway epithelial cells," *Journal of Virology*, vol. 87, no. 12, pp. 7075–7092, 2013.

[43] J. Yang, A. Mitra, N. Dojer, S. Fu, M. Rowicka, and A. R. Brasier, "A probabilistic approach to learn chromatin architecture and accurate inference of the NF-kappaB/RelA regulatory network using ChIP-Seq," *Nucleic Acids Research*, vol. 41, no. 15, pp. 7240–7259, 2013.

[44] L. Yang, L. Gu, Z. Li, and M. Zhou, "Translation of TRAF1 is regulated by IRES-dependent mechanism and stimulated by vincristine," *Nucleic Acids Research*, vol. 38, no. 13, pp. 4503–4513, 2010.

[45] S. Choudhary, S. Boldogh, R. Garofalo, M. Jamaluddin, and A. R. Brasier, "Respiratory syncytial virus influences NF-κB-dependent gene expression through a novel pathway involving MAP3K14/NIK expression and nuclear complex formation with NF-κB2," *Journal of Virology*, vol. 79, no. 14, pp. 8948–8959, 2005.

[46] Y.-J. Chen, S.-J. Zeng, J. T. Hsu et al., "Amantadine as a regulator of internal ribosome entry site," *Acta Pharmacologica Sinica*, vol. 29, no. 11, pp. 1327–1333, 2008.

[47] K. K. Kim, M. C. Kugler, P. J. Wolters et al., "Alveolar epithelial cell mesenchymal transition develops in vivo during pulmonary fibrosis and is regulated by the extracellular matrix," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 35, pp. 13180–13185, 2006.

[48] Y. C. Yang, N. Zhang, K. van Crombruggen, G. H. Hu, S. L. Hong, and C. Bachert, "Transforming growth factor-beta1 in inflammatory airway disease: a key for understanding inflammation and remodeling," *Allergy*, vol. 67, no. 10, pp. 1193–1202, 2012.

[49] H. Kasai, J. T. Allen, R. M. Mason, T. Kamimura, and Z. Zhang, "TGF-β1 induces human alveolar epithelial to mesenchymal cell transition (EMT)," *Respiratory Research*, vol. 6, article 56, 2005.

[50] C. G. Lee, B. Ma, S. Takyar et al., "Studies of vascular endothelial growth factor in asthma and chronic obstructive pulmonary disease," *Proceedings of the American Thoracic Society*, vol. 8, no. 6, pp. 512–515, 2011.

[51] S. S. Sidhu, S. Yuan, A. L. Innes et al., "Roles of epithelial cell-derived periostin in TGF-β activation, collagen production, and collagen gel elasticity in asthma," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 32, pp. 14170–14175, 2010.

[52] S. Al-Muhsen, J. R. Johnson, and Q. Hamid, "Remodeling in asthma," *Journal of Allergy and Clinical Immunology*, vol. 128, no. 3, pp. 451–462, 2011.

[53] H. L. Chua, P. Bhat-Nakshatri, S. E. Clare, A. Morimiya, S. Badve, and H. Nakshatri, "NF-κB represses E-cadherin expression and enhances epithelial to mesenchymal transition of mammary epithelial cells: potential involvement of ZEB-1 and ZEB-2," *Oncogene*, vol. 26, no. 5, pp. 711–724, 2007.

[54] H. J. Kim, B. C. Litzenburger, X. Cui et al., "Constitutively active type I insulin-like growth factor receptor causes transformation and xenograft growth of immortalized mammary epithelial cells and is accompanied by an epithelial-to-mesenchymal transition mediated by NF-κB and snail," *Molecular and Cellular Biology*, vol. 27, no. 8, pp. 3165–3175, 2007.

[55] M. J. Barberà, I. Puig, D. Domínguez et al., "Regulation of Snail transcription during epithelial to mesenchymal transition of tumor cells," *Oncogene*, vol. 23, no. 44, pp. 7345–7354, 2004.

[56] Y. Wu, J. Deng, P. G. Rychahou, S. Qiu, B. M. Evers, and B. P. Zhou, "Stabilization of Snail by NF-κB Is Required for inflammation-induced cell migration and invasion," *Cancer Cell*, vol. 15, no. 5, pp. 416–428, 2009.

[57] A. R. Brasier, A. Garcia-Sastre, and S. M. Lemon, *Cellular Signaling and Innate Immune Response to RNA Virus Infections*, ASM Press, Washington, DC, USA, 2008.

[58] R. J. Sims III, R. Belotserkovskaya, and D. Reinberg, "Elongation by RNA polymerase II: the short and long of it," *Genes and Development*, vol. 18, no. 20, pp. 2437–2468, 2004.

[59] B. M. Peterlin and D. H. Price, "Controlling the elongation phase of transcription with P-TEFb," *Molecular Cell*, vol. 23, no. 3, pp. 297–305, 2006.

[60] S. Tourdot, S. Mathie, T. Hussell et al., "Respiratory syncytial virus infection provokes airway remodelling in allergen-exposed mice in absence of prior allergen sensitization," *Clinical and Experimental Allergy*, vol. 38, no. 6, pp. 1016–1024, 2008.

[61] L. S. van Rijt, N. Vos, M. Willart et al., "Persistent activation of dendritic cells after resolution of allergic airway inflammation breaks tolerance to inhaled allergens in mice," *American Journal of Respiratory and Critical Care Medicine*, vol. 184, no. 3, pp. 303–311, 2011.

[62] E. Kaltenborn, S. Kern, S. Frixel et al., "Respiratory syncytial virus potentiates ABCA3 mutation-induced loss of lung epithelial cell differentiation," *Human Molecular Genetics*, vol. 21, pp. 2793–2806, 2012.

[63] C. Palena, D. E. Polev, K. Y. Tsang et al., "The human T-box mesodermal transcription factor Brachyury is a candidate target for T-cell—mediated cancer immunotherapy," *Clinical Cancer Research*, vol. 13, no. 8, pp. 2471–2478, 2007.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 512086, 14 pages
http://dx.doi.org/10.1155/2013/512086

Research Article

Systems Approaches Evaluating the Perturbation of Xenobiotic Metabolism in Response to Cigarette Smoke Exposure in Nasal and Bronchial Tissues

Anita R. Iskandar, Florian Martin, Marja Talikka, Walter K. Schlage, Radina Kostadinova, Carole Mathis, Julia Hoeng, and Manuel C. Peitsch

Philip Morris International R&D, Philip Morris Products SA, Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland

Correspondence should be addressed to Julia Hoeng; julia.hoeng@pmi.com

Received 17 June 2013; Revised 14 August 2013; Accepted 16 August 2013

Academic Editor: Tao Huang

Capturing the effects of exposure in a specific target organ is a major challenge in risk assessment. Exposure to cigarette smoke (CS) implicates the field of tissue injury in the lung as well as nasal and airway epithelia. Xenobiotic metabolism in particular becomes an attractive tool for chemical risk assessment because of its responsiveness against toxic compounds, including those present in CS. This study describes an efficient integration from transcriptomic data to quantitative measures, which reflect the responses against xenobiotics that are captured in a biological network model. We show here that our novel systems approach can quantify the perturbation in the network model of xenobiotic metabolism. We further show that this approach efficiently compares the perturbation upon CS exposure in bronchial and nasal epithelial cells *in vivo* samples obtained from smokers. Our observation suggests the xenobiotic responses in the bronchial and nasal epithelial cells of smokers were similar to those observed in their respective organotypic models exposed to CS. Furthermore, the results suggest that nasal tissue is a reliable surrogate to measure xenobiotic responses in bronchial tissue.

1. Introduction

Humans and other mammals are equipped with a sophisticated machinery to handle carcinogens and other xenobiotic compounds. In studies assessing the effects of cigarette smoke (CS) exposure, a particular interest is given to the metabolism of xenobiotics. The metabolism of xenobiotics includes oxidative reactions by phase I enzymes that convert lipophilic chemical compounds into their hydrophilic forms, followed by phase II conjugation enzymes, and finally the phase III membrane transporters [1]. The second and the last play a role in the elimination of xenobiotic metabolites [1]. The most prominent phase I enzymes are cytochrome P450s (also known as CYPs) that detoxify or activate xenobiotic compounds [1]. The phase I enzymes are also known to be responsible for the metabolism of compounds present in CS, such as nicotine, benzene, polycyclic aromatic hydrocarbons (PAHs), and tobacco-specific nitrosamines (TSNAs) [1, 2]. The induction of a specific CYP has been utilized for the identification of a specific chemical exposure (e.g., induction of CYP1 family specifies the exposure to PAHs) [1, 2]. The roles of various CYPs on the metabolism of CS toxicants have been discussed elsewhere in great detail [3–7]. The metabolization of PAHs and TSNAs can lead to the generation of carcinogenic metabolites that can interact with genomic DNA (i.e., leading to the formation of DNA adducts) [8]. Subsequently, unrepaired DNA adducts would cause gene mutations that lead to the development of cancer (carcinogenesis) [9, 10]. Furthermore, the phase II enzymes (mainly the transferases) catalyze conjugation reactions, such as glucuronidation, sulfation, methylation, and acetylation. These reactions are aimed to detoxify xenobiotic compounds [1, 5]. Moreover, the phase III enzymes refer to the active membrane transporters responsible for the translocation of xenobiotic metabolites across cellular membranes [1, 11]. The initial member of this enzyme family is the ATP-binding

cassette (ABC) family of drug transporters [1]. Nonetheless, the effects of CS on the phase III response have been mainly studied in *in vitro* systems [12, 13].

The expression of CYPs in a specific tissue may suggest a tissue-specific mechanism in response to xenobiotics [14]. Although the liver is known to be the main organ responsible for the metabolism of xenobiotics, the liver is mostly processing toxicants in blood circulation, which come directly from the digestive tract [15]. Consequently, airborne toxicants that come via breathing, including CS exposure, bypass the initial liver detoxification pathway [15]. Therefore, compared to the liver, the respiratory system is exposed to a higher concentration of these toxicants [16]. Thus, the lung and respiratory tract are relevant and valuable for the risk assessment of CS toxicants. Many lung cell types, including bronchial epithelial cells, Clara cells, type II pneumocytes, and alveolar macrophages are capable in metabolizing xenobiotic compounds [14]. Normally, the levels of CYPs in the lung are expressed at trace levels, but they are induced upon CS exposure [14]. Studies have reported that bronchial tissues of smokers exhibit higher levels of CYPs (e.g., CYP1A1 and CYP1B1) as compared to nonsmokers [16–20]. Smoking cessation can reverse the induction of CYP expression upon smoking [20].

CS generates a field of tissue injury throughout the respiratory tract [21]. Tissue injury in the respiratory tract of healthy smokers may precede the development of CS-associated lung diseases [21]. Alteration of the genes encoding the xenobiotic metabolism enzymes has been reported to occur in a similar manner in the nasal as compared to bronchial epithelia, thus supporting the tissue injury hypothesis. For example, increases in expression of CYP1A1 and CYP1B1 were also reported in the nasal epithelium in addition to bronchial epithelium of smokers [17, 18]. Sampling of nasal epithelia by brushing/scraping is less invasive as compared to lung biopsy, thus, providing a better opportunity to screen for respiratory diseases and understand the possible mechanism associated with CS exposure [17]. Nonetheless, identifying gene expression profiles associated with xenobiotic metabolism remains challenging because the expression of genes encoding the xenobiotic enzymes is highly variable within an individual because it may change over time [15]. In this regard, we propose that our system approaches using a network model could potentially be useful to characterize the perturbation in xenobiotic metabolism upon CS exposure.

In a qualitative manner, our network model can be used to gain insight into possible biological mechanisms pertaining to xenobiotic metabolism that are associated to a given exposure [22, 23]. The network model is built to capture biological mechanisms of the xenobiotic metabolism based on evidence from the scientific literature [24] using causal relationships encoded in Biological Expression Language (BEL) [25]. The BEL framework is an open-source technology for managing, publishing, and utilizing a structured life-science knowledge (http://www.openbel.org/) [25]. Previously, we have published the first version of the xenobiotic metabolism network model in the context of the Cellular Stress Network Model [26]. Furthermore, the network model was modified to capture a more comprehensive xenobiotic metabolism response [23] and is shown in Figure 1. The biological network model consists of backbone nodes that are connected by causal edges that carry directional information encoded in BEL. In this current xenobiotic metabolism network model, the central backbone node is the transcriptional activity of aryl hydrocarbon receptor (taof (AHR)). Aryl hydrocarbon receptor (AHR) is a transcription factor known to be activated by xenobiotic compounds. AHR regulates the expression of several target genes (e.g., CYP1A1, CYP1B1, among others). The network model uses transcriptomic data as input that are used to computationally predict the activity/functionality of the backbone nodes (Figure 1, inset) [26–29]. The blue ovals represent the activity of the backbone nodes (i.e., the functional layer) and the green balls represent the expression of genes (i.e., the transcriptional layer). The expression of a given gene can be modulated by one or more backbone nodes as depicted by black arrows. Our network model illustrates the fundamental paradigm shift from forward to backward reasoning. The former considers that the gene transcript abundance is a direct surrogate entity for its protein (or protein function). In contrast, our model considers the latter, in which the changes in gene expression are the consequence of the upstream biological processes embedded in the backbone nodes (the functional layer). Using the backward reasoning, we develop the network perturbation amplitude (NPA) algorithm that provides a quantification of the backbone nodes [23, 25, 29], which is called the "differential network backbone value" (illustrated in Figure 1, inset). Our NPA approach (described in Section 2) aims at scoring functional biological processes based on the fold changes of the gene expression. Thus, the quantification of the backbone nodes (i.e., the "differential network backbone values") in this model reflects the biological mechanisms pertaining to xenobiotic metabolism. The NPA algorithm also integrates the network topology and directionality of edges in the network [23, 25, 29]. Both of which are taken into account for the computation of the NPA score of the entire network along with its companion statistics (described in Section 2). The NPA score of the entire network can be used to evaluate the degree of perturbation between experimental measures (e.g., exposed versus unexposed samples) [23, 29]. Thus, the differential network backbone values in this specific network exemplify the activity of biological mechanisms pertaining to xenobiotic metabolism. Moreover, using transcriptomic data derived from exposed and unexposed samples, the NPA algorithm provides a quantitative measurement of the network perturbation affecting the xenobiotic metabolism.

Our NPA approach has been used previously to compare the perturbation of the xenobiotic metabolism between an *in vivo* dataset derived from bronchial epithelium of smokers obtained by brushing and an *in vitro* dataset derived from organotypic bronchial model exposed to CS [23]. In that study, we demonstrated that at the level of backbone nodes (i.e., the differential network backbone values), the *in vivo* and *in vitro* samples were significantly correlated [23]. Here, we further extended the use of the xenobiotic metabolism network model by presenting some new use cases to probe the comparability not only between the network perturbation derived from *in vivo* and *in vitro*

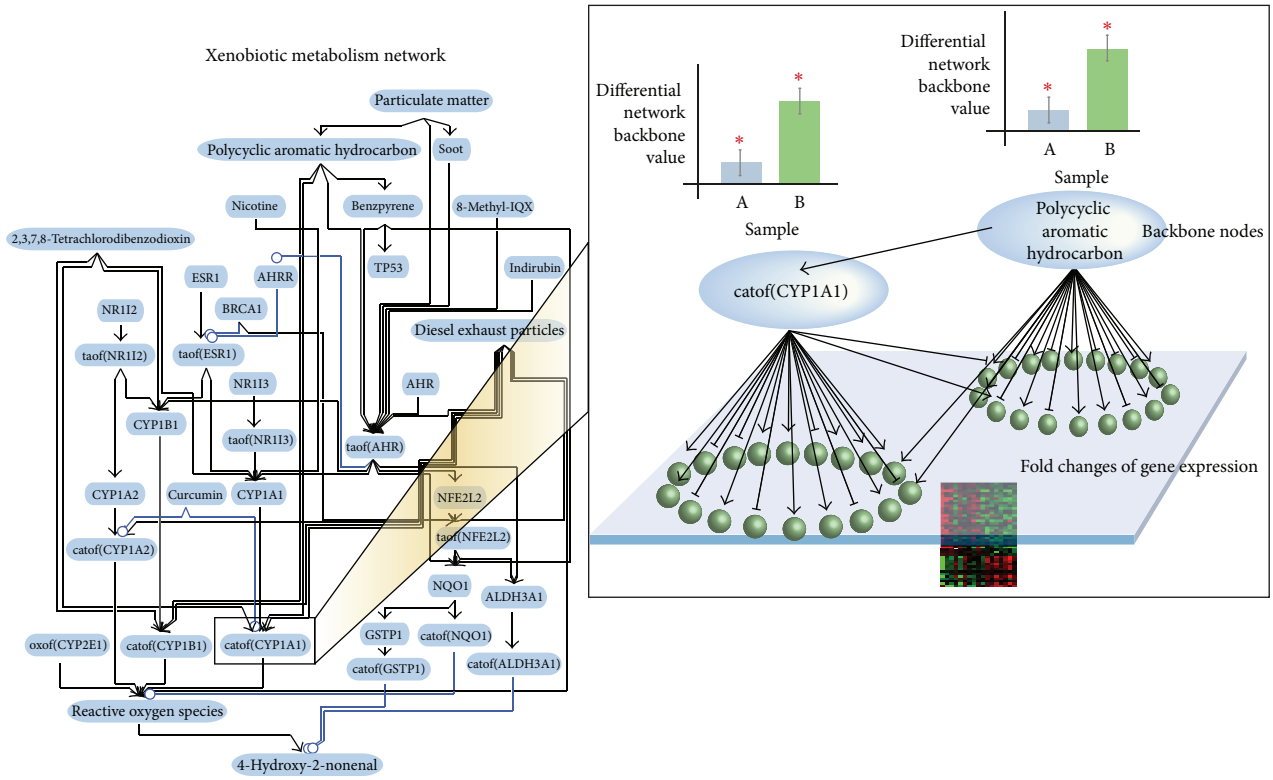


Figure 1: A network model representing the mechanism of xenobiotic metabolism and an illustration of network perturbation amplitude (NPA) approach.

data but also from bronchial and its surrogate nasal tissue.

We and others have recently reported that *in vitro* organotypic human-derived tracheal/bronchial epithelium pseudostratified models resemble human respiratory tract epithelium at the morphological and molecular levels [30–33]. Our group has previously reported using Gene Set Enrichment Analysis; CS exposure affects similar biological changes in bronchial epithelial cells obtained from smokers as compared to the organotypic bronchial epithelium model exposed to CS [31]. In this present study, we not only examined the *in vitro* organotypical bronchial model but also nasal model. Both the nasal and bronchial epithelia organotypic models (Figure 2) contain ciliated cells and express the airway lineage markers, such as p63—a marker of basal epithelials cells that is required for the normal development of epithelial tissues [34]—and Muc5AC that is specifically produced by airway mucous-secreting epithelial cells [35]. Specifically, in this present work, using the NPA approach and the network model, we compare the CS-induced perturbations of the xenobiotic metabolism network in: (1) nasal versus bronchial tissues *in vivo*, (2) nasal versus bronchial tissues *in vitro*, (3) nasal and bronchial tissues *in vivo* versus *in vitro*.

2. Materials and Methods

2.1. Organotypic Tissue Culture Models. MucilAir-human fibroblasts-bronchial and MucilAir-human fibroblast-nasal full-thickness tissue models were generated from primary human respiratory epithelial cells cocultured with primary human airway fibroblasts. The MucilAir models were purchased from Epithelix Sárl (Geneva, Switzerland) and maintained according to the manufacturer's protocol. MucilAir model is a ready-to-use 3D model of differentiated human epithelium [36]. The organotypic tissues were primary human epithelial cells isolated from healthy, nonsmoking, Caucasian donors that were reconstituted using fibroblasts. Coculture of fibroblasts has been shown to contribute to the growth and differentiation of epithelial cells in 3D cultures [37]. The bronchial epithelial cells were obtained from one particular donor, and the nasal epithelial cells were obtained from another donor. Quality control assessments were performed on both models (data not shown). The tissue models were cultured at the air-liquid interface in 0.7 mL media in cell culture inserts (24-well format). The organotypic models were maintained at 37°C for 14 days at the air-liquid interface with fresh medium replaced every 2 days.

2.2. Vitrocell Cigarette Smoke Exposure to the Respiratory Organotypic Tissue Culture Models. After cell culture models grown in culture for 2-3 days, the tissues (in triplicate) were exposed at the air-liquid interface to 16% (vol/vol) mainstream CS exposure (a total of 4 cigarettes, 3R4F) with 1 hour rest between each cigarette and 60% humidified air in the Vitrocell systems (Waldkirch, Germany). The 60% humidified air exposure was used as a control exposure.

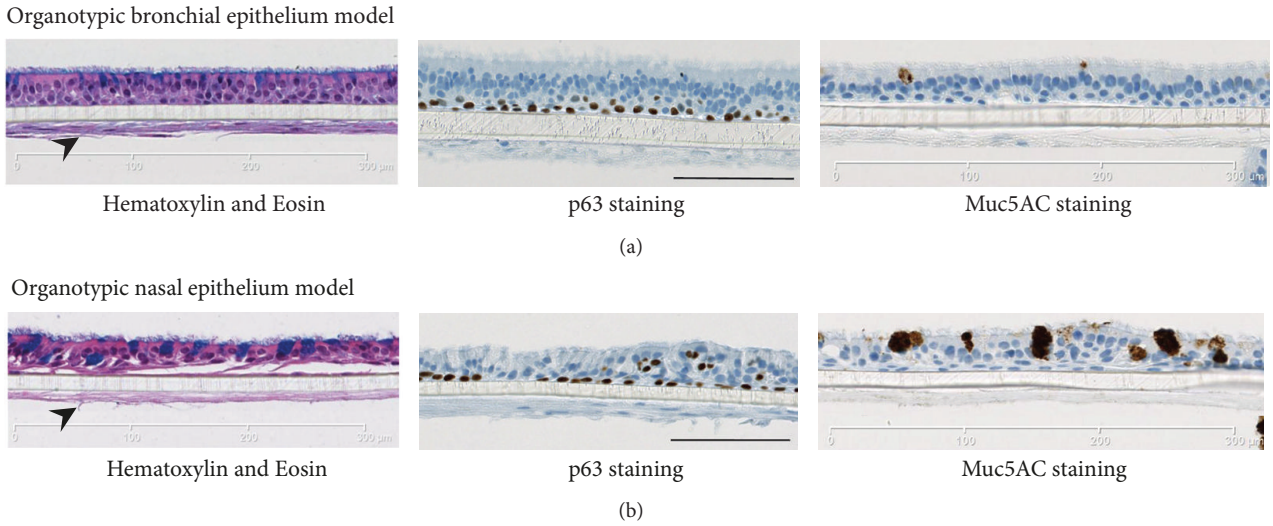


Figure 2: Organotypic bronchial (a) and nasal (b) models. The *in vitro* models contained ciliated cells shown in the apical layer of the Hematoxylin and Eosin stained cells (left). The models were cocultured with fibroblasts that are important for the growth and differentiation of epithelial cells (indicated by arrows). Staining of airway lineage markers: p63 and Muc5AC are shown (center and right).

The Total Particulate Matter (TPM) inside the exposure chamber has been measured for each CS concentration (the mean TPM deposition measured after each cigarette was 2842.4 ng/cm^2± SEM = 570.7, $N = 24$). The reference cigarette 3R4F was obtained from the University of Kentucky (http://www.ca.uky.edu/refcig/) and smoked on the 30-port carousel smoking machine (SM2000, Philip Morris, Int.) according to the Health Canada regimen [38]. After exposure, the organotypic models were incubated with fresh culture medium immediately (0 h after exposure). Additionally various durations of postexposure were implemented (4, 24, and 48 h) before tissues were harvested for further analyses.

2.3. RNA and Microarray Hybridization. Exposed tissues ($n = 3$) at 0, 4, 24, and 48 h postexposure time were washed 3 times with ice-cold PBS and subsequently lysed using Qiazol lysis reagent (miRNeasy Mini Kit, Qiagen) and frozen at −80°C for up to 1 week. The miRNeasy Mini Kit was used for the extraction and purification of mRNA. Total RNA quantity was measured using NanoDrop ND1000 and qualitatively verified using an Agilent 2100 Bioanalyzer profile (A RIN number greater than 8). For the mRNA analysis, total RNA (100 ng) was processed according to the GeneChip HT 3′ IVT Express User Manual (Affymetrix). Genechip Human Genome U133 Plus 2 Arrays were used for microarray hybridization. The dataset has been submitted to Arrayexpress (Accession code = E-MTAB-1721).

2.4. Microarray Data Processing. Data processing and scoring methods were implemented using the *R* statistical environment version 2.14 [39]. Raw RNA expression data were analyzed using the *affy* and *limma* packages of the Bioconductor suite of microarray analysis tools (version 2.9) available in the *R* statistical environment [40, 41]. Robust Multichip Average (GCRMA) background correction and quantile normalization were used to generate probe set expression values [42]. For each data set, an overall linear model was fitted to the data for the specific contrasts of interest (e.g., the comparisons of "treated" and "control" conditions) generating raw P values for each probe set on the microarray, which were further adjusted using the Benjamini-Hochberg procedure. A blocking factor (the exposure plate) from the experiment design was accounted in the model for data processing.

2.5. Network-Based Analysis. Leveraging the "cause-and-effect" network models together with Network Perturbation Amplitude (NPA) algorithms ([29, 43]), the fold changes of gene expression were translated into differential network backbone value for each backbone node in the network (Figure 1). The "differential network backbone value" was the result of a fitting procedure between the network model and the gene expression fold changes, where the smoothest function (accounting for the sign of the causal edges) was derived by further imposing a boundary condition on the backbone nodes corresponding to the gene expression changes. Statistical correlations were computed for the differential network backbone values and fold changes of gene expression, including R^2, Pearson correlation, and Spearman correlation, along with the P values. For a negative control analysis, we computed a permutation test to assess if the correlation obtained between the differential network backbone values were solely due to the dimension reduction effect. Genes underlying the network were randomly permuted 1000 times for each comparison group to decorrelate the fold changes of gene expression (the GSE16008 nasal versus bronchial data were used in this example). Subsequently, correlations between the backbone values were computed. This approach leads to a P value 0.002 two sided (see Supplementary Figure 1 in the Supplementary Material available online at http://dx.doi.org/10.1155/2013/512086).

The differential network backbone values were in turn summarized into a quantitative measure of NPA score for the entire network. The NPA is computed as a (semi-)Sobolev-type norm on the signed directed graph underlying the network (N), (which can be expressed as a quadratic form). In summary, the NPA algorithm considers two main input components. First the "cause-and-effect" network model describing the mechanism and second, the gene expression dataset from a well-designed experiment.

In addition to the confidence intervals of the NPA scores, which account for the experimental error (e.g., the biological variation between samples in an experimental group), companion statistics were derived to describe the specificity of the NPA score to the biology described in the network. Because NPA is a quadratic form of the fold changes, its variance can be computed based on the fold-change estimated variances. A confidence interval is subsequently derived using the central limit theorem. Two permutation tests were implemented [43], whereby first, to assesses if the results were specific to the underlying evidence (i.e., gene fold-changes) in the model, leading to a permutation P value (denoted by $^{*}O$ in the figures when P-value < 0.05). Second, to assess whether the "cause-and-effect" layer of the network significantly contributed to the amplitude of the network perturbation (denoted by K^{*} in the figures when P value < 0.05). The network was considered to be specifically perturbed if both P values mentioned before were <0.05, and the perturbation was called significant when the confidence interval was greater than 0.

2.6. Cytochrome Activity Assay. We measured the activity of CYP1A1 and CYP1B1 using nonlytic P450-Glo assays (CYP1A1 assay cat number V8752; CYP1B1 assay cat number V8762; Promega) based on luminescence at the 48 h after exposure on the human organotypic nasal and bronchial models. The assay was performed according to the manufacturer's recommendations. Briefly, both nasal and bronchial epithelia models were incubated in medium with luminogenic CYP-Glo substrate, such as luciferin-CEE for 3 h (CYP1A1 and CYP1B1), to produce a luciferin product that can be quantified in the supernatant by a light-generating reaction upon the addition of luciferin detection reagent.

3. Results and Discussion

3.1. Comparison between Xenobiotic Metabolism Responses in Bronchial and Nasal Epithelia In Vivo upon CS Exposure. Another group has reported that alterations of xenobiotic metabolism in the bronchial epithelium obtained from human donors are similar to those in the nasal epithelium [21]. This observation supports the field of tissue injury hypothesis, in which changes in the respiratory tract of smokers precede the development of CS-associated lung diseases [21]. To further examine this hypothesis, we used the NPA approach and the xenobiotic network model to compare the differential network backbone values derived from the bronchial and nasal samples in the *in vivo* dataset GSE16008 (smokers versus nonsmokers). This approach was taken to compare the biological mechanisms associated to xenobiotic metabolism that were perturbed by CS exposure in these two tissues.

We have reported before that the xenobiotic metabolism network model can capture the common perturbation from several independent datasets [22, 23]. In this study, the publicly available dataset (GSE16008) was used to compare the xenobiotic responses in the human bronchial and nasal epithelia upon CS exposure. The GSE16008 dataset contains gene expression from nasal and bronchial epithelial cells obtained from healthy current smokers and nonsmokers. The bronchial epithelial cells were collected by bronchoscopy, whereas the nasal epithelial cells were collected by brushing the inferior turbinate [18]. Because we have not analyzed GSE16008 before, we first probed the comparability of this particular dataset to other publicly available datasets (i.e., GSE7895, GSE19667, and GSE14633). These datasets contain gene expression of bronchial epithelial cells obtained by bronchoscopy from smokers and nonsmokers (GSE7895, [20]), (GSE19667, [44]), and (GSE14633, [45]). Figure 3 shows the correlation of the differential network backbone values in the xenobiotic metabolism network model (the model is depicted in Figure 1) using the NPA approach between GSE16008 and the aforementioned datasets.

Figure 4(a) shows that the differential network backbone values were well correlated between the *in vivo* bronchial and nasal brushing epithelia. Furthermore, we illustrated how the backbone AHR was computed from the gene expression data (Figure 4(a), inset). Figure 4(a) shows a correlation of the differential network backbone values derived from the bronchial and nasal samples in the *in vivo* dataset GSE16008 (smokers versus nonsmokers) computed using the NPA approach in the xenobiotic metabolism network model. Each data point represents a backbone node in the xenobiotic network model. Representative node labels are shown. The blue line is the linear regression line computed by least squares fit with significant P value < 0.05. The 95%-confidence intervals of the differential backbone values are shown for the two perturbations (axes). The differential network backbone values generated from the *in vivo* bronchial data were in general greater than those from the *in vivo* nasal data (inferred from the regression line), with the exception of aryl hydrocarbon receptor (AHR) (Figure 4(a)).

Figure 4(b) captures the differential network backbone values in the xenobiotic metabolism network model using the *in vivo* bronchial (left) and nasal (right) data. The different colors reflect the quantification of the backbone nodes (i.e., the "differential network backbone values") derived from the NPA algorithm that demonstrate the biological mechanism pertaining to xenobiotic metabolism. The values less than 0 indicate downregulation of the backbone node activity; whereas, the values more than 0 indicate upregulation of the backbone node activity. $^{*}P$ values < 0.05. Furthermore, the nature of the network perturbation, which was reflected on the differential network backbone values, was similar between the bronchial and nasal epithelia. For example, cigarette smoking was associated with decreased activation of the aryl hydrocarbon receptor repressor (AHRR) in both the bronchial and nasal samples (Figure 4(b)). AHRR is known

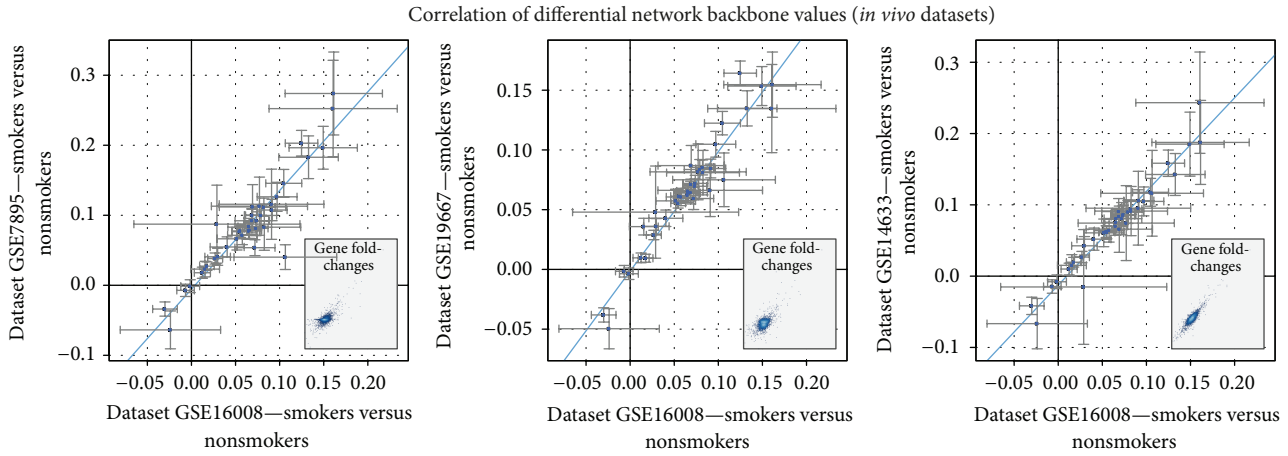


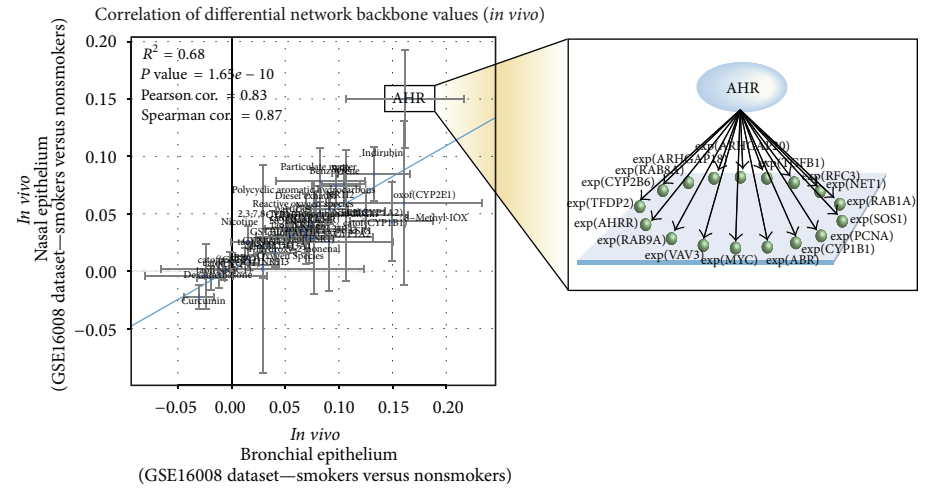
Figure 3: Comparability of GSE16008 dataset to other publicly available datasets. Correlations among the differential network backbone values from different human datasets in the xenobiotic metabolism network were shown. The human datasets comprise smoker versus nonsmoker data. Each data point represents a backbone node in the network. The 95%-confidence intervals of the differential network backbone values are shown for the two perturbations (axes). Blue lines show the linear regression lines computed by least squares fit. All the regression models were significant (P value < 0.05). Insets illustrate the correlation of the fold change of gene expressions.

to inhibit the binding of AHR to xenobiotic-responsive elements (XRE), thus, suppressing the transcription of AHR-dependent genes, including CYP1A1, CYP1A2, and CYP1B1 [46]. Consistently, we observed the upregulated differential network backbone values for these aforementioned CYPs (Figure 4(b)). Despite the expression of AHR falls farther away from the regression line (Figure 4(a) as mentioned before); interestingly, the differential network backbone values of the AHR node were similar between the bronchial and nasal, indicating that the activity of AHR upon CS exposure was increased to the same extend in both bronchial and nasal (Figures 4(a) and 4(b)). This result supports the notion that AHR plays an important role in the activation of CS toxicants not only in the lower respiratory tract but also in the upper respiratory tract. Studies have indicated that AHR is a promiscuous receptor that is capable in binding to diverse chemicals, leading to their activations [1]. In regard to smoking, this observation further supports that CS exposure generates a field of tissue injury throughout the respiratory tract [21].

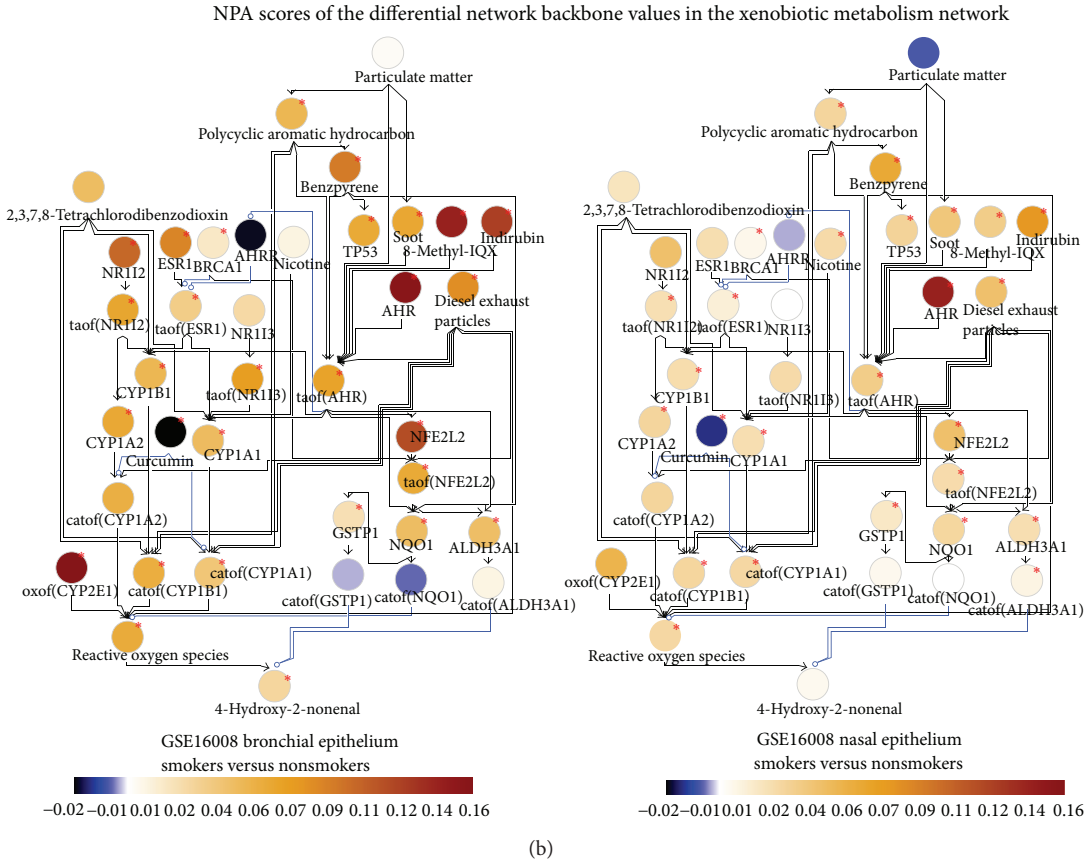
Figure 4(c) shows the bar plot of the NPA scores for the entire networks along with their companion statistics along with their companion statistics ${}^{*}O$ and K^{*} as described in Section 2 (P values < 0.05). These significant statistics suggest that both *in vivo* nasal and bronchial samples from the dataset significantly demonstrate the biological mechanisms represented in the xenobiotic metabolism network model (described in Section 2). These results suggest that the nasal as a surrogate tissue of the bronchial epithelium elicits similar xenobiotic responses upon CS exposure, which were reflected by the similar changes of the differential network backbone values in the xenobiotic metabolism network model. This result also supports the overall CS exposure-related impact on the tissues lining the respiratory tract [21].

Moreover, unlike the correlation at the backbone levels (i.e., functional layer) (Figure 4(a)), a correlation between the gene expression generated from the dataset GSE16008 (i.e., transcriptional layer) was not observed (Figure 4(d)). These results indicate that the utilization of our NPA approach using the network model, which comprised these two layers (i.e., functional and transcriptional layers), could facilitate a high-resolution comparison of high-throughput transcriptomic data and to understand the biological insight ingrained in the data.

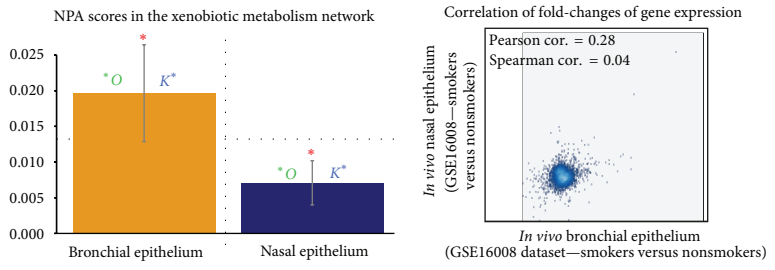
3.2. Comparison between Xenobiotic Metabolism Responses in Organotypic Bronchial and Nasal In Vitro Models upon CS Exposure. We further determined whether the same information could be observed *in vitro*. Development of a reliable *in vitro* system that mimics the *in vivo* condition has been challenging. Recently, organotypic culture models of human cells have been developed and utilized to understand biological processes [30–33, 47]. In this present study, we compared the network perturbations that occurred in *in vitro* organotypic bronchial and nasal epithelia models that were exposed to whole CS (see Section 2). The gene expression from these tissue models was measured in cells that were immediately harvested after the last exposure (0 h after exposure). Figure 5(a) shows that the differential network backbone values were well correlated between the *in vitro* bronchial and nasal epithelia. This comparability at the functional layer was in agreement with what was observed using the data generated from the *in vivo* dataset. Each data point represents a backbone node in the xenobiotic network model. The blue line is the linear regression line computed by least squares fit with significant P value < 0.05. The 95%-confidence intervals of the differential backbone values are shown for the two perturbations (axes). Figure 5(a), inset



(a)



(b)



(c) (d)

FIGURE 4: Correlation between the differential network backbone values in response to CS exposure generated from *in vivo* human bronchial and nasal datasets in the xenobiotic metabolism network model.

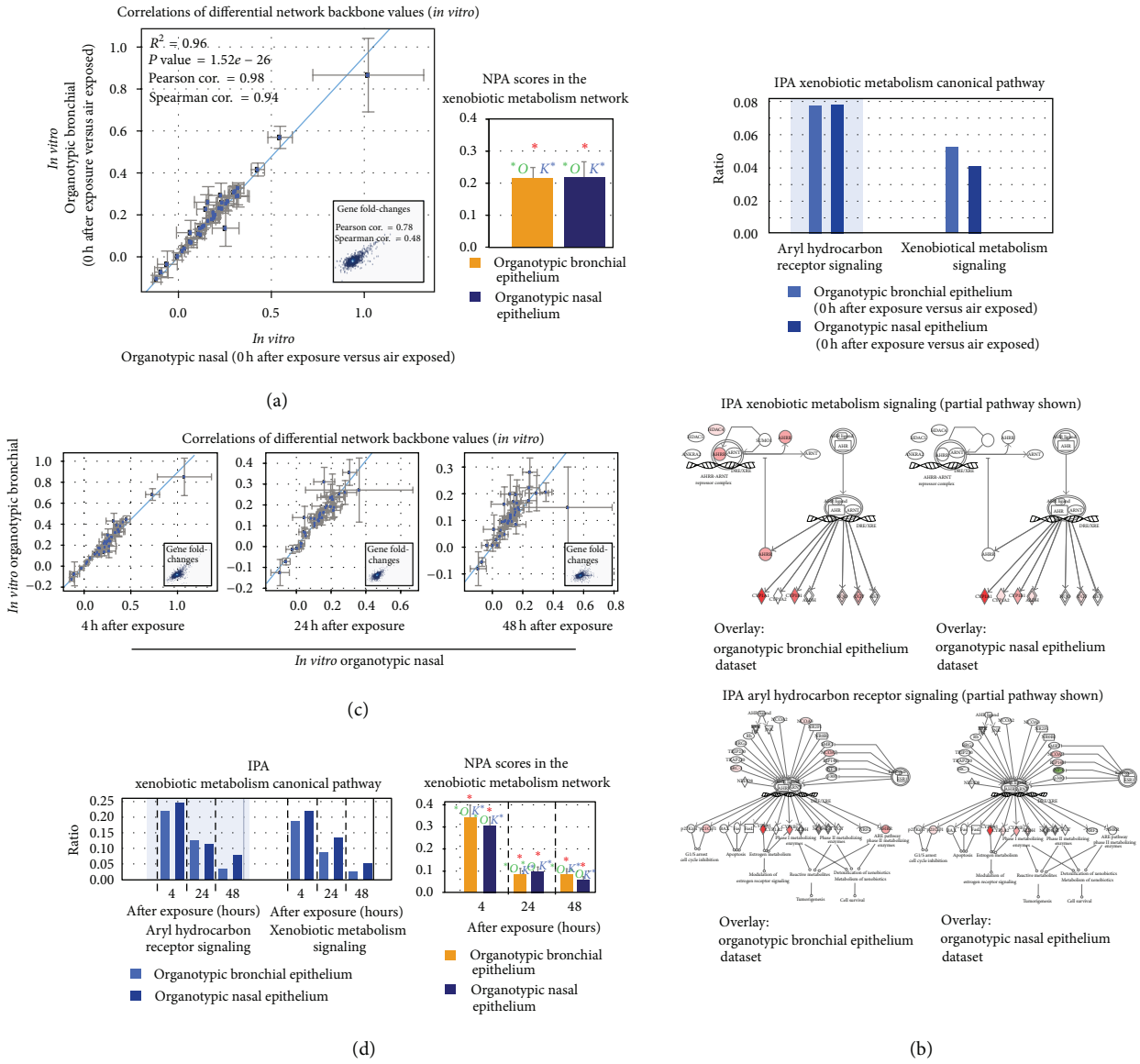


FIGURE 5: Comparison between xenobiotic metabolism responses in organotypic bronchial and nasal epithelia *in vitro* models upon CS exposure.

illustrates the correlation between the fold changes of the gene expression (correlation at the transcriptional layer). The NPA scores (bar plot) indicating the statistical significance of the perturbation of the xenobiotic metabolism network model in response to smoking are shown: *indicates significance of NPA scores of the entire network level generated from the *in vitro* bronchial and nasal datasets as well as their companion statistics $^*O$ and K^* as described in Section 2 (P values < 0.05). The bar plot (Figure 5(a)) shows the NPA scores for the entire networks along with their companion statistics. These significant statistics suggest that both *in vitro* nasal and bronchial samples from the dataset significantly demonstrate the biological mechanisms represented in the xenobiotic metabolism network model (described in Section 2).

Furthermore, to investigate how our analysis using the xenobiotic metabolism network model was comparable to the commercially available data analysis and interpretation tool (Ingenuity Pathways Analysis (IPA)), the same datasets generated from the *in vitro* organotypic samples were uploaded to IPA. Within the IPA's knowledge base, the "Xenobiotic Metabolism" canonical pathways are comprised of two signaling pathways: the "Aryl Hydrocarbon Receptor Signaling" and the "Xenobiotic Metabolism Signaling." Figure 5(b) shows significant associations between the datasets and the two IPA's canonical pathways within the category of "Xenobiotic Metabolism" (P value < 0.05). The y-axis displays the ratio calculated as follows: the number of genes in the associated pathways that meet cutoff criteria, divided by the total number of genes that make up that specific pathway.

The taller the bars, the more genes were associated with the pathway. Representative pathways overlayed with the two datasets generated from the organotypic bronchial and nasal models are shown. Interestingly, the bronchial and nasal data were associated to the two signaling pathways in a similar manner, which was indicated by the similarity of the ratios (Figure 5(b), top). This observation was in agreement with our approach using the NPA analyses and the network model (Figure 5(a)).

Additionally, we examined the effects of various postexposure time points to assess the ability of cells to recover from CS exposure. We hypothesized that the longer the duration of postexposure, the less perturbed the xenobiotic metabolism would be. The differential network backbone values continue to be correlated between the bronchial and nasal at the 4, 24, and 48 h after exposure time (Figure 5(c)). Nevertheless, the correlations were reduced as the duration of the postexposure increased (Figure 5(c) and Table 1). Each data point represents a backbone node in the xenobiotic network model as indicated with the node labels. The blue line is the linear regression line computed by least squares fit with significant P value < 0.05. The 95%-confidence intervals of the differential backbone values are shown for the two perturbations (axes). Figure 5(c), inset, illustrates the correlation between the fold changes of the gene expression (correlation at the transcriptional layer, Table 1). The reduced responses that were inferred from the reduced correlations between the differential network backbone values, were also reflected from the analysis using IPA, in which decreased ratios (i.e., the association between the datasets with the two pathways) were observed in the datasets at the later time point of postexposure (Figure 5(d), left bar plot). The results from IPA analysis were also in agreement to those reflected by the NPA scores for the entire network (Figure 5(d), right bar plot), in which the later time point of postexposure had reduced scores. Taken together, this data suggests that the shorter the postexposure time, the more perturbed the xenobiotic metabolism in both bronchial and nasal tissues. This observation is consistent with a previous study in which a transient induction of phase I xenobiotic metabolism enzymes (e.g., *cyp1A1* and *aldh3A1*) is observed in CS-exposed lung tissues of Sprague-Dawley rats [48]. Furthermore, this could offer a likely explanation for why we observed a better correlation of the differential network backbone values between the *in vitro* organotypic bronchial and nasal models with shorter postexposure time (Figure 5(c) and Table 1).

3.3. In Vivo and In Vitro Comparison between Xenobiotic Metabolism Responses in Bronchial and Nasal Epithelia upon CS Exposure. We further examined whether *in vitro* organotypic models could reveal a similar xenobiotic response upon CS exposure as compared to that observed *in vivo*. Thus, we determined whether the differential network backbone values generated from the *in vivo* datasets were correlated to those from the *in vivo*. The NPA approach that quantifies the changes at the backbone levels (i.e., the differential network backbone values) could indicate the potential biological mechanisms that were perturbed upon exposure to CS. Therefore, whether similar biological responses occurred in *in vivo* situation were comparable to those in *in vitro* models can be inferred from the correlation between the differential network backbone values. Figures 6(a) and 6(b) show the correlations, in bronchial and nasal samples, respectively, between the differential network backbone values generated from *in vivo* dataset to those generated from *in vitro* models. This observation is in agreement with our other publication, in which a similar biological alteration was observed in *in vivo* bronchial epithelial cells as compared to an *in vitro* organotypic bronchial epithelial model (EpiAirway system, MatTeK Corporation) [31]. Nonetheless, this present study further suggests that the *in vitro* organotypic nasal model would also be useful to investigate the mechanisms occur in the *in vivo* nasal situation upon smoking.

Moreover, we compared the data derived from organotypic *in vitro* models at various postexposure time to the *in vivo* datasets at the backbone nodes level in the xenobiotic metabolism network model; the xenobiotic responses were better correlated (Table 2) in the bronchial (Figure 6(a)) as compared to the nasal samples (Figure 6(b)). The differential network backbone values were derived from the bronchial and nasal data in the *in vivo* dataset GSE16008 (smokers versus nonsmokers) and from the data generated from CS-exposed *in vitro* organotypic bronchial and nasal models computed using the NPA approach in the xenobiotic metabolism network model. Each data point represents a backbone node in the xenobiotic network model. The blue line is the linear regression line computed by least squares fit with significant P value < 0.05. The 95%-confidence intervals of the differential backbone values are shown for the two perturbations (axes). The insets illustrate the correlations between the fold-changes of the gene expression (correlation at the transcriptional layer). Zhang and colleagues have previously reported that the effect of smoking is less pronounced in the nasal epithelium when compared to bronchial epithelium obtained from smokers [18], which could explain why the correlation observed in the nasal samples was weaker. Moreover, to better assess the *in vitro* organotypic nasal model, we tested the effects of CS exposure on the enzymatic activity of CYP1A1 and CYP1B1. We found that CS exposure significantly increased the activity of both CYP1A1 and CYP1B1 measured in the nasal epithelium *in vitro* model (Figure 6(c)), supporting the potential of the nasal model to be utilized for toxicity assessment against airborne exposure. The CYP activities (luminescence, RLU) of the CS-exposed nasal tissues as compared to the air-exposed tissues were measured at 48 h after exposure (see Section 2). Shown are CYP1A1 and CYP1B1 activities obtained from triplicate measurements ($N = 3$), $^{*}P < 0.05$ as compared to the air-exposed tissue. Additionally, although the xenobiotic responses generated from the *in vitro* organotypic models at the later time of postexposure were reduced (Figure 5(d) and Table 2), the differential network backbone values remained well correlated as compared to those generated from the *in vivo* datasets (Figure 6(d) and Table 2). However, the *in vivo*/*in vitro* correlations of the gene expression (Figure 6(d), insets and Table 2) were weak.

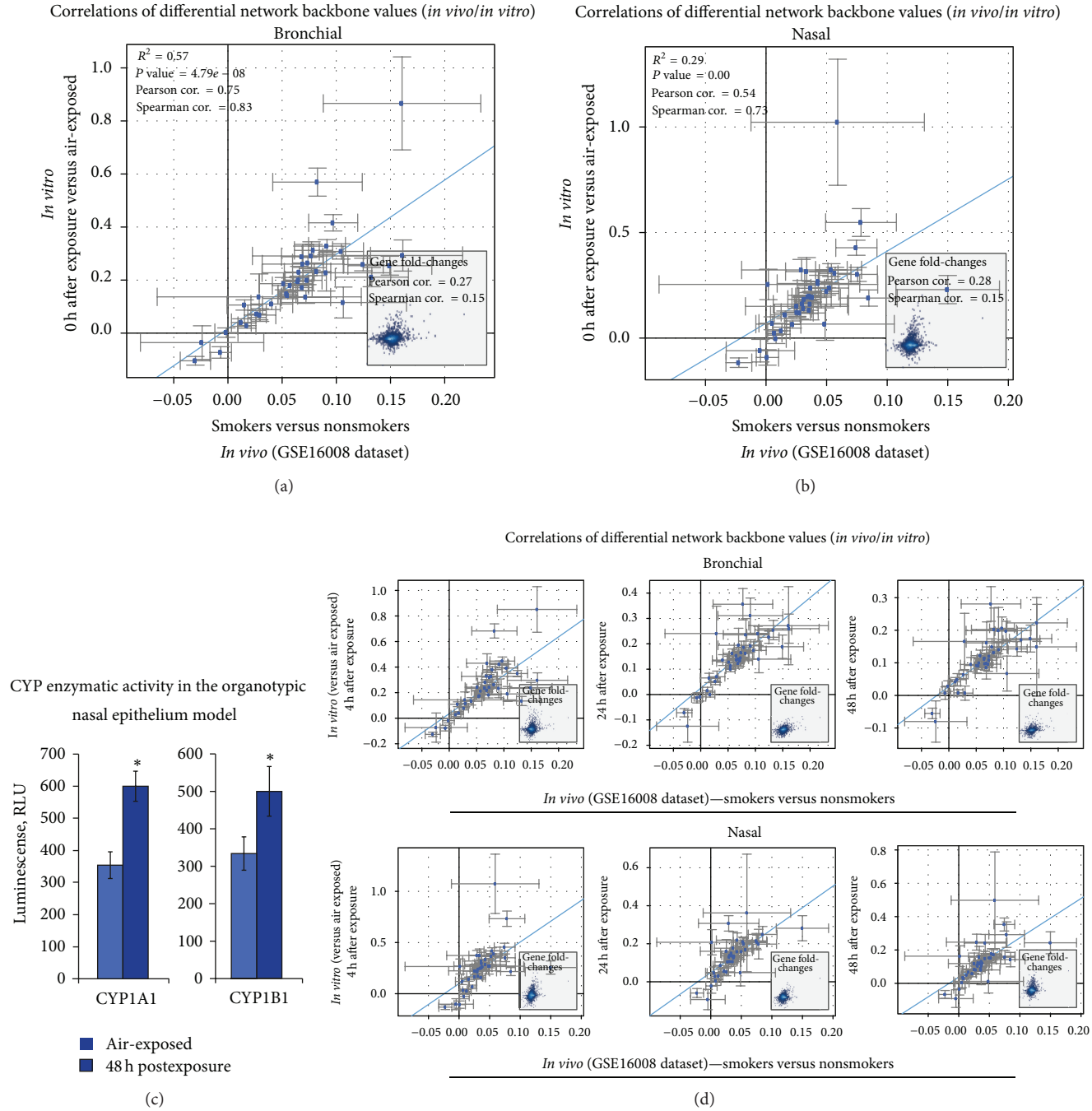


FIGURE 6: Correlations between the differential network backbone values in response to CS exposure generated from *in vivo* datasets and *in vitro* organotypic models in the xenobiotic metabolism network model.

4. Conclusion

Here we show that our quantitative systems-level approach utilizing the xenobiotic metabolism network model allowed a robust comparison derived from transcriptional data. This approach could provide a mechanistic insight that occurred in response to CS exposure, which is reflected from the differential network backbone values. The quantification of the xenobiotic network model perturbation using the NPA approach not only could compare the responses observed from datasets generated from *in vivo* samples and *in vitro* organotypic models but also from bronchial and its surrogate nasal epithelia. Furthermore, our results suggested that the organotypic nasal *in vitro* model could be useful as a risk assessment tool in understanding biological mechanisms leading to lung diseases associated to airborne exposure. Our

Table 1: Statistical correlation between the bronchial versus nasal *in vitro* data.

Comparison group	Between the backbone values		Between the fold change of genes expression	
	Pearson correlation	Spearman correlation	Pearson correlation	Spearman correlation
Bronchial *in vitro* versus nasal *in vitro* (4 h after exposure)	0.97	0.95	0.72	0.55
Bronchial *in vitro* versus nasal *in vitro* (24 h after exposure)	0.93	0.94	0.62	0.49
Bronchial *in vitro* versus nasal *in vitro* (48 h after exposure)	0.77	0.86	0.39	0.37

P values < 0.05 for all comparisons.

Table 2: Statistical correlation between the *in vivo* versus *in vitro* data at various postexposure times.

Comparison group	Between the backbone values		Between the fold change of genes expression	
	Pearson correlation	Spearman correlation	Pearson Correlation	Spearman Correlation
Bronchial				
In vivo versus 4 h after exposure *in vitro*	0.73	0.77	0.25	0.13
In vivo versus 24 h after exposure *in vitro*	0.81	0.83	0.37	0.29
In vivo versus 48 h after exposure *in vitro*	0.77	0.80	0.35	0.30
Nasal				
In vivo versus 4 h after exposure *in vitro*	0.57	0.76	0.35	0.27
In vivo versus 24 h after exposure *in vitro*	0.71	0.74	0.31	0.09
In vivo versus 48 h after exposure *in vitro*	0.65	0.73	0.26	0.14

P values < 0.05 for all comparisons.

results are consistent with an overall CS exposure-related impact on the tissues lining the respiratory tract, and thus supporting the field of tissue injury theory [21].

Studies have reported that CS exposure is associated with increased expression of genes encoding the xenobiotic metabolism enzymes, such as CYP1A1 and CYP1B1 in both the nasal [17, 18] and buccal epithelia [49, 50]. Similar to the nasal epithelium, buccal epithelium has been postulated as a suitable surrogate tissue for the lung, which could be useful to determine disease risk biomarkers [51]. Because collections of both nasal and buccal epithelial samples are relatively simpler and less invasive as compared to the collection of bronchial epithelium, these tissues become attractive surrogate tissues for toxicology assessment in response to CS exposure. Sampling of the lung is usually done by brushing or biopsy [52]. However, these methods are invasive, thus, unfeasible for large clinical studies [51]. Additionally, the use of *in vitro* organotypic models provides an attractive tool for toxicology assessments of specific airborne exposures. In this study, we also demonstrated that the perturbation of the xenobiotic metabolism in the CS-exposed organotypic nasal *in vitro* epithelia models resembled that in the nasal epithelial cells obtained by brushing from smoker donors. Nonetheless, whether similar results would be observed in organotypic model derived from different donors is unknown. Donor-dependent variability is expected [36] and should be addressed in future studies. Furthermore, future studies should also investigate whether other bronchial surrogate tissues (e.g., buccal epithelial *in vivo* samples and organotypic *in vitro* models) could be utilized to assess and compare the perturbation in the xenobiotic metabolism upon CS exposure. Such data would further highlight the relevance and practicality of *in vitro* organotypic models for toxicology assessment.

Our present work provided a useful example for the utilization of transcriptomic data for impact assessment that focuses on xenobiotic responses against airborne exposure. However, theories have been developed supporting how the entire respiratory tract exhibits genomic, epigenomic (e.g., methylation of genes encoding the xenobiotic metabolizing enzymes), transcriptomic, and proteomic modifications [17, 53, 54]. Additionally, CS exposure has often been associated with adduct formation not only in the lung tissue but also in the blood circulation [55–62]. Figure 7 depicts how transcriptomic data could be leveraged into various systems approaches that implement the larger spectrum of "omics" technologies. Although this present study described the utilization of transcriptomic data, further information from genome and its derivatives, including proteins, metabolites, and adducts would be useful for the overall assessment of CS exposure on the metabolism of xenobiotic.

Disclosure

The authors and the research described in this paper were supported by Philip Morris International.

Authors' Contribution

Anita R. Iskandar and Florian Martin equally contributed.

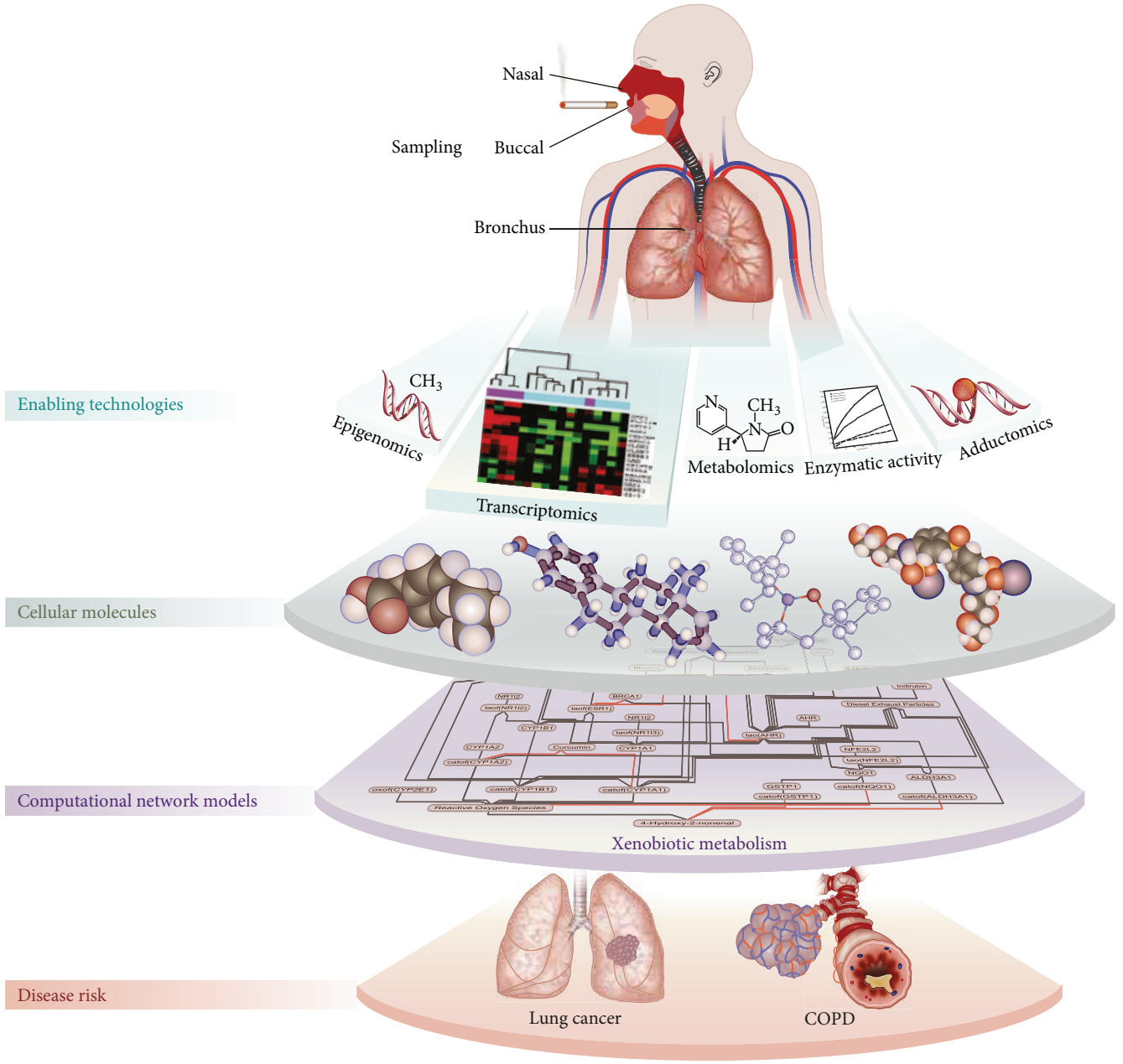


FIGURE 7: Our present work could be implemented to the new generation of "omics" technology for the overall assessment of CS exposure pertaining to the perturbation of xenobiotic metabolism. This current work provided an useful example for the utilization of transcriptomic data for impact assessment that focuses on xenobiotic responses against airborne exposure.

Acknowledgments

The authors would like to thank Diana Kuehn for the cell culture work and CYP activity assays, both Remi Dulize and Emmanuel Guedj for the nucleic acid research as well as Carine Poussin and Nikolai Ivanov for reviewing the paper. The authors would like to thank Shoaib Majeed for the well conducted whole smoke exposed experiment in the organotypic culture.

References

[1] C. J. Omiecinski, J. P. Vanden Heuvel, G. H. Perdew, and J. M. Peters, "Xenobiotic metabolism, disposition, and regulation by receptors: from biochemical phenomenon to predictors of major toxicities," *Toxicological Sciences*, vol. 120, supplement 1, pp. S49–S75, 2011.

[2] A. Sharma, K. Saurabh, S. Yadav, S. K. Jain, and D. Parmar, "Expression profiling of selected genes of toxication and

detoxication pathways in peripheral blood lymphocytes as a biomarker for predicting toxicity of environmental chemicals," *International Journal of Hygiene and Environmental Health*, 2012.

[3] S. S. Hecht, "DNA adduct formation from tobacco-specific N-nitrosamines," *Mutation Research*, vol. 424, no. 1-2, pp. 127–142, 1999.

[4] D. Li, P. F. Firozi, L.-E. Wang et al., "Sensitivity to DNA damage induced by benzo(a)pyrene diol epoxide and risk of lung cancer: a case-control analysis," *Cancer Research*, vol. 61, no. 4, pp. 1445–1450, 2001.

[5] T. Shimada, "Xenobiotic-metabolizing enzymes involved in activation and detoxification of carcinogenic polycyclic aromatic hydrocarbons," *Drug Metabolism and Pharmacokinetics*, vol. 21, no. 4, pp. 257–276, 2006.

[6] N. M. DeVore and E. E. Scott, "Nicotine and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone binding and access channel in human cytochrome P450 2A6 and 2A13 enzymes," *The Journal of Biological Chemistry*, vol. 287, no. 32, pp. 26576–26585, 2012.

[7] J. Hukkanen, P. Jacob III, and N. L. Benowitz, "Metabolism and disposition kinetics of nicotine," *Pharmacological Reviews*, vol. 57, no. 1, pp. 79–115, 2005.

[8] J. H. Kim, K. H. Stansbury, N. J. Walker, M. A. Trush, P. T. Strickland, and T. R. Sutter, "Metabolism of benzo[a]pyrene and benzo[a]pyrene-7,8-diol by human cytochrome P450 1B1," *Carcinogenesis*, vol. 19, no. 10, pp. 1847–1853, 1999.

[9] R. Piipari, K. Savela, T. Nurminen et al., "Expression of CYP1A1, CYP1B1 and CYP3A, and polycyclic aromatic hydrocarbon-DNA adduct formation in bronchoalveolar macrophages of smokers and non-smokers," *International Journal of Cancer*, vol. 86, no. 5, pp. 610–616, 2000.

[10] D. H. Phillips, B. Schoket, A. Hewer, E. Bailey, S. Kostic, and I. Vincze, "Influence of cigarette smoking on the levels of DNA adducts in human bronchial epithelium and white blood cells," *International Journal of Cancer*, vol. 46, no. 4, pp. 569–575, 1990.

[11] E. Croom, "Metabolism of xenobiotics of human environments," *Progress in Molecular Biology and Translational Science*, vol. 112, pp. 31–88, 2012.

[12] Y. An, A. Kiang, J. P. Lopez et al., "Cigarette smoke promotes drug resistance and expansion of cancer stem cell-like side population," *PLoS ONE*, vol. 7, no. 11, Article ID e47919, 2012.

[13] M. van der Deen, E. G. E. de Vries, H. Visserman et al., "Cigarette smoke extract affects functional activity of MRP1 in bronchial epithelial cells," *Journal of Biochemical and Molecular Toxicology*, vol. 21, no. 5, pp. 243–251, 2007.

[14] X. Ding and L. S. Kaminsky, "Human extrahepatic cytochromes P450: function in xenobiotic metabolism and tissue-selective chemical toxicity in the respiratory and gastrointestinal tracts," *Annual Review of Pharmacology and Toxicology*, vol. 43, no. 1, pp. 149–173, 2003.

[15] N. Finnström, B. Ask, M.-L. Dahl, M. Gadd, and A. Rane, "Intra-individual variation and sex differences in gene expression of cytochromes P450 in circulating leukocytes," *Pharmacogenomics Journal*, vol. 2, no. 2, pp. 111–116, 2002.

[16] T. Thum, V. J. Erpenbeck, J. Moeller, J. M. Hohlfeld, N. Krug, and J. Borlak, "Expression of xenobiotic metabolizing enzymes in different lung compartments of smokers and nonsmokers," *Environmental Health Perspectives*, vol. 114, no. 11, pp. 1655–1661, 2006.

[17] S. Sridhar, F. Schembri, J. Zeskind et al., "Smoking-induced gene expression changes in the bronchial airway are reflected in nasal and buccal epithelium," *BMC Genomics*, vol. 9, article 259, 2008.

[18] X. Zhang, P. Sebastiani, G. Liu et al., "Similarities and differences between smoking-related gene expression in nasal and bronchial epithelium," *Physiological Genomics*, vol. 41, no. 1, pp. 1–8, 2010.

[19] A. Spira, J. Beane, V. Shah et al., "Effects of cigarette smoke on the human airway epithelial cell transcriptome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 27, pp. 10143–10148, 2004.

[20] J. Beane, P. Sebastiani, G. Liu, J. S. Brody, M. E. Lenburg, and A. Spira, "Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression," *Genome Biology*, vol. 8, no. 9, article R201, 2007.

[21] K. Steiling, J. Ryan, J. S. Brody, and A. Spira, "The field of tissue injury in the lung and airway," *Cancer Prevention Research*, vol. 1, no. 6, pp. 396–403, 2008.

[22] J. Hoeng, M. Talikka, F. Martin et al., "Toxicopanomics: applications of genomics, transcriptomics, proteomics and lipidomics in predictive mechanistic toxicology," in *Principle and Methods on Toxicology*, A. W. Hayes, Ed., Taylor & Francis, 2013.

[23] J. Hoeng, M. Talikka, F. Martin et al., "Case study: the role of mechanistic network models in systems toxicology," *Drug Discovery Today*, 2013.

[24] Selventa, "Reverse Causal Reasoning Methods Whitepaper," http://www.selventa.com/technology/white-papers.

[25] T. M. Thomson, A. Sewer, F. Martin et al., "Quantitative assessment of biological impact using transcriptomic data and mechanistic network models," *Toxicology and Applied Pharmacology*, 2013.

[26] W. K. Schlage, J. W. Westra, S. Gebel et al., "A computable cellular stress network model for non-diseased pulmonary and cardiovascular tissue," *BMC Systems Biology*, vol. 5, article 168, 2011.

[27] S. Gebel, R. B. Lichtner, B. Frushour et al., "Construction of a computable network model for DNA damage, autophagy, cell death, and senescence," *Bioinformatics and Biology Insights*, vol. 7, pp. 97–117, 2013.

[28] J. W. Westra, W. K. Schlage, B. P. Frushour et al., "Construction of a computable cell proliferation network focused on non-diseased lung cells," *BMC Systems Biology*, vol. 5, article 105, 2011.

[29] F. Martin, T. M. Thomson, A. Sewer et al., "Assessment of network perturbation amplitude by applying high-throughput data to causal biological networks," *BMC Systems Biology*, vol. 6, no. 1, article 54, 2012.

[30] P. H. Karp, T. Moniger, S. P. Weber et al., "An *in vitro* model of differentiated human airway epithelia: methods for establishing primary cultures," in *Epithelial Cell Culture Protocols*, C. Wise, Ed., vol. 188 of *Methods in Molecular Medicine*, chapter 11, pp. 115–137, 2002.

[31] C. Mathis, C. Poussin, D. Weisensee et al., "Human bronchial epithelial cells exposed *in vitro* to cigarette smoke at the air-liquid interface resemble bronchial epithelium from human smokers," *American Journal of Physiology: Lung Cellular and Molecular Physiology*, vol. 304, no. 7, pp. L489–L503, 2013.

[32] H. Maunders, S. Patwardhan, J. Phillips, A. Clack, and A. Richter, "Human bronchial epithelial cell transcriptome: gene expression changes following acute exposure to whole cigarette smoke *in vitro*," *American Journal of Physiology: Lung Cellular and Molecular Physiology*, vol. 292, no. 5, pp. L1248–L1256, 2007.

[33] A. A. Pezzulo, T. D. Starner, T. E. Scheetz et al., "The air-liquid interface and use of primary cell cultures are important to recapitulate the transcriptional profile of *in vivo* airway epithelia," *American Journal of Physiology: Lung Cellular and Molecular Physiology*, vol. 300, no. 1, pp. L25–L31, 2011.

[34] C. J. di Como, M. J. Urist, I. Babayan et al., "p63 expression profiles in human normal and tumor tissues," *Clinical Cancer Research*, vol. 8, no. 2, pp. 494–501, 2002.

[35] J. L. McQualter, K. Yuen, B. Williams, and I. Bertoncello, "Evidence of an epithelial stem/progenitor cell hierarchy in the adult mouse lung," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 4, pp. 1414–1419, 2010.

[36] S. Huang, L. Wiszniewski, and S. Constant, "The use of *in vitro* 3D cell models in drug development for respiratory diseases," in *Drug Discovery and Development—Present and Future*, I. M. Kapetanovic, Ed., chapter 8, InTech, 2011.

[37] S. Parrinello, J.-P. Coppe, A. Krtolica, and J. Campisi, "Stromal-epithelial interactions in aging and cancer: senescent fibroblasts alter epithelial cell differentiation," *Journal of Cell Science*, vol. 118, no. 3, pp. 485–496, 2005.

[38] Health Canada, "Determination of "tar", nicotine, and carbon monoxide in mainstream tobacco smoke," Official method T-115, 1999.

[39] R Development Core Team, *R: A Language and Environment For Statistical Computing*, 2009.

[40] R. Gentleman, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer Science, Business Media, New York, NY, USA, 2005.

[41] R. C. Gentleman, V. J. Carey, D. M. Bates et al., "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, article R80, 2004.

[42] R. A. Irizarry, B. Hobbs, F. Collin et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[43] F. Martin, "Systems and methods for network-based biological activity assessment," WO Patent 2,013,034,300, 2013.

[44] Y. Strulovici-Barel, L. Omberg, M. O'Mahony et al., "Threshold of biologic responses of the small airway epithelium to low levels of tobacco smoke," *American Journal of Respiratory and Critical Care Medicine*, vol. 182, no. 12, pp. 1524–1532, 2010.

[45] F. Schembri, S. Sridhar, C. Perdomo et al., "MicroRNAs as modulators of smoking-induced gene expression changes in human airway epithelium," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 7, pp. 2319–2324, 2009.

[46] L. Stejskalova, L. Vecerova, L. M. Peréz et al., "Aryl hydrocarbon receptor and aryl hydrocarbon nuclear translocator expression in human and rat placentas and transcription activity in human trophoblast cultures," *Toxicological Sciences*, vol. 123, no. 1, pp. 26–36, 2011.

[47] Y. Bosse, D. S. Postma, D. D. Sin et al., "Molecular signature of smoking in human lung tissues," *Cancer Research*, vol. 72, no. 15, pp. 3753–3763, 2012.

[48] S. Gebel, B. Gerstmayer, P. Kuhl, J. Borlak, K. Meurrens, and T. Müller, "The kinetics of transcriptomic changes induced by cigarette smoke in rat lungs reveals a specific program of defense, inflammation, and circadian clock gene expression," *Toxicological Sciences*, vol. 93, no. 2, pp. 422–431, 2006.

[49] J. O. Boyle, Z. H. Gümüş, A. Kacker et al., "Effects of cigarette smoke on the human oral mucosal transcriptome," *Cancer Prevention Research*, vol. 3, no. 3, pp. 266–278, 2010.

[50] S. D. Spivack, G. J. Hurteau, R. Jain et al., "Gene-environment interaction signatures by quantitative mRNA profiling in exfoliated buccal mucosal cells," *Cancer Research*, vol. 64, no. 18, pp. 6805–6813, 2004.

[51] D. Sidransky, "The oral cavity as a molecular mirror of lung carcinogenesis," *Cancer Prevention Research*, vol. 1, no. 1, pp. 12–14, 2008.

[52] J. Dionísio, "Diagnostic flexible bronchoscopy and accessory techniques," *Revista Portuguesa de Pneumologia*, vol. 18, no. 2, pp. 99–106, 2012.

[53] S. Anttila, J. Hakkola, P. Tuominen et al., "Methylation of cytochrome P4501A1 promoter in the lung is associated with tobacco smoking," *Cancer Research*, vol. 63, no. 24, pp. 8623–8628, 2003.

[54] V. Tamási, K. Monostory, R. A. Prough, and A. Falus, "Role of xenobiotic metabolism in cancer: involvement of transcriptional and miRNA regulation of P450s," *Cellular and Molecular Life Sciences*, vol. 68, no. 7, pp. 1131–1146, 2011.

[55] D. H. Phillips, A. Hewer, C. N. Martin, R. C. Garner, and M. M. King, "Correlation of DNA adduct levels in human lung with cigarette smoking," *Nature*, vol. 336, no. 6201, pp. 790–792, 1988.

[56] B. Schoket, D. H. Phillips, S. Kostic, and I. Vincze, "Smoking-associated bulky DNA adducts in bronchial tissue related to CYP1A1 MspI and GSTM1 genotypes in lung patients," *Carcinogenesis*, vol. 19, no. 5, pp. 841–846, 1998.

[57] L. Anna, K. Kovács, E. Gyorffy, B. Schoket, and J. Nair, "Smoking-related O^4-ethylthymidine formation in human lung tissue and comparisons with bulky DNA adducts," *Mutagenesis*, vol. 26, no. 4, pp. 523–527, 2011.

[58] E. Gyorffy, L. Anna, Z. Gyori et al., "DNA adducts in tumour, normal peripheral lung and bronchus, and peripheral blood lymphocytes from smoking and non-smoking patients: correlations between tissues and detection by ^{32}P-postlabelling and immunoassay," *Carcinogenesis*, vol. 25, no. 7, pp. 1201–1209, 2004.

[59] M. Lodovici, V. Akpan, L. Giovannini, F. Migliani, and P. Dolara, "Benzo[a]pyrene diol-epoxide DNA adducts and levels of polycyclic aromatic hydrocarbons in autoptic samples from human lungs," *Chemico-Biological Interactions*, vol. 116, no. 3, pp. 199–212, 1998.

[60] R. W. L. Godschalk, D. E. M. Feldker, P. J. A. Borm, E. F. M. Wouters, and F.-J. van Schooten, "Body mass index modulates aromatic DNA adduct levels and their persistence in smokers," *Cancer Epidemiology Biomarkers & Prevention*, vol. 11, no. 8, pp. 790–793, 2002.

[61] A. Besaratinia, L. M. Maas, E. M. C. Brouwer, J. C. S. Kleinjans, and F. J. van Schooten, "Comparison between smoking-related DNA adduct analysis in induced sputum and peripheral blood lymphocytes," *Carcinogenesis*, vol. 21, no. 7, pp. 1335–1340, 2000.

[62] S. Pavanello, A. Pulliero, B. O. Saia, and E. Clonfero, "Determinants of anti-benzo[a]pyrene diol epoxide-DNA adduct formation in lymphomonocytes of the general population," *Mutation Research*, vol. 611, no. 1-2, pp. 54–63, 2006.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 409658, 7 pages
http://dx.doi.org/10.1155/2013/409658

Research Article

An Accurate Method for Prediction of Protein-Ligand Binding Site on Protein Surface Using SVM and Statistical Depth Function

Kui Wang,[1] Jianzhao Gao,[1] Shiyi Shen,[1] Jack A. Tuszynski,[2] Jishou Ruan,[1] and Gang Hu[1]

[1] *College of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China*
[2] *Division of Experimental Cancer, Cross Cancer Institute, 115660 University Avenue, Edmonton, AB, Canada T6G 2V4*

Correspondence should be addressed to Gang Hu; huggs.hg@gmail.com

Received 17 May 2013; Revised 15 August 2013; Accepted 29 August 2013

Academic Editor: Bing Niu

Since proteins carry out their functions through interactions with other molecules, accurately identifying the protein-ligand binding site plays an important role in protein functional annotation and rational drug discovery. In the past two decades, a lot of algorithms were present to predict the protein-ligand binding site. In this paper, we introduce statistical depth function to define negative samples and propose an SVM-based method which integrates sequence and structural information to predict binding site. The results show that the present method performs better than the existent ones. The accuracy, sensitivity, and specificity on training set are 77.55%, 56.15%, and 87.96%, respectively; on the independent test set, the accuracy, sensitivity, and specificity are 80.36%, 53.53%, and 92.38%, respectively.

1. Introduction

With the development of the technology to solve the protein structures, the rate of deposition of protein structure in the PDB [1] grows very fast. Unfortunately, there are still a lot of resolved proteins with unknown function. For the proteins always carry out their functions through interactions with other molecules, such as other proteins, peptides, nucleotides, compounds, and so forth, identifying the residues involved in these interactions is an important step towards characterizing protein function. The protein functional sites consist of various types of binding site including protein-ligand binding site, protein-protein binding site, and protein-DNA binding site. Since the ligands (here we refer to small organic compounds as ligands) constitute most of the drugs approved by FDA [2], the prediction of protein-ligand binding site also plays an important role in rational drug discovery. Therefore, we focus on the protein-ligand binding site in this study.

In the past 15 years, many methods were developed to predict protein-ligand binding sites. There are mainly two type of methods, geometry- and energy-based methods. The energy-based methods identify the binding site using model of energetics [3–6], such as PocketFinder [5] and Q-SiteFinder [6]. Most of the algorithms are based on geometry, for the binding sites always locate on the concave surface, which likes a pocket or a cleft. Actually, many methods were proposed to detect the pocket on protein surface using geometric criteria, and the pocket with largest volume is often returned as prediction of binding surface. POCKET [7] and Ligsite [8] map the entire protein structure to 3D grid and cluster the grids with special event, like protein-solvent-protein event or surface-solvent-surface event. Surfnet [9] places empty spheres which separate any two atoms of protein, and cluster these spheres to describe the pocket. PASS [10] uses sphere probes to fill the cavities layer by layer and detect the pocket. CASTp [11] applies the alpha shape theory [12] from computational geometry to detect and measure the pockets. The approaches mentioned above often use pure geometric criteria without additional information like conservation or physic-chemical information. Ligsitecsc [13], ConSurf [14], and ConCavity [15] made a big progress after combining the evolutionary conservation to pocket detections.

Besides methods based on geometry or energy, researchers also developed algorithms based on machine

learning model for functional sites prediction. Zhang et al. [16] predict catalytic site using SVM based on sequence information. Ansari and Raghava [17] identify the NAD interacting residues in proteins using SVM. Cheng et al. [18] also use SVMs to predict RNA-binding sites of proteins. Ofran and Rost [19] use neural networks to identify the protein-protein interaction sites. In contrast to the geometry- or energy-based algorithm, the prediction made by machine learning always employed only sequence information and a few structural information like content of secondary structure.

Our study in this paper has two motivations. First, although existent geometry-based methods provided considerable accurate prediction of binding site, improvement could be made by integrating other information, such as evolutionary conservation. Many pure geometry-based methods only return the largest pocket, which is not always true for ligand-binding pocket. Ligsitecsc and ConSurf rerank the pockets by conservation and perform better than previous methods, which rank the pockets by volume. It implies that additional information should be took into account to bind site prediction. On the other hand, machine learning based approaches only focus on sequence information. Here we want to build a method based on comprehensive features which are available and useful for identifying binding site. Not only sequence information but also structural information will be combined together. Second, it is inherently difficult to define negative samples no matter which method employed. It is easy to define positive samples from the interactions of protein-ligand complexes which are experimentally approved. Not all interactions are known, so it is hard to say which residue cannot bind ligands. In this paper, we introduce statistical depth function (details in Section 2) to define the negative samples. The idea is from intuition that binding residues are always locate on concave protein surface and the residues on convex surface are unlikely (not impossible) to bind ligands.

In this paper we present a novel method base on SVM model integrating both sequence and structural information to predict protein-ligand binding site. To test and validate our method, a benchmark dataset including 373 complexes is built from PDBbind [20]. The validation on the independent test set shows that the accuracy of our method is about 80.36%. For the top 3 pockets provided by Ligsitecsc and CASTp, we rerank them according to our prediction. Then top1 success rate is improved from 41.6% to 75.3% (for Ligsitecsc) and from 61.0% to 77.9% (for CASTp), respectively.

2. Material and Methods

2.1. Dataset. The PDBbind [20] database provides a collection of experimentally measured binding affinity data exclusively for the biomolecular complexes available in the PDB. We select the "refined set" in PDBbind, which is compiled to provide a high quality set of protein-small ligand complexes. There are 1741 entries in the "refined set" of PDBbind. After removing the redundancy complexes with more than 30% sequence similarity, 373 nonredundant complexes are remained as our dataset.

We divided these 373 entries into two datasets randomly, one is training set and the other is test set. The training set consists of 296 complexes, and the test set contains 77 complexes. The training set is divided into five subsets (each one has about 60 proteins) randomly for 5-fold cross-validation. The PDB ids of the training set and test set are shown in Support Information STable 1, available online at http://dx.doi.org/10.1155/2013/409658.

2.2. Statistical Depth Functions. We employ the statistical depth function to measure the depth of the residue on the protein surface. The statistical depth function gives the residues in the pocket deeper depth values; and for the residues on convex protein surface, it gives them lower depth values. Statistical depth functions assign a point its degree of centrality with respect to a dataset. They are "order statistics" in higher dimension space (≥2). In statistics, statistical depth functions have become increasingly pursued as a useful tool in nonparametric inference for multivariate data. The statistical depth functions have been used to measure the residue depth and analyze the protein structure. There are several statistical depth functions to measure the degree of centrality of a point. In this study, we use half-space depth function to measure the depth of the residues because the concept and the definition of the half-space depth are simple and easy to implement.

Half-Space Depth. Tukey [21] introduced the half-space depth to order the high dimensional data. The half depth (HD) of a point x in R^d with respect to a probability measure P on R^d is defined as the minimum probability mass carried by any closed half space containing x; that is,

$$\mathrm{HD}(x, P) = \inf\{P(H) : H \text{ is a closed half space}, \ x \in H\}, \quad x \in R^d. \tag{1}$$

For a probability measure P, the half-space depth of any point in R^d with respect to P can be defined. For a dataset, such as all the atoms in a protein, we can use the empirical distribution to estimate the probability $P(H)$. We consider every point in dataset is equiprobable; then $P(H) = \sum P(x), x$ in H, where $P(x) = 1/n, n$ is the number of the points in dataset. For simplifying, we define $P(x) = 1$; then $P(H)$ represents the number of the points in H. In addition, we define H by an open half space. Thus, the depth of the points at the boundary of dataset is zero and the point which has the maximal depth value is the center of the dataset. Figure 1 shows some examples in 1 and 2 dimensions. Figure 2 shows the depth function is applied to measure the depth of protein atoms.

Depth and Relative Accessible Surface Area (RSA). According to the definition of half-space depth above, the depth values of buried residues are always greater than zero. And the residues with small depth values locate on the convex of the protein. It is notable that not all the residues with depth values greater than zero are buried, that is, the depth values of the points b, c, d, and e in Figure 1(d). These residues locate in the "pockets" on the protein surface and their RSAs are greater than 0, too. On the other hand, the residues the depth values of which are

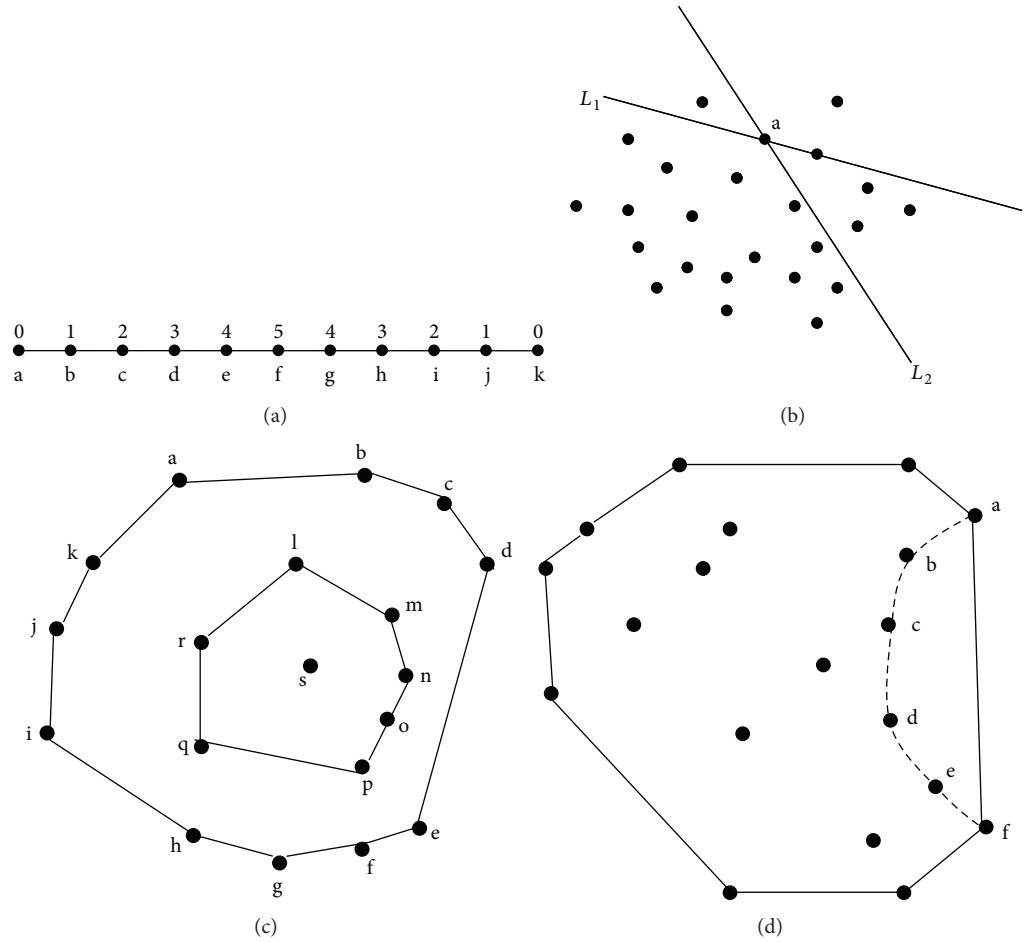


Figure 1: An illustration of the concept of half-space depth for 1- and 2-dimensional datasets.

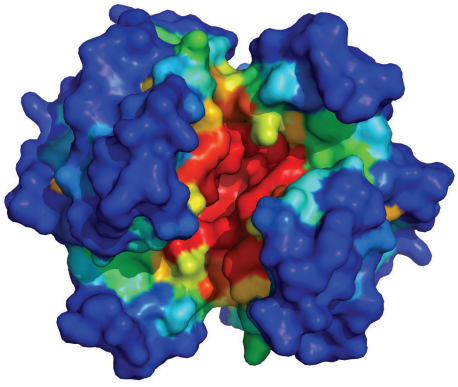


Figure 2: An illustration of the half-space depth function. The protein surface (PDBID: 10 gs) is colored by the statistical depth value of the residues. The color is gradually change from blue to red according to the depth values of the residues. The deepest residues are red, and the convex residues are blue. All the figures of the protein surface are created by Pymol.

zero and RSAs of which are greater than zero will locate on the protein surface and will not locate in "pockets" on the protein surface. In addition, using both depth and RSA, the residues in the pockets can be found easily, too. Thus, the residues on the protein surface can be divided into two types according to the RSA and depth value. The first class includes the residues whose RSAs are greater than 0 and depths close to 0, which means these residues locate on the convex of the protein. And the second class includes the residues whose RSAs and depths are both greater than 0, which means the residues in this class locate in pockets on the protein surface.

2.3. Sample Selection. Firstly, we introduce some concepts to define the samples. We remove the buried residues and only consider the residues on the protein surface. The residue is considered as being on the surface if its RSA is greater than 10%. For a given residue on the surface, its neighbors are the residues on the surface with distance to the given residue <10 angstroms. We call a residue on the protein surface and its neighbors together a patch. As a result, each residue on protein surface has a patch. A sample just refers to a residue on protein surface or its corresponding patch. In this study, we have three types of samples: positive samples, negative samples, and the others (we call this kind of samples as not positive and not negative samples, NP & NN for short).

For each residue on the surface, if the distance between any nonhydrogen atom of the residue and any atom of the

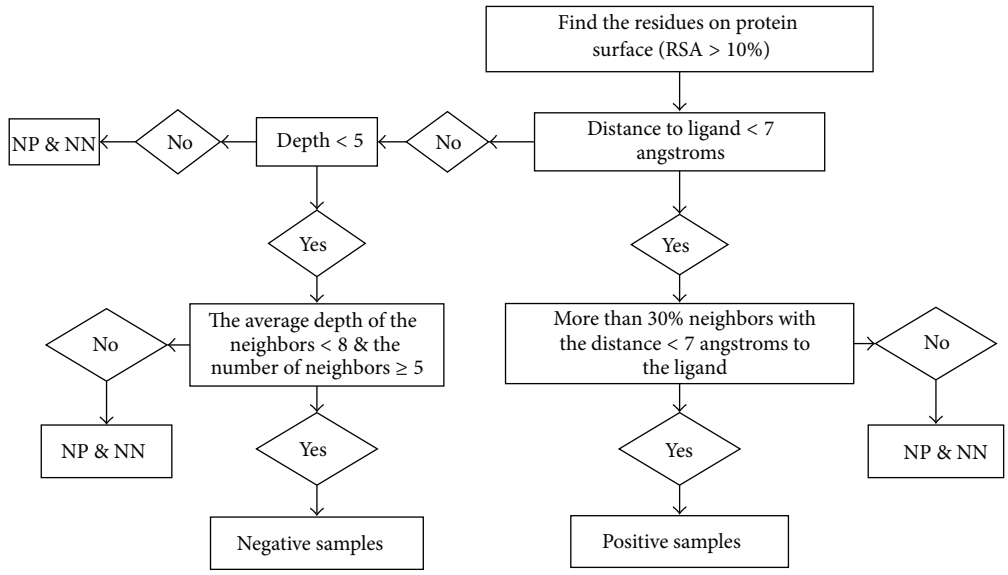


Figure 3: The framework of samples selection.

ligand is less than 7 angstroms, this residue is kept as a positive sample candidate; otherwise, the residue is kept as a negative sample candidate. For any positive sample candidate, if more than 30% neighbors of this residue are the positive sample candidates, the sample is considered as a positive sample; otherwise, it is considered as a NP & NN sample.

The negative samples are not easy to define like positive samples. The proteins of complexes might bind some ligands somewhere and we cannot distinguish the potential functional binding residues from the others. We consider the fact that the residues which locate on the protein surface and not in a "pocket" or "cavity" are hard to bind a ligand. So the residues on the convex protein surface are regarded as negative samples in our study. Here we use statistical depth function (which is defined above) to distinguish whether the residue is or is not on the convex surface. Firstly, we remove the residues with half-space depth greater than 5 from the negative sample candidates. The negative sample candidate is defined as negative sample if the average depth value of its neighbors is less than 8 and the number of its neighbors is greater than 5. Figure 3 shows the framework to select the samples.

Here, the standard of negative samples is a little strict. There are two reasons. One is that we want to make sure our negative samples would not bind to a ligand. The other reason is that the definition can balance the positive samples and negative samples to avoid the training bias. After the samples selection, the radio of positive samples to negative samples is about 1 : 2 in the training set and 1 : 3 in the test set.

2.4. Feature Design. There are totally 330 features for each sample. These features include two-aspect information of protein: global information and local information. We explain every feature as follows.

Global Information. The global information of the sample includes the length of the protein sequence, the distance to P terminal, the distance to N terminal, the residue components, and the global secondary structure content. The secondary structures of proteins are calculated by DSSP [22]. We use a four-dimensional binary vector to represent the length and the two distances, which are discredited in four intervals [0, 60), [60, 120), [120, 240), and ≥240. A 20-dimensional vector and a 3-dimensional vector are used for residue components and secondary structure content, respectively. So we have $4 * 3 + 20 + 3 = 35$ features of global information.

Local Information. The local information of the sample comes from two aspects: sequence information and structure information. A sliding window is used for the residue to describe the properties of the neighbors on the protein sequence. Each position in the sliding window includes 30 features. The first 20 features are from position (PSSM-specific scoring matrix) and the 21th is conservation score. The 22th–24th are the 3-dimensional binary vector for the secondary structure of the residue. Then the 25th–29th are positive charge, negative charge, PI value, polarity, and hydrophobicity. The last one is an indicator. If the data do not exist (the side is out of window), all the features are zero and the last feature is assigned 1. We use a window with length 9; thus we have $30 * 9 = 270$ features about sequence information for the central residue in the sliding window.

The structure information of the sample comes from the patch of this sample. The patch includes the central residue and its neighbors. Because the number of neighbors is not equal, it is impossible to describe every neighbor as the features. As a simplification, we use the minimum, maximum, and the average of the patch properties, which include positive charge, negative charge, PI vale, polarity, hydrophobicity, hydrogen bond tendency, the conservation score, and the secondary structure content. Besides the hydrogen bond tendency, which we use four features to describe (minimum, maximum, average, and the value of the central residue), the secondary structure content needs 3 features to describe and

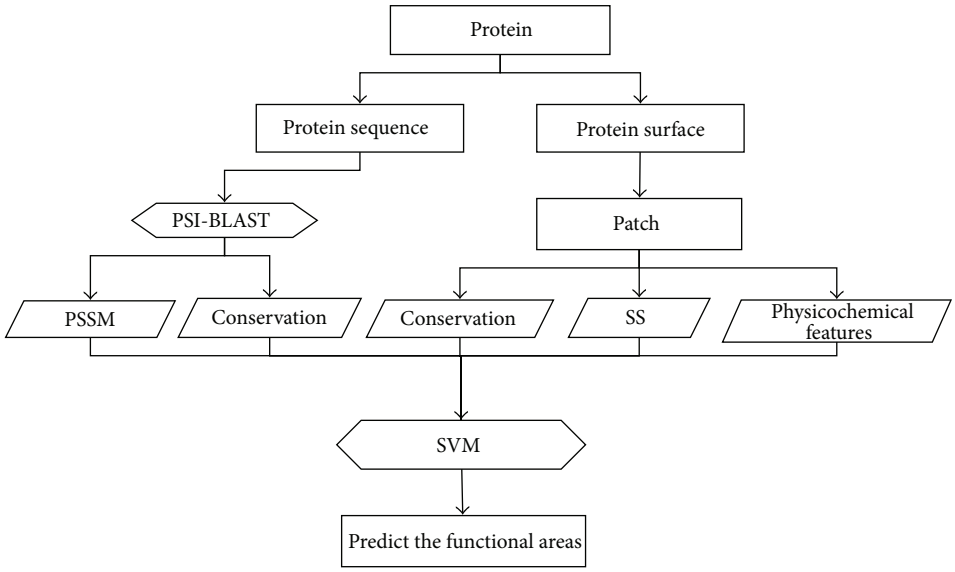


Figure 4: Framework of the proposed prediction system.

other properties also have three values (minimum, maximum and average of the patch); the total features of local structure information are $4 + 6 * 3 + 3 = 25$.

Totally, we have $35 + 270 + 25 = 330$ features for each sample.

The charges, polarity, and hydrophobicity of amino acid are extracted from AAIndex [23] database. Sarkhel and Desiraju [24] calculated the frequency distribution of hydrogen bonds for amino acids which act as donors or acceptors between proteins and ligands. We use these frequencies as hydrogen bond tendency. The conservation score of the residue is calculated using Shannon Entropy from PSI-blast [25] profile.

2.5. Perform the Training and Prediction. As we mentioned in dataset subsection, we divide the dataset into training set and test set. Every feature is normalized to $[0, 1]$. Using Libsvm [26] package and selecting RBF (Radial Basis Function) as the kernel function of SVM, we select 30 proteins from training set to train the SVM parameters using a grid method. Then we train the SVM with the best parameters on training set and predict the functional surfaces on test set. On the training set, we do 5-fold cross-validation to train our model. Figure 4 shows the framework of the proposed prediction system.

3. Results and Discussions

3.1. Binding Sites Predictions. Using Libsvm package and selecting RBF as the kernel function of SVM, we train SVM model based on the sequence and structural features mentioned above. Totally, we get 15385 positive samples and 31663 negative samples based on the training set (the ratio between positive samples and negative samples is about 1 : 2). Similarly, we get 3510 positive samples and 12201 negative samples based on the test set (the ratio is about 1 : 3).

The accuracy, precision, sensitivity, specificity, and MCC are used to validate the performance of our method. Then these five indices of training set and the test set are shown in Table 1. We do 5-fold cross-validation to avoid overfitting problem. And the accuracy of the train set is shown similar to the accuracy of the test set. It can also be observed in the case of precision, sensitivity, specificity, and MCC.

CASTp returns the exact binding residues and the residues forming the pocket mouth. We can compare the binding prediction between our method and CASTp directly, which is shown in Table 1. Our method clearly outperforms CASTp.

Unfortunately different approaches always return the predictions in different ways. For example, Ligsitecsc only returns the geometry center coordinates of the pocket. To compare our method with other algorithms, we have to evaluate them using the same standards. Therefore, we define the residues close enough to the pocket center as the binding site/binding residues returned by Ligsitecsc. The distance threshold to pocket center ranges from 1 A to 75 A, because the binding residues will not increase after the threshold greater than 75 A. We can also obtain the ROC curve of the Ligsitecsc (Figure 5) by this way. Since the SVM model returns a real value range from 0 to 1, the ROC curve of our method is also shown in Figure 5. From the comparison of ROC curves, our method performs much better than Ligsitecsc. The AUC value of our method is 0.71 greater than 0.63 which is the AUC vale of Ligsitecsc.

3.2. Rerank the Pockets. Although our method can identify the binding sites more accurately than CASTp and Ligsitecsc, we cannot provide a pocket-level comparison. Unlike CASTp and Ligsitecsc, we do not detect pocket. However, we can also apply our model in the postprocess of geometry-based models to improve the rank of the pockets. The motivation is that using conservation score not pocket volume has made a big progress in the search of pocket.

We focus on the top 3 pockets obtained from CASTp and Ligsitecsc. For Ligsitecsc, it only gives three pockets per

Table 1: The values of five indices on the training set and test set for well-trained SVM.

Dataset	Accuracy	Precision	Sensitivity	Specificity	MCC
Train set	77.55%	69.38%	56.15%	87.96%	**49.68%**
Test set (our method)	**80.36%**	**75.88%**	**53.53%**	**92.38%**	**50.85%**
Test set (CASTp)	**56.49%**	**20.81%**	**41.67%**	**60.21%**	**2.0%**

Table 2: Top n success rates for test set.

Methods	Top 1	Top 2	Top 3
Ligsitecsc	40.3% (31/77)	71.4% (55/77)	83.1% (64/77)
Ligsite + our method	**75.3% (58/77)**	**81.8% (63/77)**	**83.1% (64/77)**
CAStp	60.0% (47/77)	88.3% (68/77)	92.2% (71/77)
CAStp + our method	**77.9% (60/77)**	**90.9% (70/77)**	**92.2% (71/77)**

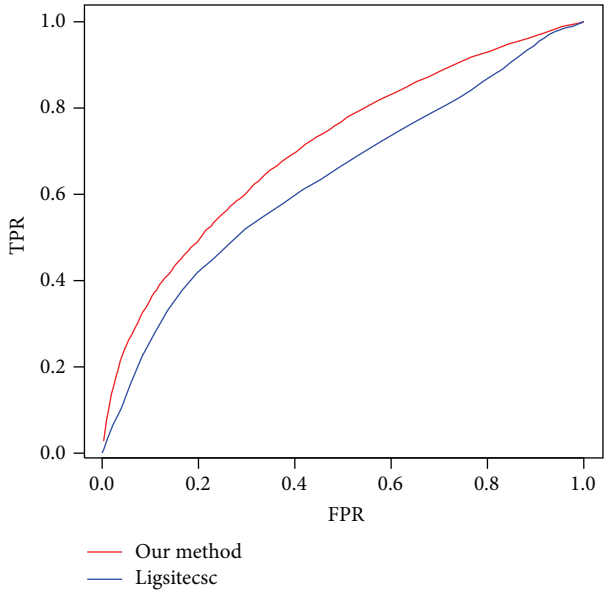


Figure 5: The ROC curves of our method and Ligsitecsc. The blue line is ROC curve of Ligsitecsc, and the red line is of our method.

protein and ranks them using the conservation score. The top 1 pocket in Ligsitecsc means the most conservative pocket in three of them. The CASTp provides more pockets than Ligsitecsc, but the pockets are not ranked and many of them are too small to bind ligand. We rank them using volume and select the three largest ones as the top 3 pockets returned by CASTp. The pocket from prediction is considered as a functional or binding pocket when the distance between the pocket geometry center and any atom of ligands is less than 8 A. We define the top n ($n = 1, 2, 3$) success rates as the number of functional pockets in top n divided by the number of proteins. On our test set, we have 77 proteins and 231 prediction pockets in total.

Our method is combined to pocket detection algorithm by a simple way. For each pocket from prediction, we only count the residues which are predicted as binding sites by our SVM model. Then this number is used to rank the pocket. The result is shown in Table 2. Top 3 success rates among the four methods are the same according to the definition, but the top 1 success rates are totally different. The methods combined with our model gain a much improvement from the CASTp or Ligsitecsc. It implies our method included the complementary information for CASTp and Ligsitecsc. And the model can be applied in other geometry-based pocket detection methods.

4. Conclusion

In this paper, we introduce statistical depth function to define the negative sample of the protein-ligand binding site. The further analysis shows negative sample defined by this way is reasonable and helpful for the model training (shown in Support Information Sections 2–4). Then we propose an SVM model including sequence and structural information; the results show the method significantly outperforms the existing methods based on pure geometry or only combining evolutionary conservation. Our method can also provide the complementary information for geometry-based methods such as CASTp and Ligsitecsc in the postprocess.

Acknowledgments

Jianzhao Gao, Kui Wang, and Gang Hu were supported by Fundamental Research Funds for the Central Universities (Grants nos. 65011491, 65011521, and 65011441); Gang Hu was also supported by NSFC (Grant no. 11101226); Shiyi Shen was supported by NSFC (Grant no. 20836005); Jishou Ruan was supported by the International Development Research Center, Ottawa, Canada (no. 104519-010), NSFC (Grants nos. 31050110432, and 31150110577), and Tianjin Science and Technology Support Project 08ZCHHZ00200. Jack A. Tuszynski acknowledges generous support for this research provided by NSERC, Alberta Cancer Foundation, the Allard Foundation, the Canadian Breast Cancer Foundation, and the Alberta Advanced Education and Technology.

References

[1] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[2] D. S. Wishart, C. Knox, A. C. Guo et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, no. 1, pp. D901–D906, 2008.

[3] A. H. Elcock, "Prediction of functionally important residues based solely on the computed energetics of protein structure," *Journal of Molecular Biology*, vol. 312, no. 4, pp. 885–896, 2001.

[4] P. Bate and J. Warwicker, "Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods," *Journal of Molecular Biology*, vol. 340, no. 2, pp. 263–276, 2004.

[5] J. An, M. Totrov, and R. Abagyan, "Pocketome via comprehensive identification and classification of ligand binding envelopes," *Molecular and Cellular Proteomics*, vol. 4, no. 6, pp. 752–761, 2005.

[6] A. T. R. Laurie and R. M. Jackson, "Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites," *Bioinformatics*, vol. 21, no. 9, pp. 1908–1916, 2005.

[7] D. G. Levitt and L. J. Banaszak, "POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids," *Journal of Molecular Graphics*, vol. 10, no. 4, pp. 229–234, 1992.

[8] M. Hendlich, F. Rippmann, and G. Barnickel, "LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins," *Journal of Molecular Graphics and Modelling*, vol. 15, no. 6, pp. 359–363, 1997.

[9] R. A. Laskowski, "SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions," *Journal of Molecular Graphics*, vol. 13, no. 5, pp. 323–330, 1995.

[10] G. P. Brady Jr. and P. F. W. Stouten, "Fast prediction and visualization of protein binding pockets with PASS," *Journal of Computer-Aided Molecular Design*, vol. 14, no. 4, pp. 383–401, 2000.

[11] T. A. Binkowski, S. Naghibzadeh, and J. Liang, "CASTp: computed atlas of surface topography of proteins," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3352–3355, 2003.

[12] H. Edelsbrunner and E. Mucke, "Three-dimensional alpha shapes," *ACM Transactions on Graphics*, vol. 13, pp. 43–72, 1994.

[13] B. Huang and M. Schroeder, "LIGSITEcsc: predicting ligand binding sites using the connolly surface and degree of conservation," *BMC Structural Biology*, vol. 6, article 19, 2006.

[14] M. Landau, I. Mayrose, Y. Rosenberg et al., "ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures," *Nucleic Acids Research*, vol. 33, no. 2, pp. W299–W302, 2005.

[15] J. A. Capra, R. A. Laskowski, J. M. Thornton, M. Singh, and T. A. Funkhouser, "Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure," *PLoS Computational Biology*, vol. 5, no. 12, Article ID e1000585, 2009.

[16] T. Zhang, H. Zhang, K. Chen, S. Shen, J. Ruan, and L. Kurgan, "Accurate sequence-based prediction of catalytic residues," *Bioinformatics*, vol. 24, no. 20, pp. 2329–2338, 2008.

[17] H. R. Ansari and G. P. S. Raghava, "Identification of NAD interacting residues in proteins," *BMC Bioinformatics*, vol. 11, article 160, 2010.

[18] C.-W. Cheng, E. C.-Y. Su, J.-K. Hwang, T.-Y. Sung, and W.-L. Hsu, "Predicting RNA-binding sites of proteins using support vector machines and evolutionary information," *BMC Bioinformatics*, vol. 9, supplement 12, article S6, 2008.

[19] Y. Ofran and B. Rost, "ISIS: interaction sites identified from sequence," *Bioinformatics*, vol. 23, no. 2, pp. e13–e16, 2007.

[20] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, and S. Wang, "The PDBbind database: methodologies and updates," *Journal of Medicinal Chemistry*, vol. 48, no. 12, pp. 4111–4119, 2005.

[21] J. W. Tukey, "Mathematics and picturing data," in *In Proceedings of International Congress of Mathematicians*, vol. 2, pp. 523–531, Vancouver, Canada, 1975.

[22] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[23] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "AAindex: amino acid index database, progress report 2008," *Nucleic Acids Research*, vol. 36, no. 1, pp. D202–D205, 2008.

[24] S. Sarkhel and G. R. Desiraju, "N–H. . .O, O–H. . .O, and C–H. . .O hydrogen bonds in protein-ligand complexes: strong and weak interactions in molecular recognition," *Proteins*, vol. 54, no. 2, pp. 247–259, 2004.

[25] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[26] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 687607, 13 pages
http://dx.doi.org/10.1155/2013/687607

Research Article

Computer-Assisted System with Multiple Feature Fused Support Vector Machine for Sperm Morphology Diagnosis

Kuo-Kun Tseng,[1] Yifan Li,[1] Chih-Yu Hsu,[2] Huang-Nan Huang,[3] Ming Zhao,[1] and Mingyue Ding[4]

[1] *Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen Graduate School, Shenzhen, Guangdong 518055, China*
[2] *Department of Information and Communication Engineering, Chaoyang University of Technology, Taichung 41349, Taiwan*
[3] *Department of Mathematics, Tunghai University, Taichung 40704, Taiwan*
[4] *Huazhong University of Science and Technology, Wuhan 430074, China*

Correspondence should be addressed to Chih-Yu Hsu; tccnchsu@gmail.com

Received 15 April 2013; Revised 4 July 2013; Accepted 23 July 2013

Academic Editor: Lei Chen

Sperm morphology is an important technique in identifying the health of sperms. In this paper we present a new system and novel approaches to classify different kinds of sperm images in order to assess their health. Our approach mainly relies on a one-dimensional feature which is extracted from the sperm's contour with gray level information. Our approach can handle rotation and scaling of the image. Moreover, it is fused with SVM classification to improve its accuracy. In our evaluation, our method has better performance than the existing approaches to sperm classification.

1. Introduction

With the development of modern computer technology, medical imaging has played an important role in clinical diagnosis and treatment. Medical image analysers are facing the challenge of precisely extracting information from the medical image with the help of computer-assisted systems. Since people have become more and more concerned about the health of the next generation, morphology would be one important technique to identify the health of sperms. To examine whether or not the sperms are healthy, it is essential to inspect the sperms to assess their appearance. Currently, sperm quality is mostly judged by experts and doctors. Because of the numerous types of sperm shape, the efficiency and accuracy relying on human assessment are not ideal. As computer morphology technology develops, quantitative analysis of sperm morphology is demanded to assist doctors in their diagnoses. Thus, this research is intended to design a helpful sperm classification system.

Sperm morphology is an image classification problem in sperm imaging. It first detects a segment of the sperm image, after which feature extraction and analysis is possible, for example, sperm length, width, and size, followed by further classification according to sperm features [1]. As a result, solving the problem of sperm image recognition and classification can be valuable for aspects of sperm diagnosis.

Our sperm morphology system is equipped with a microscope connected to a computer to observe the real-time sperm image. The microscope helped us to take photos of the sperm images and input them into our computer. With the input we managed to obtain all the results and conclusions. The system and its equipment are shown in Figure 1.

In addition to system implementation, this research has made the following contributions.

(1) We proposed two approaches to transform the sperm contour into a one-dimensional waveform as an analysis feature. The first algorithm takes advantage of the distance between two points on the edge to produce a waveform. The second computes the distance from the geometric centre to the edge as the vertical value of the waveform.

(2) After extraction, we proposed an SVM classification on these waveforms with rank and grey level features.

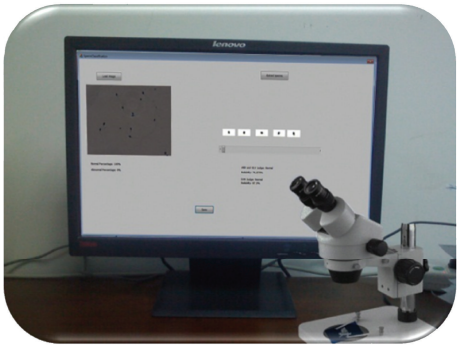


Figure 1: The sperm morphology diagnosis system with a microscope.

According to our comprehensive survey, this has not yet been used in sperm classification.

(3) We also conducted a complete comparison. We compared our approaches with the K-nearest neighbour, Scale-Invariant Feature Transform (SIFT), and the elliptic model. The experiment results show better performance than previous methods.

In our evaluation, we applied our approach to a sperm database. The results show that our idea is feasible and gives better performance than the existing approaches.

The rest of this paper comprises four parts. Section 2 introduces some other research studied to help our work. Section 3 proposes the architecture and algorithm, introducing the details of our algorithm which leads to a more detailed understanding of our approach. Section 4 provides the results and discussion proving that our work is feasible, and finally, the conclusion is given in Section 5.

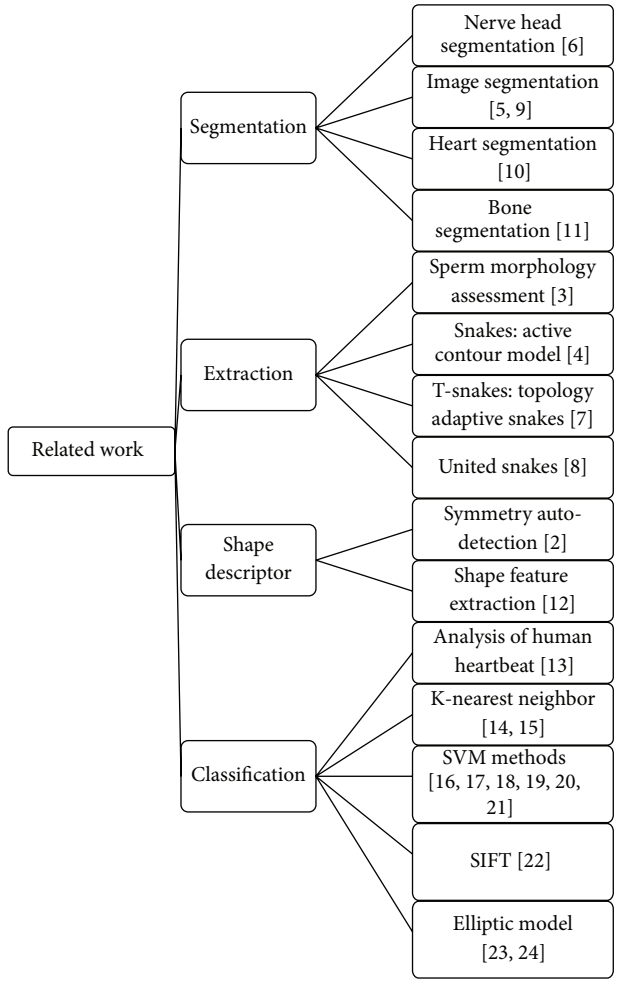


Figure 2: Related work.

2. Related Works

For this research on sperm morphology, we reviewed the related works on segmentation, extraction, shape descriptor, and the classification algorithm, as shown in Figure 2.

2.1. Segmentation. First, segmentation takes place, so we have to look at several pieces of the literature [2–6] related to image segmentation. The first article [3] presents a method for optic nerve head segmentation and its validation. The method is based on the Hough transform and anchored active contour model. The results were validated by comparing the performance of different classifiers and showed that this approach is suitable for automated diagnosis of and screening for glaucoma.

Considering that there is no guarantee that the sperms we observed will appear with the posture and position we need, it is absolutely necessary for us to investigate how to deal with active contours. In research [2], they introduce a geometrical, variation frame that uses active contours to segment and obtain features from images at the same time.

To obtain a better result of image segmentation, paper [4] enhanced the lane for interactive image segmentation by incremental path map construction, a modified version of the live lane that can extract objects from an image interactively with efficiency and repeatability. It guarantees a strictly bounded response time and follows the target boundary with little digression.

Furthermore, in paper [5], a new method was proposed for the local assessment of boundary detection by a simulated search. Its boundary detection can be optimized per landmark during model training. The success of the method was demonstrated for cardiac image segmentation and it was shown to improve the capture range and accuracy of the boundary detection. Another paper, [6], evaluated various image features and different search strategies for fitting active shape models (ASMs) to bone object boundaries in digitized radiographs. It proposed an improved search procedure that is more robust against outlier configurations in the boundary target points.

2.2. Extraction. The next topic is extraction. Because the results of the sperm-head contour extraction have an essential influence on the classification, we studied some issues which provide further information. As it is of great importance to obtain the sperm-head contour precisely, we studied articles on how to abstract contours. The first was "Sperm

morphology assessment using David's classification" [7]. This paper aimed to compare assessment of sperm morphology by using David's classification (DC) based on the strict criteria (computer-assisted sperm analysis (CASA) SC) for their ability to predict fertilization in a selected in vitro fertilization (IVF) population. Their results showed that the DC sperm morphology analysis was less indicative of fertilization than CASA SC.

After the DC analysis, we reviewed the paper on the technology of extracting objects' edges. The first article [8], "Snakes," is an energy-minimizing spline guided by external constraint forces and influenced by image forces that pull it towards features such as lines and edges. Snakes are active contour models; they lock onto nearby edges, localizing them accurately. Scale-space continuation can be used to enlarge the capture region surrounding a feature. Snakes provide a unified account of a number of visual problems, including detection of edges, lines, and subjective contours; motion tracking; and stereo matching. We have used snakes successfully for interactive interpretation, in which user-imposed constraint forces guide the snake near features of interest.

As the snake may not do its job well enough for current research, we reviewed some improved algorithms such as T-snakes [9], topology adaptive snakes. In this paper, they present a new class of deformable contours (snakes) and are applied to the segmentation of medical images. They enable topological flexibility among other features. The resulting topology adaptive snakes, or "T-snakes," can be used to segment some of the most complex-shaped biological structures from medical images in an efficient and highly automated manner.

Moreover, other authors [10] present a framework called united snakes, which has two key features. First, it unifies the most popular snake variants, expanding the range of object modelling capabilities. Second, it embodies the idea of the technique known as live wire or intelligent scissors. The two techniques can be combined advantageously by introducing an effective hard constraint mechanism. They apply united snakes to several different medical image analysis tasks, demonstrating the generality, accuracy, and robustness of the tool.

2.3. Shape Descriptor. The third topic, the most related work, is called shape descriptor. We focused on how to transform it into a one-dimensional feature. To achieve the goal we studied further related articles. The first, paper [11], presents a new symmetry autodetection approach. The symmetry can be detected automatically by using corner detection. During the process, the contour can be transferred to a waveform.

For additional study of the shape descriptor, paper [12] focused on presenting the existing approaches of shape-based feature extraction. Paper [13] introduced the extraction of waveform features by reduced binary features, used to reduce complexity and storage.

2.4. Classification. Another common image matching approach, the K-nearest neighbour method as a compared target, proposes a method to fuse real-value K-nearest neighbour classifiers by feature grouping [14]. The real-value K-nearest neighbour classifier can approximate continuous-valued target functions. In addition, it is sensitive to feature perturbation. Therefore, when the multiple real-value K-nearest neighbour classifier is fused by feature grouping, the performance of the fusion will be better than the single classifier. Another K-nearest neighbour method [15] presents a novel improvement to the K-nearest neighbour mean classifier (K-NNMC). K-NNMC finds the K-nearest neighbours for each class of training patterns separately and finds the means for each of these K-neighbours (class-wise). Classification is undertaken according to the nearest mean pattern. In experiments using several standard datasets, it has been shown that the proposed classifier provides better classification accuracy over the conventional K-nearest neighbour method, and thus, it is a suitable method to be used in data mining applications.

As we have been using the SVM as an advanced method to improve the performance of our approach, in order to achieve a deep understanding of SVM, we paid attention to related SVM research. Research work [16] presents a new valid edge detection algorithm based on an SVM to avoid the disadvantages of traditional image edge detection methods. In another SVM related work [17], the authors compared the performance of artificial-immune-system- (AIS) based algorithms to a Gaussian kernel-based SVM. Their experimentation indicates that the AIS-based classification paradigm has the intrinsic property of dealing more efficiently with highly skewed datasets. In addition, research [18] takes advantage of SVM to characterize the sperm population structure related to freezability. The SVM was generated using sperm motility information captured by CASA from thawed semen. This SVM method was used to characterize the motile sperm subpopulations for Iberian red deer.

Research [19] provided more information on the effect of SVM on sperm research. An automated, quantitative method that objectively classifies five distinct motility patterns of mouse sperm using the SVM method, was developed. Its parameters are associated with the classified tracks and were incorporated into established SVM algorithms to generate a series of equations. These equations were integrated into a binary decision tree that sequentially sorts uncharacterized tracks into distinct categories.

Once we had finished reading about the sperm related SVM, we moved on to research [20] which reveals the advantages of the proposed mixed-feature model and presents the capability of identifying human facial expressions from static images. The subsequent framework is a multistage discrimination model based on global appearance features extracted from two-dimensional principal component analysis (2DPCA) and local texture represented by a local binary pattern (LBP). The experimental results indicate that the proposed mixed-feature model is feasible and outperforms the single-feature model.

We then tried to look for research on feature extraction and the SVM classifier [21]. This paper introduces a new method for the early detection of colon cancer using a combination of feature extraction based on wavelets for Fourier

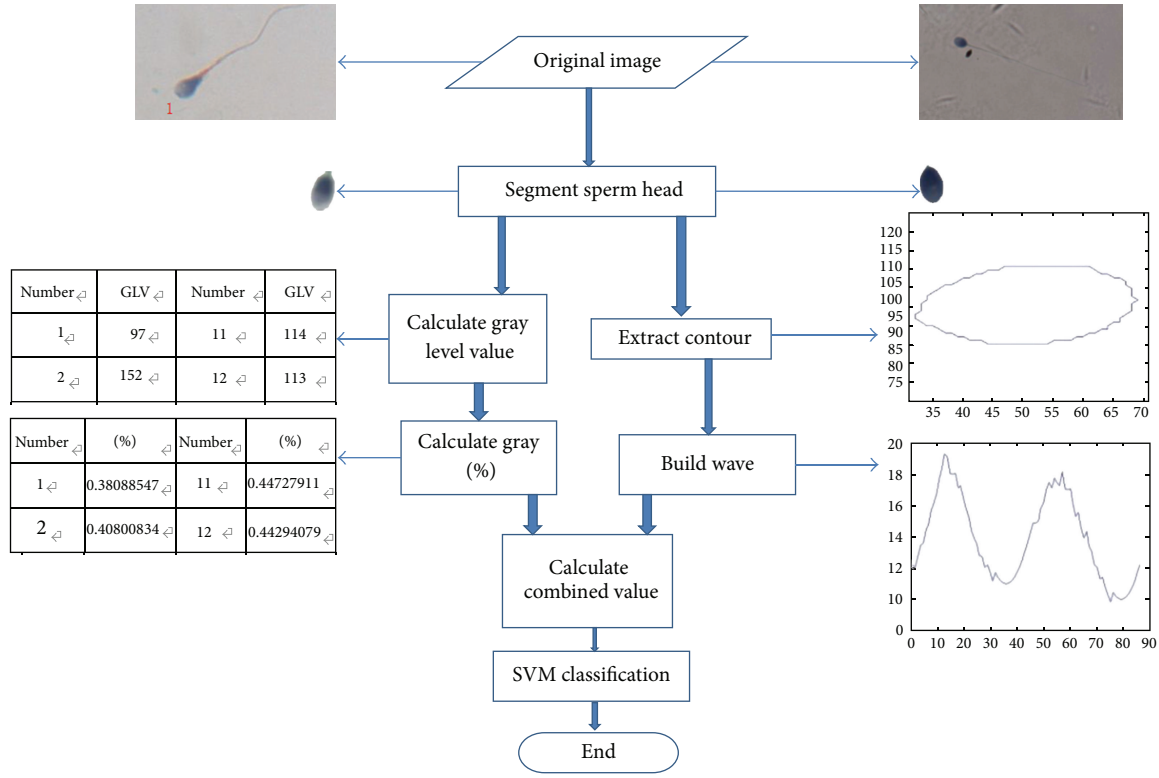


Figure 3: The flow chart of the sperm classification system.

transform infrared spectroscopy (FTIR) and classification with SVM.

One popular robust image-matching approach is SIFT. In paper [22], it performs a reliable matching between different views of an object or scene. Its features are invariant to image scale and rotation. In that sense, its images can be matched with high reliability against a large database of features from many images.

With regard to the elliptic model, it tried to estimate the contour of a sperm by an ellipse shape. Research work [23] used an ellipse to classify sperm. Another article, [24], proposed a new method of sperm morphological classification using the elliptic shape parameterized by the discrete Fourier transform and reconstructed with dyadic data points. The enclosed area of boundaries as a classification feature was calculated and transformed by wavelet transform.

Although some researchers focused on sperm classification, none has used our approach for sperm imaging. This paper presents a system and a novel approach which uses a one-dimensional contour and gray level features to diagnose different sperms according to their characteristics.

3. Proposed Approaches

3.1. Overall Procedure. In this section, we present the algorithm of our work. First of all, we provide the flow of our approach as Figure 3. After image segmentation, the sperm head contour is extracted; we transform the edge of the sperm image into separated coordinate points and divide the coordinate into two vectors.

3.2. Proposed Algorithms. For the classification, we applied two methods. In the first, the bilateral symmetrical function of (n), for the continuous situation, is defined as follows:

$$S(n) = \int_{C_1} \left\|\overline{P_n P}\right\|^2 ds - \int_{C_2} \left\|\overline{P_n P}\right\|^2 ds, \quad (1)$$

where C_1 denotes the arc on the left of the axis while C_2 is the right arc.

The symmetry function is proposed as follows:

$$\text{Percent}_{\text{sym}}\left(A_n(k)\right) = 1 - \frac{\sum \text{abs}\left(A(i) - A(2k-i)\right)}{\sum 2 * \text{sqrt}\left(A(i) * A(2k-i)\right)}. \quad (2)$$

$A_n(k)$ denotes the corner point array while the kth point is the smallest value of $S(n)$.

In our second method, the geometric centre is calculated. We used the following equation to compute the horizontal coordinate:

$$x_{\text{mid}} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad (3)$$

where x_{mid} represents the horizontal coordinate of the geometric centre, x_i represents the horizontal coordinate of each point on the contour, and n represents the number of points

on the edge. The vertical coordinate could be calculated in the same way.

To obtain the distance from edge to the centre, we take advantage of the following format:

$$d(i) = \sqrt{(x_i - x_{\text{mid}})^2 + (y_i - y_{\text{mid}})^2}. \quad (4)$$

Considering the difficulty in choosing the starting point, we had to avoid the problem. Therefore, we took rank algorithm [13] which ignores where to start into consideration in order to classify the results. The rank algorithm transforms a sequence of numbers $\{x_1; x_2; x_3; \ldots; x_n\}$ into a sequence composed of 1 and 0. The transform format is as follows:

$$I_n = \begin{cases} 0; & \text{if } x_n \le x_{n-1}; \\ 1; & \text{if } x_n > x_{n-1}. \end{cases} \quad (5)$$

Therefore, we achieved a binary system with 1 and 0. We transformed 5 consecutive numbers in I_n into decimal numbers. By computing the times each decimal number appeared, we ascertained the rank and the probability of the original sequence. Taking advantage of rank and probability, we calculated the similarity between two sequences. The format is as follows:

$$D_m(S_1, S_2) = \frac{\sum_{k=0}^{2^m-1} |R_1(w_k) - R_2(w_k)| \, p_1(w_k) \, p_2(w_k)}{(2^m - 1) \sum_{k=0}^{2^m-1} p_1(w_k) \, p_2(w_k)}. \quad (6)$$

By calculating the D_m we distinguished the ten most representative sequences and took them as the criteria.

Using the rank algorithm [13], we achieved another binary system with 1 and 0. Then we transformed 8 consecutive numbers in I_n into decimal numbers and finally we obtained the rank and the probability of the sequence. With format (6) we obtained the similarity among the sequences. Similar to the previous method, we achieved another criterion.

To avoid the problem of a single criterion being too lopsided, we took the gray level value of the sperm into consideration. First, we calculated the gray level value of all points within the sperm using the following format:

$$G = \sum_{1}^{n} g_i, \quad (7)$$

where G represents the summation of the gray level value of each point, g_i represents the gray level value of each point, and n represents the total number of points within the sperm.

Then we computed the gray percentage using the following format:

$$P = 1 - \frac{G}{n * 255}, \quad (8)$$

where P represents the level of darkness.

3.2.1. Joint Rank Difference and Gray Level Method. By calculating the average rank difference of normal and abnormal sperms, we found the dividing line between them. Thus, the rank difference itself can work as a judgment as to whether or not the test sperm is normal. We combined the two rank differences originating from the distance from centre to the contour and the grey level value of the pixels in a sperm by calculating the sum of the test sperm's distance average rank difference and the average grey level value rank difference while each of them takes a certain weight. The format is as follows:

$$C = \alpha * \text{dARD} + (1 - \alpha) * 255 * P. \quad (9)$$

C represents the value that is the combination of the two standards; α represents the weight of average rank difference; dARD represents the average rank difference of distance from the centre to the contour.

With the combination value we enhanced the original judgment by considering more elements and the importance of each. To find the dividing line, we collected all the combination values of the sperms and chose one of their average values to provide the best accuracy for the dividing line.

3.3. Fused SVM Method. To achieve a better result, we fused the SVM method as an advanced classifier. SVM Methods are supervised learning models with associated learning algorithms that analyse data and recognize patterns and are used for classification and regression analysis. The basic SVM takes a set of input sperm features and predicts, for each given input, the possible class form, normal or abnormal, making it a nonprobabilistic binary linear classifier.

Given a training dataset $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $x_i \in \mathbb{R}$ and y_i is either 1 or -1 indicating the class to which the point x_i belongs, let $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$. The construction of the hyperplane for a linearly separable problem is $\mathbf{w}^T\mathbf{x} + b = 0$, where $\mathbf{w}$ is the normal vector to the hyperplane and the parameter $b/\|\mathbf{w}\|$ determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$. Thus, the margin between the hyperplane and the nearest point is maximized and can be posed as the following problem:

$$\begin{aligned} &\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ &\text{subject to} \quad y_i\left(w^T x_i + b\right) \ge 1 - \xi_i, \quad i = 1, 2, \ldots, n, \\ &\qquad\qquad\qquad\qquad\qquad\qquad \xi_i \ge 0, \end{aligned} \quad (10)$$

where C is a user-defined constant as the penalty parameter of the error term.

The SVM requires the optimal solution. We use the LIBSVM [25] to solve this optimization problem with the user guide given in [26].

As for the input of the SVM approach, we fused the grey level value and dARD as well. First of all, it extracts grey level value of the pixels within the sperm head. It starts from the centre of the sperm, and then it computes the following grey level values in a square roundabout path. The path map is shown in Table 1 as the sequence of numbers.

Table 1: Grey level value extraction path.

43	42	41	40	39	38	37	64
44	21	20	19	18	17	36	63
45	22	7	6	5	16	35	62
46	23	8	1	4	15	34	61
47	24	9	2	3	14	33	60
48	25	10	11	12	13	32	59
49	26	27	28	29	30	31	58
50	51	52	53	54	55	56	57

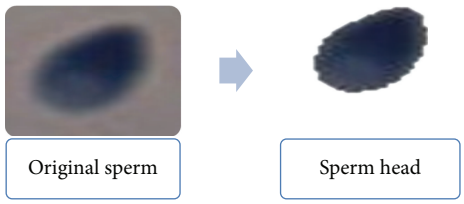


Figure 4: Extract sperm head from normal sperm.

When the path collides with the contour, the sequence of grey level values takes advantage of the Rank algorithm in transforming 8 consecutive numbers I_n into the decimal numbers during the ends of extraction. Then we achieved rank sequences and D_m between every pair of images which can represent their similarities. As a result, we can tell which image represents the rest of the images the best by calculating the sum of the D_m value from one image to all the other images. Thus we can put the rank sequences in order as well.

With the rank sequences, we added the dARD sequences at the end of corresponding rank sequence. The new sequences consist of the training samples and test samples. The ten training samples include five most representative sequences of normal sperms and five ones of abnormal sperms. The rest sperms are the test samples.

4. Fundamental Evaluation

In this fundamental evaluation, we have undertaken various experiments on 80 normal and 80 abnormal sperms. They are from a hospital; the images are provided on https://code.google.com/p/support-vector-machine-for-sperm-morphology-diagnosis/ as Supplementary material. See Supplementary Material available online at http://dx.doi.org/10.1155/2013/687607. Before the classification of sperms, the segmentation is done by two steps. (1) We pick the area where sperms do not overlay manually. (2) We segment the sperms by reading the gray scale value of the picked area. With knowing the location of the sperms we copy them from the area and build a new image to store each one of them. With a single sperm in an image, it can cut off its tail easily.

Figures 4 and 5 present an example of our segmentation result.

In this way, we segmented all the 80 normal and 80 abnormal sperm heads from 80 pieces of sperm images.

4.1. Results of Our Waveform Extraction Methods. In this section, we present part of the results of the steps in our experiments which explain how we made a choice and their performances. In Table 2, the column *Number* lists the number of each image. The columns *Symmetry* and *Mid* show the result of each transformation from image into waveform.

Table 3 presents the waveform obtained by the first method using four different starting points for the same image. With the differences among four waves, the problem arises from choosing one for the classification. The result of the second approach avoids the problem of picking the starting point. Thus, we take the second result to classify sperms.

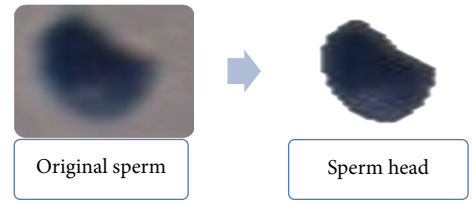


Figure 5: Extract sperm head from abnormal sperm.

4.2. Average Rank Difference. In Figure 6, the horizontal axis represents the serial number of each sperm image. The vertical axis represents the average difference of rank among the 80 normal or 80 abnormal sperms and the ten model sperms. As the result did not provide a clear dividing line between normal and abnormal sperms, we adopted another method to examine their differences. We calculated the difference in rank between the ten normal sperms and the other sperms. Taking a look at Figure 6, if we take horizontal value 181 as the dividing line between normal and abnormal, meaning when the value of the average rank difference of a sperm is below 181, we take the sperm as normal, otherwise we take it as abnormal. In this case, the accuracy could reach 55%.

4.3. Grey Level Feature. Then we focused on the influence of the grey level value. Parts of the results are as shown in Figures 7 and 8.

In Figures 7 and 8, the horizontal axis represents the serial number of each sperm image. In Figures 7(a) and 8(a), the vertical axis represents the average grey level value of the points of the sperms in the images. In Figures 7(b) and 8(b), the vertical axis represents the grey level percentage of the segmentation of the sperms.

According to the percentage results, we can ascertain the percentage dividing line between normal and abnormal sperms to be 0.268228785. With the division we can achieve 72.5% accuracy.

4.4. Joint Average Result. We then tried to combine the two aspects as a judgment. First of all, we found the new parameter through format (7). By setting different α, we determined the different weight of each aspect. The new parameter is as in Figure 9. It contains the new criteria produced with rank and gray level values of normal and abnormal sperms. α is the argument that represents

Table 2: Waveform extraction from both methods.

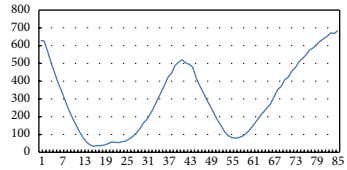
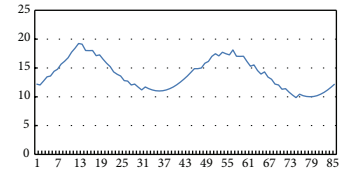
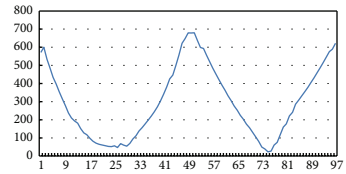
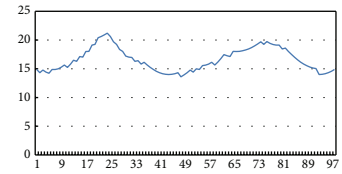
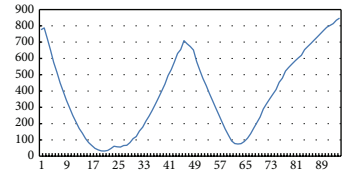
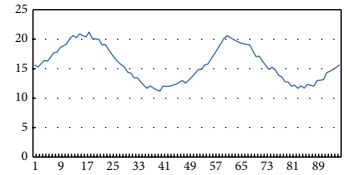
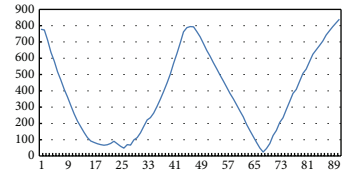
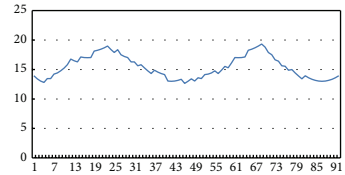
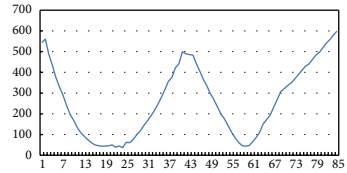
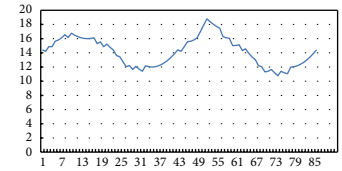
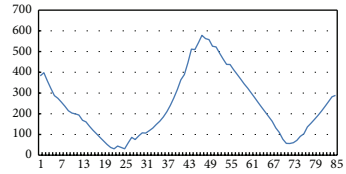
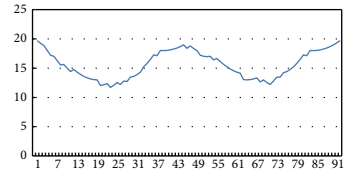
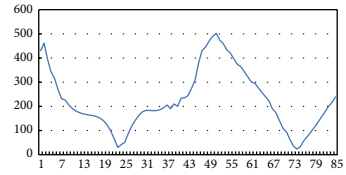
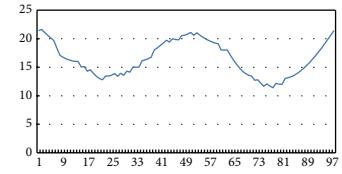
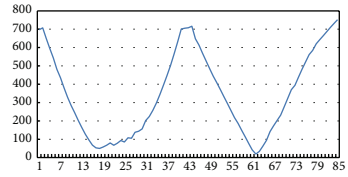
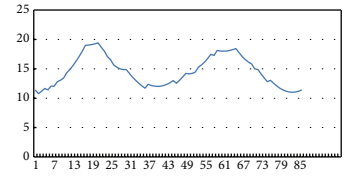
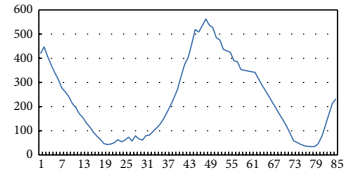
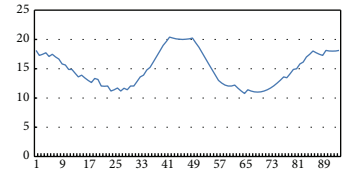
Num	Normal	Symmetry	Mid
1			
2			
3			
4			
5			
6			
7			
8			
9			

Table 2: Continued.

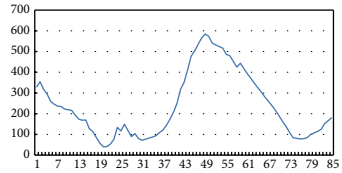
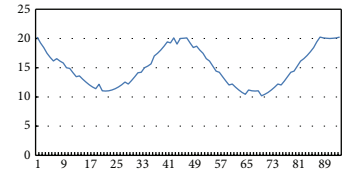
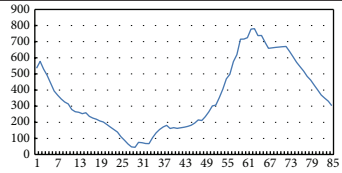
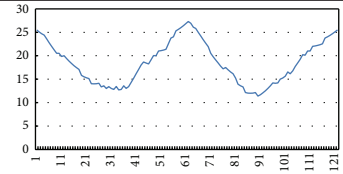
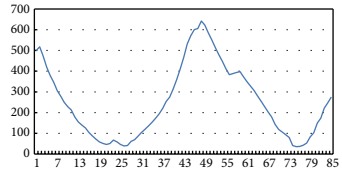
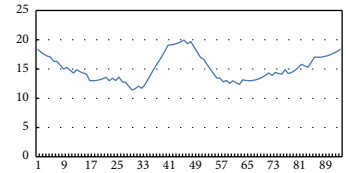
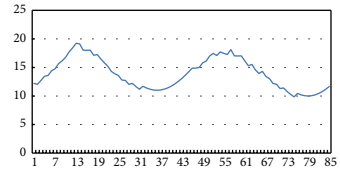
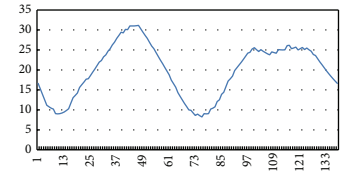
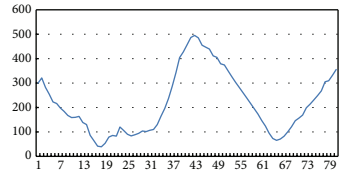
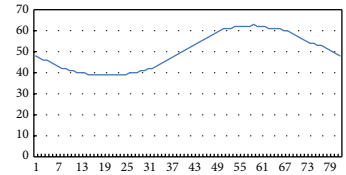
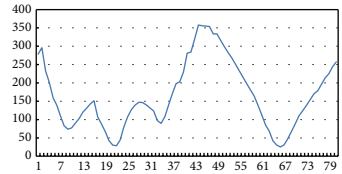
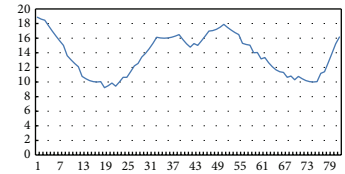
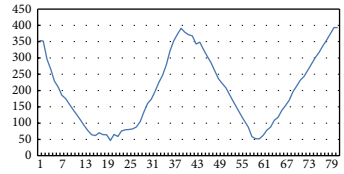
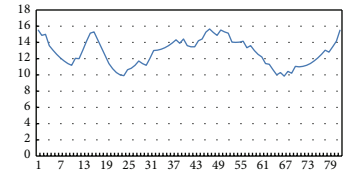
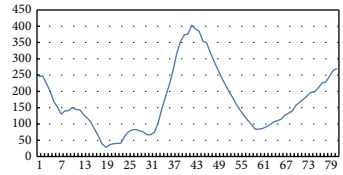
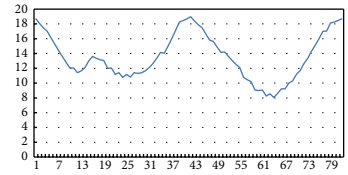
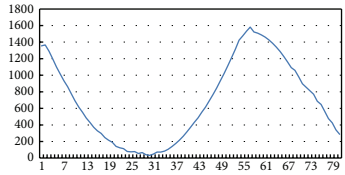
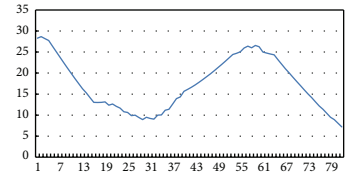
Num	Abnormal	Symmetry	Mid
10			
81			
82			
83			
84			
85			
86			
87			
88			

Table 2: Continued.

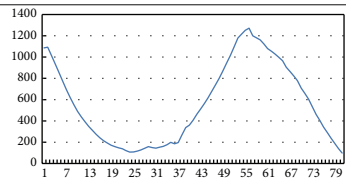
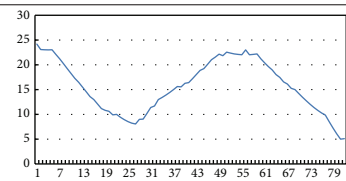
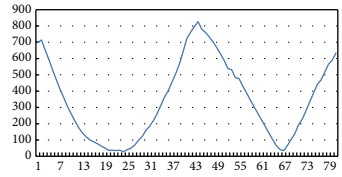
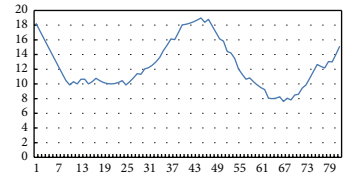
89			
90			

Table 3: The waveform of different starting points from the first method.

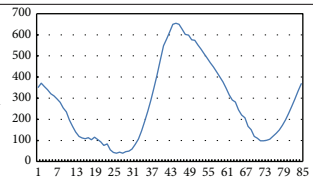
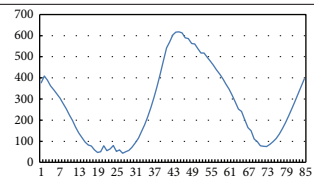
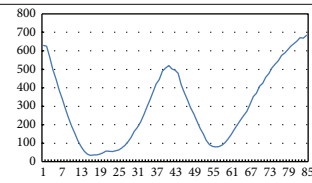
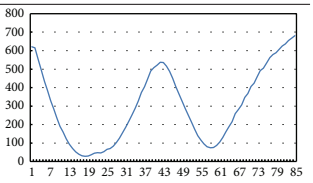
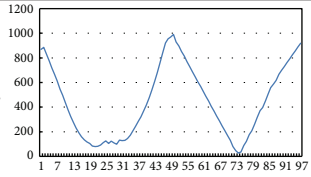
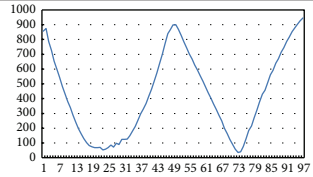
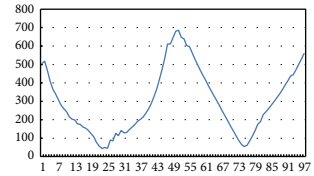
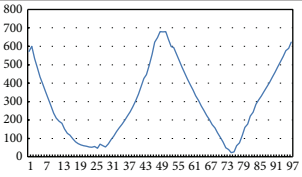
Start point	Result			
Start 1				
Start 2				

Table 4: Dividing line of C.

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
DL	177	165	160	149	134	123	114	101	95

Table 5: Accuracy of different α.

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Acc	0.61	0.62	0.64	0.66	0.69	0.72	0.74	0.77	0.73

the weighted average rank difference taken in the criteria. With the new parameter we can calculate its dividing line (DL) as a threshold.

In Table 4, row DL represents the dividing line of each α value. With the dividing line we can classify the sperms. Assuming we have already known the exact classification of each sperm, we can tell how many of them are correctly classified and thus, we can come up with the accuracy. The accuracy (Acc) of the different α is as shown in Table 5.

In Table 5, the row Acc represents the accuracy of each α value. According to the result, we chose the value 0.8 for α. In this case, the accuracy can reach 76.875%.

4.5. Fused SVM Result. With all the data provided above, we took advantage of the SVM method to undertake the classification and used LIBSVM Software from http://www.csie.ntu.edu.tw/~cjlin/libsvm. As for the parameters, the default values are used in the implementation, and the SVM type is C-SVC; the kernel function type is RBF function: $e^{-\gamma(u-v)^2} = 1/k$; cost = 1; cache size = 40; eps = 0.001; shrinking = 1; weight = 1.

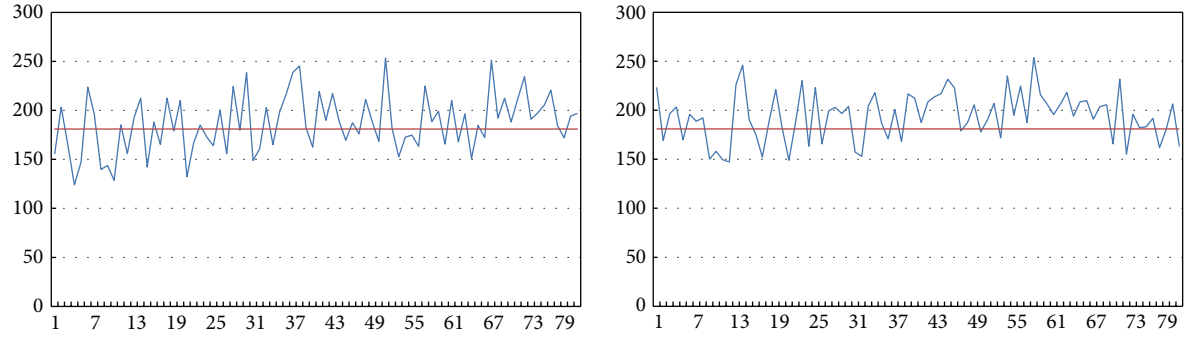
We input the training samples and test samples as mentioned in Section 3.3. With ten training samples and test samples of 160 sperms which consist of 80 normal and 80 abnormal ones, we obtain accuracy with 88.9%.

5. Comparison Evaluation

In this section, we introduce the three methods we used as comparisons to our method: K-nearest neighbor method works quite effective as well, the other two fail to satisfy us.

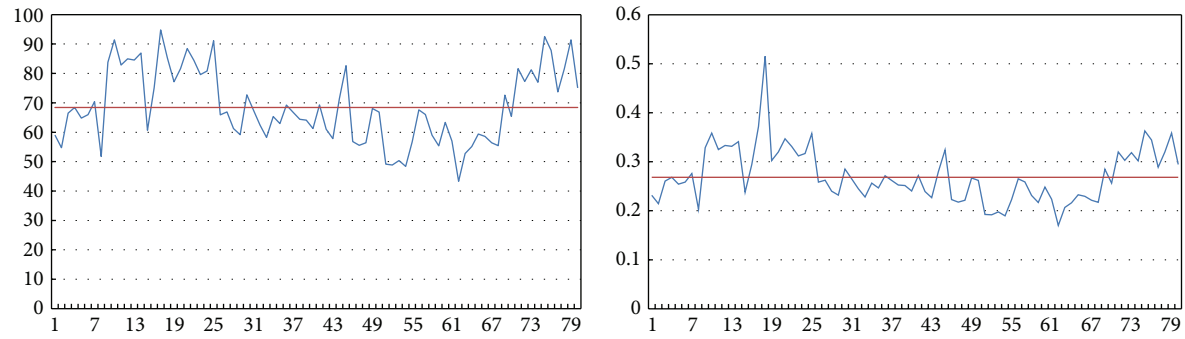
5.1. Compared K-Nearest Neighbour Approach. The K-nearest neighbour method showed the following differences.

We took the 160 images of sperms as the original classifications, and then we input the test image, transformed all the images into gray, and then calculated the difference between the test image and the original images. Then we picked the ten images with the smaller differences. After counting the classifications of the ten images, we specified the one occurring most often to be the classification of the test



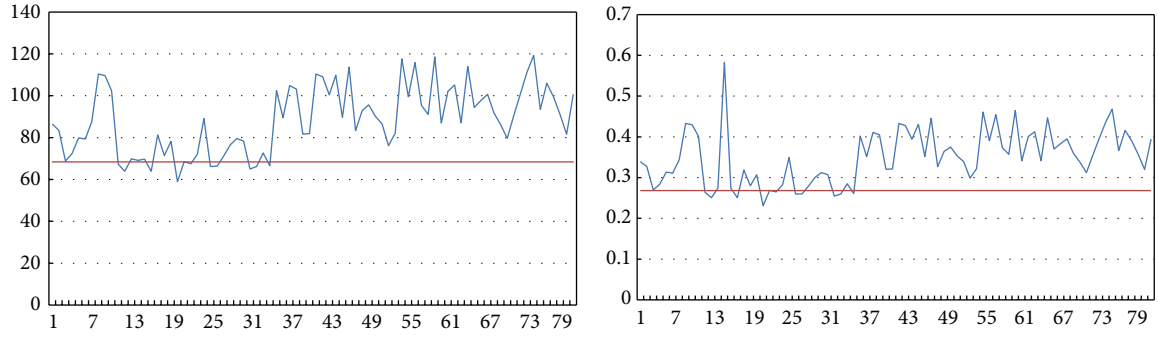
(a) Normal sperms (b) Abnormal sperms

Figure 6: Normal and abnormal sperms and their corresponding data.



(a) (b)

Figure 7: Grey level value and percentage of normal sperms.



(a) (b)

Figure 8: Grey level value and percentage of abnormal sperms.

image. The classification of the top ten nearest neighbours is as in Figure 10.

Taking the sperms whose sum value is larger than 3 as normal and the sperms whose sum value is smaller than 4 as abnormal, we can tell that only 39 sperms are mistakenly classified, providing accuracy of 75.625%.

5.2. Compared SIFT Approach. Lowe summed up the existing feature detection method based on invariants technology, in 2004, and formally proposed an image scaling, rotation, and even affine transformation for invariant image with local feature description operator based on scale space SIFT [22]. The SIFT algorithm first undertakes feature detection in scale space and defines the key points' positions and the scale of the key points, and then it uses the main direction of the neighbourhood gradient of the key points as the direction features of the points in order to achieve the operator independence of scale and the direction. The format which produces the scale space is as follows:

$$L(x, y; t) = \int_{\xi=-\infty}^{\infty} \int_{\eta}^{\infty} \frac{1}{2\pi t} e^{-(\xi^2+\eta^2)/2t} f(x-\xi, y-\eta)\, d\xi\, d\eta. \tag{11}$$

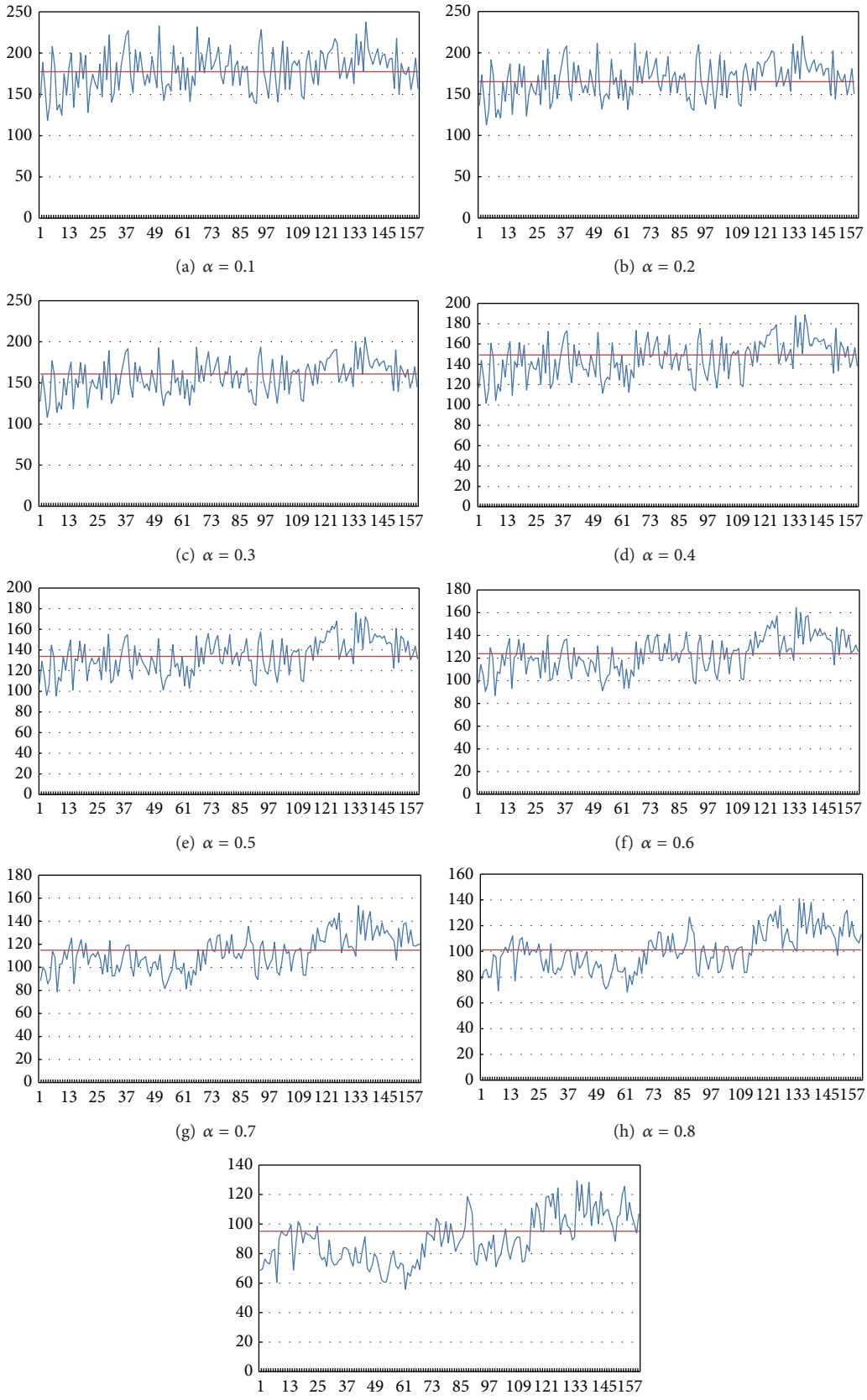
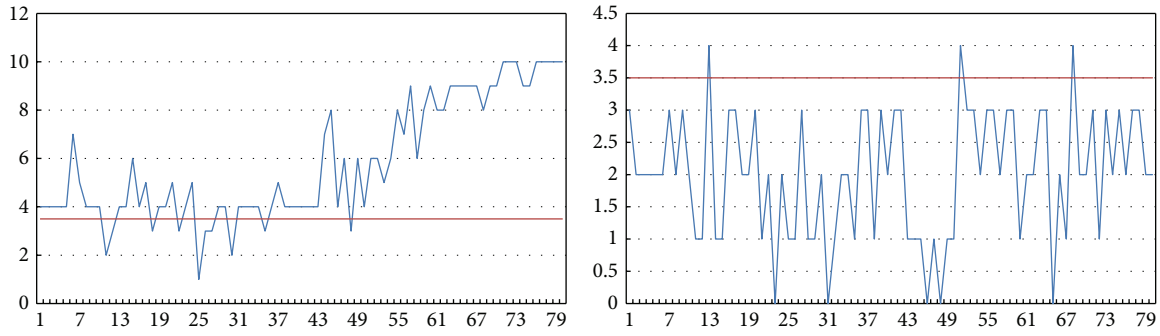


Figure 9: The result of joint rank and gray level features of normal and abnormal sperms.



(a) Normal sperms (b) Abnormal sperms

Figure 10: The classification of the top ten neighbours of (a) normal and (b) abnormal sperms.

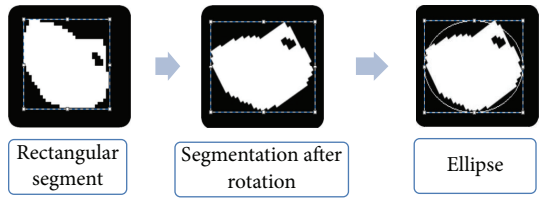


Figure 11: The ellipse model to estimate the contour of the sperm.

In format (11), t represents the scaling parameter. By undertaking convolution in the whole domain with a two-dimensional Gaussian kernel and input image, we can achieve scaling correspondent to t.

The SIFT feature vector has the following features: (a) it is the local feature of an image which maintains invariance not only on rotation, scale, and brightness variation but also on the viewing angle, the affine transformation, and the noise; (b) it is distinctive and informative and suitable for fast, accurate matching in a mass signature database; (c) it can produce a large number of SIFT feature vectors with few objects; (d) it is of high speed.

Matlab source code of SIFT is from http://www.vlfeat.org/index.html [27]. We used the SIFT method to extract feature points of all the sperm images and achieved the ten normal and ten abnormal sperm images as in the training model. Then, we calculated the difference between the training models and the other 70 normal and 70 abnormal sperm images. With the dividing line we came up with accuracy of 50%. The result turned out to be bad because the input data are simple sperm images and the SIFT algorithm cannot extract enough feature descriptors for classification.

5.3. Compared Ellipse Model Approach. With the image of a single sperm, we can use the ellipse model to estimate the contour of the sperm. First of all, we used a rectangular segment to cut out the sperm. In order to make the rectangle close to the sperm edge, we needed to choose the rectangle with the smallest area; we rotated the sperm so that the segmentation would be easier. We calculated the left-most, upper-most, downward-most, and right-most points on the edges and built the rectangular segment based on them as depicted in Figure 11.

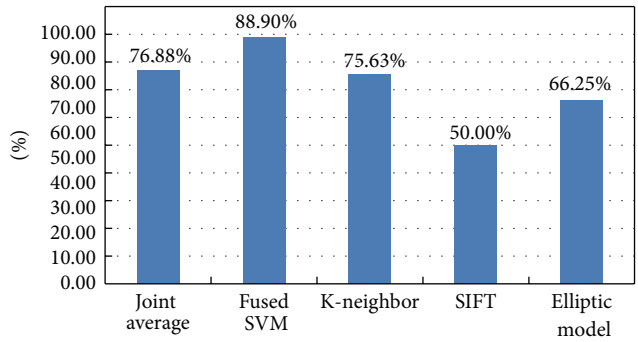


Figure 12: Accuracy comparison of all methods.

With the rectangle we achieved an ellipse to estimate the sperm whose long axis radius is half the length of the rectangle's long side and whose short axis radius is half the length of the rectangle's short side. As a result, we ascertained the ovality, which is the ratio of the length of the long axis to the short axis. Through our experiment on 80 sperms, we could tell whether the sperm was too close to a circle, whose ratio is close to 1, and whether the sperm was too slim, whose ratio is close to 2 or even larger than 2. We took those whose ratio was larger than 1.2 and smaller than 1.8 as normal, and thus we achieved accuracy of 66.25%.

6. Conclusion

As shown in Figure 12, we have presented an approach for the classification of sperm images with one-dimensional waveform and gray level features. In the comprehensive evaluation, the joint average approach method was applied

to 160 sperm image samples and provided 76.875% accuracy of judgment. We then applied the fused SVM method to the images and reached accuracy of 87.5%. The result proves that the proposed approach is superior to the previous approaches.

References

[1] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recognition*, vol. 26, no. 9, pp. 1277–1294, 1993.

[2] A. Yezzi, L. Zöllei, and T. Kapur, "A variational framework for integrating segmentation and registration through active contours," *Medical Image Analysis*, vol. 7, no. 2, pp. 171–185, 2003.

[3] R. Chrástek, M. Wolf, K. Donath et al., "Automated segmentation of the optic nerve head for diagnosis of glaucoma," *Medical Image Analysis*, vol. 9, no. 4, pp. 297–314, 2005.

[4] H. W. Kang and S. Y. Shin, "Enhanced lane: interactive image segmentation by incremental path map construction," *Graphical Models*, vol. 64, no. 5, pp. 282–303, 2002.

[5] J. Peters, O. Ecabert, C. Meyer, R. Kneser, and J. Weese, "Optimizing boundary detection via Simulated Search with applications to multi-modal heart segmentation," *Medical Image Analysis*, vol. 14, no. 1, pp. 70–84, 2010.

[6] G. Behiels, F. Maes, D. Vandermeulen, and P. Suetens, "Evaluation of image features and search strategies for segmentation of bone structures in radiographs using Active Shape Models," *Medical Image Analysis*, vol. 6, no. 1, pp. 47–62, 2002.

[7] M. Blanchard, K. Haguenoer, A. Apert et al., "Sperm morphology assessment using David's classiffication: time to switch to strict criteria? Prospective comparative analysis in a selected IVF population," *International Journal of Andrology*, vol. 34, no. 2, pp. 145–152, 2010.

[8] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[9] T. McInerney and D. Terzopoulos, "T-snakes: topology adaptive snakes," *Medical Image Analysis*, vol. 4, no. 2, pp. 73–91, 2000.

[10] J. Liang, T. McInerney, and D. Terzopoulos, "United Snakes," *Medical Image Analysis*, vol. 10, no. 2, pp. 215–233, 2006.

[11] M. Zhao, Z. Meng, K. K. Tseng, J. S. Pan, and C. Y. Hsu, "Symmetry auto-detection based on contour and corner models," in *Proceedings of the 5th International Conference on Genetic and Evolutionary Computing (ICGEC '11)*, pp. 345–349, September 2011.

[12] Y. Mingqiang, K. Kidiyo, and R. Joseph, "A survey of shape feature extraction techniques," *Pattern Recognition*, vol. 15, no. 7, pp. 43–90, 2008.

[13] Y. Wang and H. Huang, "Analysis of Human Heartbeat with EKG Signals," 2010.

[14] Q. Hua, A. Ji, and Q. He, "Multiple real-valued K nearest neighbor classifiers system by feature grouping," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '10)*, pp. 3922–3925, October 2010.

[15] P. Viswanath and T. Hitendra Sarma, "An improvement to k-nearest neighbor classifier," in *Proceedings of the IEEE Recent Advances in Intelligent Computational Systems (RAICS '11)*, pp. 227–231, September 2011.

[16] P. Wu and Q. Chen, "A novel SVM-based edge detection method," *Physics Procedia*, vol. 24, pp. 2075–2082, 2012.

[17] D. N. Sotiropoulos and G. A. Tsihrintzis, "Artificial immune system-based classification in class-imbalanced image problems," in *Proceedings of the 8th International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 138–141, July 2012.

[18] M. Ramón, F. Martínez-Pastor, O. García-Álvarez et al., "Taking advantage of the use of supervised learning methods for characterization of sperm population structure related with freezability in the Iberian red deer," *Theriogenology*, vol. 77, no. 8, pp. 1661–1672, 2012.

[19] S. G. Goodson, Z. Zhang, J. K. Tsuruta, W. Wang, and D. A. O'Brien, "Classification of mouse sperm motility patterns using an automated multiclass support vector machines model," *Biology of Reproduction*, vol. 84, no. 6, pp. 1207–1215, 2011.

[20] D. T. Lin and D. C. Pan, "Integrating a mixed-feature model and multiclass support vector machine for facial expression recognition," *Integrated Computer-Aided Engineering*, vol. 16, no. 1, pp. 61–74, 2009.

[21] C. G. Cheng, Y. M. Tian, and W. Y. Jin, "A study on the early detection of colon cancer using the methods of wavelet feature extraction and SVM classifications of FTIR," *Spectroscopy*, vol. 22, no. 5, pp. 397–404, 2008.

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[23] V. R. Nafisi, M. H. Moradi, and M. H. Nasr-Esfahani, "Sperm identification using elliptic model and tail detection," *World Academy of Science, Engineering and Technology*, vol. 6, pp. 205–208, 2005.

[24] W. J. Yi, K. S. Park, and J. S. Paick, "Parameterized characterization of elliptic sperm heads using Fourier representation and wavelet transform," in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 20, pp. 974–977, 1998.

[25] C. C. Chang and C. J. Lin, LIBSVM—a library for support vector machines, Version 3.12, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[26] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," 2010, http://www.csie.ntu.edu.tw/~cjlin.

[27] A. Vedaldi and B. Fulkerson, "VLFeat: an open and portable library of computer vision algorithms," 2008, http://www.vlfeat.org/.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 742835, 8 pages
http://dx.doi.org/10.1155/2013/742835

Review Article

Advanced Systems Biology Methods in Drug Discovery and Translational Biomedicine

Jun Zou, Ming-Wu Zheng, Gen Li, and Zhi-Guang Su

Molecular Medicine Research Center, State Key Laboratory of Biotherapy, West China Hospital, West China School of Medicine, Sichuan University, Chengdu 610041, China

Correspondence should be addressed to Zhi-Guang Su; zhiguang_su@hotmail.com

Received 17 June 2013; Accepted 26 August 2013

Academic Editor: Bing Niu

Systems biology is in an exponential development stage in recent years and has been widely utilized in biomedicine to better understand the molecular basis of human disease and the mechanism of drug action. Here, we discuss the fundamental concept of systems biology and its two computational methods that have been commonly used, that is, network analysis and dynamical modeling. The applications of systems biology in elucidating human disease are highlighted, consisting of human disease networks, treatment response prediction, investigation of disease mechanisms, and disease-associated gene prediction. In addition, important advances in drug discovery, to which systems biology makes significant contributions, are discussed, including drug-target networks, prediction of drug-target interactions, investigation of drug adverse effects, drug repositioning, and drug combination prediction. The systems biology methods and applications covered in this review provide a framework for addressing disease mechanism and approaching drug discovery, which will facilitate the translation of research findings into clinical benefits such as novel biomarkers and promising therapies.

1. Introduction

Advances in biological sciences over the past several decades have led to the generation of a large amount of omics molecular data at the level of genome, transcriptome, proteome, and metabolome. While identifying all the genes and proteins provides a catalog of individual molecular components, it is not sufficient by itself to understand the complexity inherent in biological systems. We need to know how individual components are assembled to form the structure of the biological systems, how these interacting components can produce complex system behaviors, and how changes in conditions may dynamically alter these behaviors. As a result, systems biology has emerged as an important new discipline that addresses the current challenge of interpreting the overwhelming amount of genome-scale data on a systems level.

Yet remaining in its infancy in many ways, systems biology is in an exponential development stage in recent years and has been widely used in pharmacology to better understand molecular basis of disease and mechanism of drug action [1]. It has become apparent that many diseases such as cancer are much more complex than initially anticipated, because they are often caused by a combination of multiple molecular abnormalities, which supports a novel network perspective of complex diseases [2]. In addition, many drug candidates failed clinical phases because the mechanisms of the cellular pathways they target are incompletely understood. These have significant implications in the drug discovery process because the molecular components that need to be targeted must change from single proteins to entire cellular pathways [3]. By considering the biological context of drug target, systems biology provides new opportunities to address disease mechanisms and approach drug discovery, which will facilitate the translation of preclinical discoveries into clinic benefits such as novel biomarkers and therapies [4].

In the following sections, we will first describe systems biology methods that have become commonplace; then we will examine their various applications in drug discovery and translation medicine; finally brief discussions on future directions are given.

2. Systems Biology Methods

Systems biology focuses on developing an understanding of how the phenotypic behavior of biological system as a whole emerges from individual molecular components and their interactions that constitute the biological system [5]. Thus a key feature of systems biology is that interactions among many components are studied, rather than simply the characteristics of individual molecules. Another feature is that systems biology uses a range of computational approaches to generate predictions that can be tested experimentally. Systems biology thus relies on a combination of experiments that measure multiple cellular components and computational approaches that allow the analysis of various data sets. As an iterative process, computational modeling is performed to propose nonintuitive hypotheses that can subsequently be experimentally validated, and the newly acquired quantitative experimental data can then be used to refine the computational model that recapitulated the biological system of interest.

In general, two complementary computational approaches are used in systems biology, namely, data-driven and hypothesis-driven methodologies (also called top-down and bottom-up modeling) [6]. The data-driven approaches involve the gathering of large-scale omics data sets and subsequent analyses of these data using statistical modeling techniques. Network modeling, one of the most frequently used data-driven approaches, provides insights into the interactions among hundreds or even thousands of molecular components. On the other hand, the hypothesis-driven approaches are generally applied to relatively small systems with fewer molecular components. A major challenge to this approach is that the quantitative details of the interactions are unknown and so it is necessary to hypothesize relevant forms of the equations that govern the interactions and estimate the values of the associated parameters [6]. Dynamical modeling, the major hypothesis-driven approach, can be employed to characterize the quantitative relations between molecular components and the emergent behaviors that arise from their interactions. Choosing the appropriate modeling approaches depends on the nature of the data and the level of understanding of the studied biological system.

3. Network Modeling in Biological Systems

A "network" refers to a collection of "nodes" and a collection of "edges" that connect pairs of nodes. Network representation of biological molecular systems typically considers molecular components as nodes and their interactions or relationships as edges. In biological networks, molecular components can be genes, proteins, metabolites, drugs, or even diseases and phenotypes; interactions can be direct physical interactions, metabolic coupling, and transcriptional activation. Different types of biological networks can be constructed, such as protein-protein interaction networks, cellular signaling networks, gene regulatory networks, disease gene interaction networks, and drug interaction networks [7, 8].

Network analysis of biological systems is increasingly gaining acceptance as a useful method for data integration and analysis. Assembling a network to represent the complexity of biological systems is just recognized as the beginning of the analysis. A series of advances in graph-based theory are also relied on to provide insights into the topology properties and organizational principles of biological networks, which include information about the properties of nodes and edges, global (i.e., the entire network) topological properties, hubs, motifs, and modules [2, 7]. Properties of nodes include degree (also called connectivity degree), node betweenness centrality, closeness centrality, and eigenvector centrality. Properties of edges include edge betweenness centrality, relationship types (i.e., activation or inhibition), and edge directionality. Global topological characteristics of networks include connectivity distribution, characteristic path length, clustering coefficient, grid coefficient, network diameter, and assortativity [7].

The degree of a node is the number of edges that connect to it; for example, the degree of a protein could represent the number of proteins with which it interacts. An important realization is that networks in biological systems, including protein-protein interaction and metabolic networks, are scale-free, which means that the degree distribution (i.e., the fraction of nodes with a given degree) has a power-law tail. By contrast, for a random network, most nodes have approximately the same number of edges (i.e., fits a Poisson distribution). The scale-free architecture makes biological networks robust to random failures [7].

Network motifs are recurring small subnetworks composed of a few nodes and their edges, and the topology types of these subnetworks appear in biological networks much more frequently than expected by chance [8]. Some motifs are particularly important because they are likely to be associated with some optimized biological function; examples include negative feedback loops, positive feed-forward loops, bifans, or oscillators. Another characteristic of networks is their modularity (i.e., network clustering), implying the existence of 'modules', which are network neighbourhoods with locally dense connectivity segregated by regions of low connectivity [8]. In biological networks, a module could correspond to a group of molecules that tend to interact with each other to achieve some closely related cellular functions.

Highly connected nodes in a network are called hubs. The biological role of hubs allows for their classification into "party" hubs and "date" hubs [2]. Party hubs, also called intramodule hubs, are highly coexpressed with their interacting molecules and preferentially function inside modules. While date hubs, also called intermodule hubs, appear to be more dynamically regulated relative to their interacting molecules and preferentially link different functional modules to each other. For example, Chang et al. have recently identified modules enriched in closely connected "party hubs" that all participate in the same biological process "ribosome biogenesis and assembly" [9]. Whereas *CDC28*, predicted as a "date hub," serves as an intermodule coordinator and performs important functions in the regulation of both "cell cycle" and "DNA damage" [9].

4. Dynamical Modeling in Biological Systems

Dynamical modeling, also named mechanistic modeling, can be viewed as translations of familiar pathway maps into mathematical form [10]. Equations in dynamical models, derived from established physicochemical theory (e.g., the law of mass action and Michaelis-Menten kinetics), seek to describe biomolecular processes (such as intermolecular association, catalysis and covalent modification, and intracellular localization). Kinetic parameters in dynamical models have physicochemical interpretations that define the reaction rate and binding affinity.

Provided that reasonable values for kinetic parameters and initial concentrations of cellular components can be obtained, simulation of the dynamical models yields the concentrations of each component at subsequent times, thereby facilitating comparison of simulated and experimental time courses [5]. Thus dynamical modeling uses prior knowledge to make specific predictions and works best with pathways in which components and connectivity are relatively well established. Used appropriately, dynamical modeling is much more powerful in analyzing molecular events in a cellular context, revealing the principles of biological systems, and generating novel and useful hypotheses.

The correct mathematical form for a dynamical model depends on the properties of the system being studied and the goals of the modeling effort. Ordinary differential equations (ODEs) and partial differential equations (PDEs) are the most commonly used forms. ODEs represent the rates of production and consumption of individual biomolecular component in terms of mass action kinetics, which is an empirical law stating that rates of a reaction are proportional to the concentrations of the reacting components [11]. Each biochemical transformation is therefore represented by an elementary reaction with forward and reverse rate constants. One fundamental assumption of ODEs is that the cellular compartment is well mixed; that is, the concentration of each component is high and transports instantaneously within a compartment [12]. If this assumption is not satisfied, then it is necessary to use PDEs to explicitly simulate the changes in component concentrations with respect to space. Defining a PDEs model requires assigning components and reactions to the appropriate cellular compartment where they occur, diffusion rules and constants governing the transfer of components among different compartments, and the boundary constraints of each compartment [12].

Dynamical systems can be in either deterministic or stochastic form [5]. A dynamical system is deterministic if its trajectory is uniquely determined by the initial state and a given parameter set, while a stochastic dynamical model can go to different states with different probabilities even at a given initial state. Stochastic simulations include effects arising from random fluctuation around the average behavior, such as small molecules number of given component, sufficiently low elementary reactions, or cell-to-cell variability due to intrinsic noise.

To develop a dynamical model, there are approximately four steps. (1) Model design: one of the initial stages is to specify the model scope and establish the reaction scheme of all of the molecular components of interest. This may involve a connectivity diagram listing all of the components (including their biochemically modified versions), their connections (such as stimulatory or inhibitory connections and physical interactions), and their appropriate subcellular location [12]. (2) Model construction: according to the physicochemical theory, the connectivity diagram must be converted into appropriate biochemical reactions, which are mathematically represented by differential equations [13]. Once these reactions have been established, the experimental data needed for the kinetic parameters and initial concentrations are implemented [12]. (3) Model calibration also known as model regression, is the process by which unknown kinetic parameter values in a model are estimated so as to match model performance to experimental measurements. The parameter estimation is generally based on data-fitting techniques that involve an iterative process of adjusting kinetic parameter values to minimize the difference between the model predicted value and the corresponding experimental data [14]. (4) *Model validation* is the process of evaluating the goodness of a calibrated model. This includes making predictions that can be subjected to experimental test. If the simulation results of the dynamical model recapitulate experimentally defined input-output relations, then the model can be considered to be accurate. The input-output relations may be time-course and dose-response experimental data in the presence or absence of additional perturbations [13].

For many biological systems, there are insufficient kinetic parameters for the biochemical reactions, which have posed an obstacle for the application of quantitative dynamical models based on ODEs or PDEs. To address this issue, discrete dynamical modeling has been used to provide an alternative way to qualitatively describe complex biological systems with many unknown kinetic parameters. In these models, the states of the cellular components are qualitative, and the time variable is often considered to be discrete. The main types of discrete dynamical models include Boolean networks and Petri nets [15]. Boolean networks, whose node is described by only two qualitative states (i.e., ON and OFF), have been successfully applied in modeling gene regulatory networks and signaling networks [16, 17]. Petri nets, which contain two types of nodes representing the cellular components and the biochemical reactions, are particularly suited for modeling metabolic networks and analyzing metabolic disorders [18].

5. Systems Biology Methods to Human Disease

Compared with traditional reductionist approach that attempts to explain complex diseases by studying individual gene, systems biology is characterized by the view that the underlying mechanism of complex diseases is likely to be the dysregulation of multiple interconnected cellular pathways. Therefore, biological network analyses and dynamical modeling have been increasingly used to underlie the genotype-phenotype relationships in human disease [8]. Here, we attempt to cover four recent advances in this area:

(1) studies of global relationships between human disease and associated genes, (2) predictions of treatment response, (3) investigations of the underlying mechanism of diseases, and (4) predictions of new disease-associated genes.

5.1. Human Disease Networks. Most previous studies have focused on the association between a single gene and a single disease, whereas systems biology approaches using network-based tools enable a better understanding of the relationships among multiple genes and diseases. Goh et al. used the collected gene-disease associations to build the first human disease network by linking diseases that share one or more disease genes, and it shows that similar pathophenotypes have a higher likelihood of sharing genes than do pathophenotypes that belong to different disease classes [19]. They also found that most disease genes are nonessential and are not encoded by hub proteins. Linghu et al. explored the relationships between diverse diseases and disclosed hidden associations between disease pairs having dissimilar phenotypes [20]. Suthram et al. present an integrated network approach to identify significant similarities between diseases and reveal common disease-state modules significantly enriched for drug targets [21]. Such systematic approaches have also provided a foundation for a genome-scale network analysis of complex diseases, such as cancer [22], neurodegenerative disease [23], inflammatory disease [24], and also pathogen responses [25].

5.2. Treatment Response Prediction. An important area in which systems biology approaches have been applied is biomarker discovery. Several groups have begun to integrate gene and protein expression profiles with system-wide maps of the pathways to identify biomarkers able to diagnose disease severity and predict disease outcomes. A recent study illustrated how the network-based approach that identifies subnetworks with coherent expression patterns can be used to identify novel markers for breast cancer metastasis [26]. The subnetwork-based analysis of gene expression profiles has also successfully been used to predict the relative risk for disease progression and patient survivability [27–30]. In all cases, the goal is to identify biomarkers not as lists of individual genes or proteins but as functionally related groups of genes or proteins whose aggregate properties account for the phenotypic differences between the different populations of patients [31]. Unlike conventional expression diagnostic markers based on individual genes, these network-based diagnostic markers should be inherently more reliable since they provide the biological interpretation for the association between the subnetwork biomarker and the particular type of disease [32].

5.3. Investigation of Disease Mechanisms. Based on the construction of gene regulatory networks from large-scale molecular profiles, systems biology approaches have been valuable for elucidating the mechanisms of both physiological regulation [33] and pathological processes in complex diseases [32]. Recent observations have shown that the wiring of biological networks can change from healthy to diseased states. For instance, inflammatory and immune signaling pathways show different functional wiring when comparing normal and transformed hepatocytes [34]. Rozenblatt-Rosen et al. systematically examined host interactome and transcriptome network perturbations caused by DNA tumor virus proteins [35]. The resulting integrated viral perturbation data reflects rewiring of the host biological networks and highlights pathways, such as Notch signaling and apoptosis, that go wrong in cancer. Zhang et al. constructed gene regulatory networks to characterize molecular systems associated with Alzheimer's disease [36]. Their network-based integrative analysis not only highlighted the strong association of immune pathways with the pathophysiology of the disease but also identified the key network regulators that may serve as effective targets for therapeutic intervention. Another thrust in systems biology involves combining dynamical modeling of regulatory pathways with molecular and cellular experiments as a means to understand the precise regulatory mechanisms of networks that are altered in diseases [37–39].

5.4. Disease-Associated Gene Prediction. The search for disease-causing genes is a long-standing goal of biomedical researches. Systems biology is playing an increasing role in this area through the computational integration of multiple genome-scale measurements. It is assumed that if biological networks underlie genotype-phenotype relationships, then network properties should be able to predict unidentified human disease-associated genes. In an early example, network modeling strategies have been successfully used in tumor research. Starting with known genes encoding tumor suppressors of breast cancer, Pujana et al. generate a network containing genes linked by potential functional associations, and the analysis of this network permitted identification of novel genes potentially associated with higher risk of breast cancer [40]. Mani et al. introduce a systems biology approach, based on the analysis of the network of molecular interactions that become dysregulated in specific tumors, to decipher the human B-lymphocyte interactome, which helped to identify causal oncogenic lesions in several B-cell lymphomas [41]. Similar network-based computational frameworks have been proposed to reliably predict disease-associated genes [42, 43]. It is thus suggested that studying dysregulation at a biological network level, rather than in a "gene centric" manner, can provide a highly efficient method for addressing the problems of identifications of genes playing a role in human disease.

6. Systems Biology Methods in Drug Discovery

Systems biology approaches have long been used in pharmacology to understand drug action. The application of computational and experimental systems biology methods to pharmacology allows us to introduce the definition of "systems pharmacology" [44], which describes a field of research that provides us with a comprehensive view of drug action rooted in molecular interactions between drugs and their targets in a human cellular context. Advances in systems pharmacology will, in the long term, assist in the development of new drugs and more effective therapies for

patient treatment management. There are several important clinically motivated applications in drug discovery to which systems biology approaches make significant contributions. Here, we attempt to discuss five recent advances in this field: (1) drug-target networks, (2) predictions of drug-target interactions, (3) investigations of the adverse effects of drugs, (4) drug repositioning, and (5) predictions of drug combination.

6.1. Drug-Target Networks. Analysis of drug-target networks in a systematic manner shows a rich pattern of interactions among drugs and their targets in which drugs often bind to multiple rather than single molecular targets—a phenomenon known as "polypharmacology" [45, 46]. Topological analyses of this network quantitatively showed an overabundance of "follow-on" drugs, that is, drugs that target already-targeted proteins. Likewise, many proteins are targeted by more than one drug containing distinct chemical structures. This new appreciation of the role of polypharmacology has significant implications for drug development. Although the single-target approach remains the main strategy presently, some remarkable efforts are being put into the development of "promiscuous" drugs (also called "dirty drugs") that can bind to multiple targets.

Integrating systems biology and polypharmacology holds the promise of expanding the current opportunities to improve clinical efficacy and decrease side effects and toxicity. Advances in these areas are creating the foundation of the next paradigm in drug discovery, that is, "network pharmacology" [47]. Keiser et al. related receptors to each other quantitatively based on the chemical similarity among their ligands [48]. They have shown that targets that have no obvious sequence or structure similarity are linked quantitatively based on their bioactive ligands. The unexpected relationships between drug targets could be used to predict their biological function. Li et al. developed a computational framework to build disease-specific drug-protein network, which can help study molecular signature differences between different classes of drugs in specific disease contexts [49].

6.2. Predictions of Drug-Target Interactions. In recent years, the observation of polypharmacology that drugs often bind to more than one molecular target has gained attention. To fully understand the actions of a drug, knowledge of its polypharmacology is clearly essential. Keiser et al. report a computational tool that generates predictions of the pharmacological profile of drugs [50]. Unlike conventional approaches based on sequence or structural similarity between targets, the "similarity ensemble approach" defines each target by its set of known ligands, searches for drugs with chemical structure similar to the known ligands, and then predicts new drug-target associations. Campillos et al. used phenotypic side-effect similarities to infer whether two drugs share a therapeutic target. Applied to marketed drugs, a network of side-effect-driven drug-drug relations became apparent. Several unexpected drug-drug relations are formed by chemically dissimilar drugs from different therapeutic indications, which implies new drug-target relations [51]. Integrating side-effect and pharmacogenomic similarities, Takarabe et al. made a comprehensive prediction and suggested many potential drug-target interactions that were not predicted by previous approaches [52]. Cheng et al. compared three supervised inference methods and found that network-based inference performed best on prediction of drug-target interactions [53].

6.3. Investigations of Drug Adverse Effects. Accurate prediction of the safety and toxicology of drugs in the early stage of drug development pipelines is one of the major challenges in the pharmaceutical industry. Integrating biological data and systems biology approaches could introduce a fundamental change in the way drug candidates are assessed. Lounkine et al. use a similarity ensemble approach, which calculates whether a drug will bind to a target based on the chemical features it shares with those of known ligands, to predict the activity of marketed drugs on unintended "side-effect" targets [54]. Approximately half of their predictions were confirmed by experimental assays. An association metric was developed to prioritize those new off-targets that explained side effects better than any known target of a given drug, creating a drug-target-adverse drug reaction network. Kuhn et al. recently report a large-scale analysis to systematically predict and characterize proteins that cause drug side effects [55]. They integrated clinical phenotypic data with known drug-target interactions to identify overrepresented protein-side effect relations. The results show that a large fraction of complex drug side effects are mediated by individual proteins. Yang et al. have constructed an *in silico* chemical-protein interactome, which mimics the interactions between drugs known to cause at least one type of serious adverse effects and a panel of human proteins [56, 57]. It is revealed that drugs sharing the same adverse effects possess similarities in their chemical-protein interactome profiles. By investigating the associations between drug and off-targets, their research has explored the molecular basis of several adverse events. Other studies that integrate systems biology with structural or chemoinformatics analysis have also been conducted to successfully predict drug adverse effects [58, 59].

6.4. Drug Repositioning. Drug repositioning, also called drug repurposing, is a potential alternative for drug discovery by identifying new therapeutic applications for existing drugs. The main advantage of drug repositioning is that it should drastically reduce the risks of drug development and facilitate repositioned drugs to enter clinical phases more rapidly. As one example of this utility, Iorio et al. developed an approach that exploits similarity in molecular activity signatures of all drugs to compute pair-wise similarities in drug effect and mode of action [60]. Drugs were organized into a network using the resulting similarity scores. Network theory was then applied to partition drugs into groups of densely interconnected nodes (i.e., communities). The resulting drug communities are significantly enriched for compounds with similar mode of action, which often shared the same targets and pathways. Through this approach, drug repositioning is

revealed by colocation of drugs within the network communities, which predicts a shared molecular activity with other drugs in the drug communities. Gottlieb et al. proposed "PREDICT" algorithm that can handle both approved drugs and novel compounds [61]. This new method is based on the observation that similar drugs are indicated for similar diseases and utilizes the chemical similarity of drugs and disease-disease similarity measures for the prediction of novel drug indications. Furthermore, numerous systems biology approaches based on gene expression data for *in silico* drug repositioning have been published [62, 63]. Iskar et al. identified a large set of drug-induced transcriptional modules from genome-wide microarray data of drug-treated human cell lines [64]. The identified modules reveal the conservation of transcriptional responses towards drugs, thereby providing a starting point for drug repositioning.

6.5. Predictions of Drug Combination. Combination therapies, which modulate multiple targets simultaneously, are essential to achieve greater therapeutic benefit than using a single drug [65]. Systems biology methods have been applied to explain and predict potential drug combinations [66]. Computational approaches utilizing dynamical modeling have already been used to simulate the effect of drug combinations and generate experimentally testable interventions [67, 68]. But due to the incomplete knowledge about the kinetic values of biochemical reactions, these dynamical models are currently restricted to a small scale and only suitable for investigating the action mechanisms of drug combination. Considering target information which is usually accessible, the combination effect of drugs might be evaluated by analyzing the interaction pattern of their targets from a network perspective [69]. Lehár et al. used large-scale simulations of bacterial metabolism to simulate the inhibitory effects of drug combinations and provided evidence that synergistic combinations are generally more specific to particular cellular phenotypes than are single agents [70]. Kwong et al. recently explored a gated signaling model that offers a new framework to identify nonobvious synergistic drug combination in NRAS-mutant melanomas [71]. Lee et al. reveal how the progressive rewiring of oncogenic signaling networks over time following EGFR inhibition leaves breast tumors vulnerable to a second and later hit with DNA-damaging drugs, demonstrating that time- and order-dependent drug combinations can be more efficacious in killing cancer cells [72].

7. Perspective

Systems biology is dramatically advancing our mechanistic understanding of disease progression and the discovery of novel therapeutics. Its continued success will depend on critical progress in both experimental and computational techniques. No single technique is sufficient to uncover the whole spectrum of gene-disease and drug-target relations in the context of biological systems. The main challenges that systems biology will confront over the next decade are the incompleteness of the available interactome data and the limitation of the existing computational tools. Our vision is that integrating the interactome with genome, transcriptome, proteome, and metabolome might offer a direction for the future advance of systems biology. New methodologies are also required to integrate diverse tools from systems biology, heterogeneous omics studies, chemoinformatics and bioinformatics. An integrated network that completely describes the underlying global paradigm of a cellular network should provide us with a deeper understanding of biological system. Clearly, there is much to do before systems biology can adequately demonstrate its usefulness in drug discovery and translational biomedicine, but the examples discussed here have provided a glimpse of the potential of systems biology.

Conflict of Interests

The authors declare that there is no conflict of interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (81101675 and 31071108), Program for New Century Excellent Talents in University (NCET-10-0600), and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Education of China (2011-508-4-3).

References

[1] H.-Y. Chuang, M. Hofree, and T. Ideker, "A decade of systems biology," *Annual Review of Cell and Developmental Biology*, vol. 26, pp. 721–744, 2010.

[2] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[3] R. P. Araujo, L. A. Liotta, and E. F. Petricoin, "Proteins, drug targets and the mechanisms they control: the simple truth about complex networks," *Nature Reviews Drug Discovery*, vol. 6, no. 11, pp. 871–880, 2007.

[4] A. Pujol, R. Mosca, J. Farrés, and P. Aloy, "Unveiling the role of network and systems biology in drug discovery," *Trends in Pharmacological Sciences*, vol. 31, no. 3, pp. 115–123, 2010.

[5] E. A. Sobie, Y.-S. Lee, S. L. Jenkins, and R. Iyengar, "Systems biology—biomedical modeling," *Science Signaling*, vol. 4, no. 190, article tr2, 2011.

[6] D. Faratian, R. G. Clyde, J. W. Crawford, and D. J. Harrison, "Systems pathology—taking molecular pathology into a new dimension," *Nature Reviews Clinical Oncology*, vol. 6, no. 8, pp. 455–464, 2009.

[7] A. Ma'ayan, "Introduction to network analysis in systems biology," *Science Signaling*, vol. 4, no. 190, article tr5, 2011.

[8] M. Vidal, M. E. Cusick, and A.-L. Barabási, "Interactome networks and human disease," *Cell*, vol. 144, no. 6, pp. 986–998, 2011.

[9] X. Chang, T. Xu, Y. Li, and K. Wang, "Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of "date" and "party" hubs," *Scientific Reports*, vol. 3, article 1691, 2013.

[10] B. B. Aldridge, J. M. Burke, D. A. Lauffenburger, and P. K. Sorger, "Physicochemical modelling of cell signalling pathways," *Nature Cell Biology*, vol. 8, no. 11, pp. 1195–1203, 2006.

[11] E. A. Sobie, "An introduction to dynamical systems," *Science Signaling*, vol. 4, no. 191, article tr6, 2011.

[12] S. R. Neves, "Developing models in virtual cell," *Science Signaling*, vol. 4, no. 192, article tr12, 2011.

[13] S. R. Neves, "Obtaining and estimating kinetic parameters from the literature," *Science Signaling*, vol. 4, no. 191, article tr8, 2011.

[14] K. D. Costa, S. H. Kleinstein, and U. Hershberg, "Biomedical model fitting and error analysis," *Science Signaling*, vol. 4, no. 192, article tr9, 2011.

[15] J. Fisher and T. A. Henzinger, "Executable cell biology," *Nature Biotechnology*, vol. 25, no. 11, pp. 1239–1249, 2007.

[16] R. Zhang, M. V. Shah, J. Yang et al., "Network model of survival signaling in large granular lymphocyte leukemia," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 42, pp. 16308–16313, 2008.

[17] J. Saez-Rodriguez, L. G. Alexopoulos, J. Epperlein et al., "Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction," *Molecular Systems Biology*, vol. 5, article 331, 2009.

[18] W. Materi and D. S. Wishart, "Computational systems biology in drug discovery and development: methods and applications," *Drug Discovery Today*, vol. 12, no. 7-8, pp. 295–303, 2007.

[19] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–8690, 2007.

[20] B. Linghu, E. S. Snitkin, Z. Hu, Y. Xia, and C. DeLisi, "Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network," *Genome Biology*, vol. 10, no. 9, article R91, 2009.

[21] S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie, and A. J. Butte, "Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets," *PLoS Computational Biology*, vol. 6, no. 2, Article ID e1000662, 2010.

[22] G. Wu, X. Feng, and L. Stein, "A human functional protein interaction network and its application to cancer data analysis," *Genome Biology*, vol. 11, no. 5, article R53, 2010.

[23] D. Hwang, I. Y. Lee, H. Yoo et al., "A systems approach to prion disease," *Molecular Systems Biology*, vol. 5, article 252, 2009.

[24] N. Yosef, A. K. Shalek, J. T. Gaublomme et al., "Dynamic regulatory network controlling TH17 cell differentiation," *Nature*, vol. 496, no. 7446, pp. 461–468, 2013.

[25] I. Arnit, M. Garber, N. Chevrier et al., "Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses," *Science*, vol. 326, no. 5950, pp. 257–263, 2009.

[26] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, article 140, 2007.

[27] I. W. Taylor, R. Linding, D. Warde-Farley et al., "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nature Biotechnology*, vol. 27, no. 2, pp. 199–204, 2009.

[28] D. Breitkreutz, L. Hlatky, E. Rietman, and J. A. Tuszynski, "Molecular signaling network complexity is correlated with cancer patient survivability," *Proceedings of the National Academy of Sciences*, vol. 109, no. 23, pp. 9209–9212, 2012.

[29] H. Y. Chuang, L. Rassenti, M. Salcedo et al., "Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression," *Blood*, vol. 120, no. 13, pp. 2639–2649, 2012.

[30] T. Huang, J. Wang, Y.-D. Cai, H. Yu, and K.-C. Chou, "Hepatitis c virus network based classification of hepatocellular cirrhosis and carcinoma," *PLoS ONE*, vol. 7, no. 4, Article ID e34460, 2012.

[31] J. N. Andersen, S. Sathyanarayanan, A. Di Bacco et al., "Pathway-based identification of biomarkers for targeted therapeutics: personalized oncology with PI3K pathway inhibitors," *Science Translational Medicine*, vol. 2, no. 43, p. 43ra55, 2010.

[32] M. S. Carro, W. K. Lim, M. J. Alvarez et al., "The transcriptional network for mesenchymal transformation of brain tumours," *Nature*, vol. 463, no. 7279, pp. 318–325, 2010.

[33] K. D. Bromberg, A. Ma'ayan, S. R. Neves, and R. Iyengar, "Design logic of a cannabinoid receptor signaling network that triggers neurite outgrowth," *Science*, vol. 320, no. 5878, pp. 903–909, 2008.

[34] J. Saez-Rodriguez, L. G. Alexopoulos, M. S. Zhang, M. K. Morris, D. A. Lauffenburger, and P. K. Sorger, "Comparing signaling networks between normal and transformed hepatocytes using discrete logical models," *Cancer Research*, vol. 71, no. 16, pp. 5400–5411, 2011.

[35] O. Rozenblatt-Rosen, R. C. Deo, M. Padi et al., "Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins," *Nature*, vol. 487, pp. 491–495, 2012.

[36] B. Zhang, C. Gaiteri, L. G. Bodea et al., "Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease," *Cell*, vol. 153, no. 3, pp. 707–720, 2013.

[37] E. C. Stites, P. C. Trampont, Z. Ma, and K. S. Ravichandran, "Network analysis of oncogenic Ras activation in cancer," *Science*, vol. 318, no. 5849, pp. 463–467, 2007.

[38] D. J. Klinke II, "Signal transduction networks in cancer: quantitative parameters influence network topology," *Cancer Research*, vol. 70, no. 5, pp. 1773–1782, 2010.

[39] B. Liu, J. Zhang, P. Y. Tan et al., "A computational and experimental study of the regulatory mechanisms of the complement system," *PLoS Computational Biology*, vol. 7, no. 1, Article ID e1001059, 2011.

[40] M. A. Pujana, J.-D. J. Han, L. M. Starita et al., "Network modeling links breast cancer susceptibility and centrosome dysfunction," *Nature Genetics*, vol. 39, no. 11, pp. 1338–1349, 2007.

[41] K. M. Mani, C. Lefebvre, K. Wang et al., "A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas," *Molecular Systems Biology*, vol. 4, article 169, 2008.

[42] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Molecular Systems Biology*, vol. 4, article 189, 2008.

[43] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS Computational Biology*, vol. 6, no. 1, Article ID 1000641, 2010.

[44] S. Zhao and R. Iyengar, "Systems pharmacology: network analysis to identify multiscale mechanisms of drug action," *Annual Review of Pharmacology and Toxicology*, vol. 52, pp. 505–521, 2012.

[45] M. A. Yıldırım, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal, "Drug-target network," *Nature Biotechnology*, vol. 25, no. 10, pp. 1119–1126, 2007.

[46] G. V. Paolini, R. H. B. Shapland, W. P. Van Hoorn, J. S. Mason, and A. L. Hopkins, "Global mapping of pharmacological space," *Nature Biotechnology*, vol. 24, no. 7, pp. 805–815, 2006.

[47] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," *Nature Chemical Biology*, vol. 4, no. 11, pp. 682–690, 2008.

[48] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nature Biotechnology*, vol. 25, no. 2, pp. 197–206, 2007.

[49] J. Li, X. Zhu, and J. Y. Chen, "Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts," *PLoS Computational Biology*, vol. 5, no. 7, Article ID e1000450, 2009.

[50] M. J. Keiser, V. Setola, J. J. Irwin et al., "Predicting new molecular targets for known drugs," *Nature*, vol. 462, no. 7270, pp. 175–181, 2009.

[51] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork, "Drug target identification using side-effect similarity," *Science*, vol. 321, no. 5886, pp. 263–266, 2008.

[52] M. Takarabe, M. Kotera, Y. Nishimura, S. Goto, and Y. Yamanishi, "Drug target prediction using adverse event report systems: a pharmacogenomic approach," *Bioinformatics*, vol. 28, no. 18, pp. i611–i618, 2012.

[53] F. Cheng, C. Liu, J. Jiang et al., "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Computational Biology*, vol. 8, no. 5, Article ID e1002503, 2012.

[54] E. Lounkine, M. J. Keiser, S. Whitebread et al., "Large-scale prediction and testing of drug activity on side-effect targets," *Nature*, vol. 486, no. 7403, pp. 361–367, 2012.

[55] M. Kuhn, M. Al Banchaabouchi, M. Campillos et al., "Systematic identification of proteins that elicit drug side effects," *Molecular Systems Biology*, vol. 9, p. 663, 2013.

[56] L. Yang, J. Chen, and L. He, "Harvesting candidate genes responsible for serious adverse drug reactions from a chemical-protein interactome," *PLoS Computational Biology*, vol. 5, no. 7, Article ID e1000441, 2009.

[57] L. Yang, K. Wang, J. Chen et al., "Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome—clozapine-induced agranulocytosis as a case study," *PLoS Computational Biology*, vol. 7, no. 3, Article ID e1002016, 2011.

[58] R. L. Chang, L. Xie, L. Xie, P. E. Bourne, and B. Ø. Palsson, "Drug off-target effects predicted using structural analysis in the context of a metabolic network model," *PLoS Computational Biology*, vol. 6, no. 9, Article ID e1000938, 2010.

[59] L. Chen, J. Lu, J. Zhang, K.-R. Feng, M.-Y. Zheng, and Y.-D. Cai, "Predicting chemical toxicity effects based on chemical-chemical interactions," *PLoS ONE*, vol. 8, no. 2, Article ID e56517, 2013.

[60] F. Iorio, R. Bosotti, E. Scacheri et al., "Discovery of drug mode of action and drug repositioning from transcriptional responses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 33, pp. 14621–14626, 2010.

[61] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," *Molecular Systems Biology*, vol. 7, article 496, 2011.

[62] M. Sirota, J. T. Dudley, J. Kim et al., "Discovery and preclinical validation of drug indications using compendia of public gene expression data," *Science Translational Medicine*, vol. 3, no. 96, p. 96ra77, 2011.

[63] G. Jin, C. Fu, H. Zhao, K. Cui, J. Chang, and S. T. C. Wong, "A novel method of transcriptional response analysis to facilitate drug repositioning for cancer therapy," *Cancer Research*, vol. 72, no. 1, pp. 33–44, 2012.

[64] M. Iskar, G. Zeller, P. Blattmann et al., "Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding," *Molecular Systems Biology*, vol. 9, p. 662, 2013.

[65] J. Jia, F. Zhu, X. Ma, Z. W. Cao, Y. X. Li, and Y. Z. Chen, "Mechanisms of drug combinations: interaction and network perspectives," *Nature Reviews Drug Discovery*, vol. 8, no. 2, pp. 111–128, 2009.

[66] J. B. Fitzgerald, B. Schoeberl, U. B. Nielsen, and P. K. Sorger, "Systems biology and combination therapy in the quest for clinical efficacy," *Nature Chemical Biology*, vol. 2, no. 9, pp. 458–466, 2006.

[67] S. Iadevaia, Y. Lu, F. C. Morales, G. B. Mills, and P. T. Ram, "Identification of optimal drug combinations targeting cellular networks: integrating phospho-proteomics and computational network analysis," *Cancer Research*, vol. 70, no. 17, pp. 6704–6714, 2010.

[68] J. Zou, S.-D. Luo, Y.-Q. Wei, and S.-Y. Yang, "Integrated computational model of cell cycle and checkpoint reveals different essential roles of Aurora-A and Plk1 in mitotic entry," *Molecular BioSystems*, vol. 7, no. 1, pp. 169–179, 2011.

[69] J. Zou, P. Ji, Y.-L. Zhao et al., "Neighbor communities in drug combination networks characterize synergistic effect," *Molecular BioSystems*, vol. 8, no. 12, pp. 3185–3196, 2012.

[70] J. Lehár, A. S. Krueger, W. Avery et al., "Synergistic drug combinations tend to improve therapeutically relevant selectivity," *Nature Biotechnology*, vol. 7, pp. 659–666, 2009.

[71] L. N. Kwong, J. C. Costello, H. Liu et al., "Oncogenic NRAS signaling differentially regulates survival and proliferation in melanoma," *Nature Medicine*, vol. 18, no. 10, pp. 1503–1510, 2012.

[72] M. J. Lee, A. S. Ye, A. K. Gardino et al., "Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks," *Cell*, vol. 149, no. 4, pp. 780–794, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 795676, 12 pages
http://dx.doi.org/10.1155/2013/795676

Research Article

Study of MicroRNAs Related to the Liver Regeneration of the Whitespotted Bamboo Shark, *Chiloscyllium plagiosum*

Conger Lu, Jie Zhang, Zuoming Nie, Jian Chen, Wenping Zhang, Xiaoyuan Ren, Wei Yu, Lili Liu, Caiying Jiang, Yaozhou Zhang, Jiangfeng Guo, Wutong Wu, Jianhong Shu, and Zhengbing Lv

Institute of Biochemistry, College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China

Correspondence should be addressed to Zhengbing Lv; zhengbingl@zstu.edu.cn

Received 13 June 2013; Accepted 28 July 2013

Academic Editor: Lei Chen

To understand the mechanisms of liver regeneration better to promote research examining liver diseases and marine biology, normal and regenerative liver tissues of *Chiloscyllium plagiosum* were harvested 0 h and 24 h after partial hepatectomy (PH) and used to isolate small RNAs for miRNA sequencing. In total, 91 known miRNAs and 166 putative candidate (PC) miRNAs were identified for the first time in *Chiloscyllium plagiosum*. Through target prediction and GO analysis, 46 of 91 known miRNAs were screened specially for cellular proliferation and growth. Differential expression levels of three miRNAs (xtr-miR-125b, fru-miR-204, and hsa-miR-142-3p_R-1) related to cellular proliferation and apoptosis were measured in normal and regenerating liver tissues at 0 h, 6 h, 12 h, and 24 h using real-time PCR. The expression of these miRNAs showed a rising trend in regenerative liver tissues at 6 h and 12 h but exhibited a downward trend compared to normal levels at 24 h. Differentially expressed genes were screened in normal and regenerating liver tissues at 24 h by DDRT-PCR, and ten sequences were identified. This study provided information regarding the function of genes related to liver regeneration, deepened the understanding of mechanisms of liver regeneration, and resulted in the addition of a significant number of novel miRNAs sequences to GenBank.

1. Introduction

Cartilaginous fish are jawed vertebrates that diverged from the common ancestor of humans and teleost fish approximately 530 million years ago [1]. Similar to teleost fish, cartilaginous fishes possess complex physiological systems, such as an adaptive immune system and a pressurized circulatory system [2]. The white spotted bamboo shark (*Chiloscyllium plagiosum*) is a cartilaginous fish that is widely distributed in cold seas along the coasts of the eastern Pacific. Economically, this shark is one of the most important marine animals, possessing both scientific and commercial food values. The liver of *C. plagiosum*, accounting for 75% of the weight of the viscera, possesses immune regulatory functions and contains bioactive substances. Several investigators have isolated and cloned stimulatory factors related to liver regeneration and immune regulation from *Chiloscyllium plagiosum* [3–6], but there have not yet been miRNAs reported for this species.

MicroRNAs (miRNAs) are a family of small, noncoding, and single-stranded RNAs that consist of approximately 19–25 nucleotides derived from the stem regions of hairpin transcripts (referred to as "pre-miRNAs"). These small RNAs act as regulators, leading to either mRNA cleavage or translational repression by complementarily hybridization to the 3′-untranslated regions (3′ UTRs) of their target mRNAs. The formation of mature miRNAs involves several steps. First, most miRNAs are transcribed as long primary transcripts (pri-miRNAs) by RNA Pol II, although a minor group of miRNAs associated with Alu repeats can be transcribed by RNA Pol III. The primary transcript is processed by the RNase III enzyme Drosha in the nucleus to give one or more 60–100 nt long hairpin precursor sequences (pre-miRNAs). Together with DiGeorge syndrome critical region gene 8 (DGCR8), Drosha forms a large complex known as the Microprocessor complex; DGCR8 interacts with pri-miRNAs and assists Drosha in cleaving the substrate. The hairpin

precursor is then exported by Exportion-5 to the cytoplasm, where the mature miRNA is excised by another RNase III enzyme, Dicer. Following Dicer cleavage, the resulting ≈22 nt RNA duplex is loaded into the Ago protein to create the effector complex, RISC. One strand of the ≈22 nt RNA duplex remains with Ago as a mature miRNA and targets the 3′ UTR complementary pairing area of the target gene to approximately 2–8 bases of the 5′ end of the miRNA [7–9]. Negative posttranscriptional regulation by miRNAs has been observed in various biological processes, including viral disease [10], lipid metabolism [11], and cellular proliferation and differentiation [12].

Research into the mechanisms of liver cell growth and apoptosis has focused on individual signal molecules, important transcription factors, and related target genes and signaling pathways. Research has been less comprehensive, however, regarding the more analytical aspects of gene expression regulation. A substantial body of evidence has indicated that miRNAs are abundant in the liver and play an important role in liver development, disease, and regeneration [13–15]. In this study, we carried out deep sequencing in combination with bioinformatics and genomic sequencing to identify miRNAs in *Chiloscyllium plagiosum* liver. Although sequencing of *Chiloscyllium plagiosum* genome has not yet been completed, the *Callorhinchus milii* genome was recently proposed to be a model for cartilaginous fish genomes [16]. Therefore, we used the *Callorhinchus milii* genome to predict pre-miRNA hairpin structures. In this study, we characterized small RNAs and identified miRNAs in the *C. plagiosum* liver. We further analyzed miRNA clusters and the expression of *Chiloscyllium plagiosum* miRNAs. Our results may contribute to liver miRNA research not only for fish but also for other species. This is the first comprehensive report about the miRNA of cartilaginous fish and adds several novel miRNAs to the database, which will help to identify homologs in other species. In addition, screening differentially expressed genes in normal and regenerating liver tissues by DDRT-PCR will provide valuable information for further study of the newly discovered functional genes related to liver regeneration and the relationship between differentially expressed genes and miRNAs. We have laid the foundation for further understanding of the function of miRNAs in liver regulation in *Chiloscyllium plagiosum*.

2. Materials and Methods

2.1. Sample Preparation and Liver Excision. A live *Chiloscyllium plagiosum* was purchased from a Guangdong aquatic product market. Two-thirds of the liver was surgically removed, using a 2% chloral hydrate solution (300 mg per Kg weight of shark) as an anesthetic. The wound was then sutured, and the shark was placed into seawater to continue feeding. Regenerating liver tissues were obtained 0 h and 24 h after the partial hepatectomy (PH) by opening the suture and removing the cut border of the left liver. These experiments complied with the relevant national animal welfare laws and were conducted under guidelines approved by the China Wildlife Conservation Association.

2.2. Small RNA Library Construction and Deep Sequencing. The total RNA of the liver was extracted using the mirVana miRNA Isolation Kit (Ambion, Austin, USA) according to the manufacturer's protocol. Small RNAs were size fractionated and ligated to the SRA 5′ adapter. This step was followed by SRA 3′ adapter ligation and size fractionation to isolate RNAs of 64–99 nt. The RNAs were then converted to single-stranded cDNA using Superscript II reverse transcriptase (Invitrogen, CA, USA) and Illumina's small RNA RT Primer. This pool was once again size fractionated to isolate cDNA in the 80–115 bp range that contained miRNAs. Next, the cDNA was PCR-amplified for 20 cycles using Illumina's small RNA primer set and Phusion polymerase (New England Lab, USA). The PCR products were size fractionated and recovered for sequencing on a Genome Analyzer GA-II (Illumina, San Diego, USA) following the vendor's recommended protocol for small RNA sequencing. Following real-time sequencing, image analysis and base calling by Illumina's Real-Time Analysis version 1.8.70 (RTA v1.8.70), raw sequencing reads were obtained using Illumina's Sequencing Control Studio software version 2.8 (SCS v2.8). The extracted sequencing reads were then used for standard data analysis.

2.3. Standard Data Analysis. Following the removal of adapter sequences and low-quality reads, clean reads between 15 and 30 bases in length were processed for bioinformatic analysis. A proprietary pipeline script, ACGT101-miR v4.2 (LC Sciences) [17, 18], was used for the sequencing data analysis. Various "mappings" were performed with unique sequences against pre-miRNAs and mature miRNAs sequences from selected species (*Danio rerio, Fugu rubripes, Oryzias latipes, Xenopus laevis, Xenopus tropicalis, Homo sapiens*, and *Mus musculus*) listed in the miRBase v19.0 (http://www.mirbase.org/) as well as the genome of *Callorhinchus milii* (http://esharkgenome.imcb.a-star.edu.sg/resources.html). The mapping process is presented in Figure 1. Flanking sequences of mapped reads were subjected to secondary structure analysis to predict pre-miRNA sequences using Mfold software. The criteria used for miRNA annotation and hairpin structure determination are presented in Table 1 [19, 20].

2.4. Target Prediction for miRNAs and GO Analysis. We compared 1839 ESTs of *Chiloscyllium plagiosum* with 1344 ESTs with 98% similarity using the cd-hit-454 program [21, 22]. Ninety-one mature miRNA sequences were used as custom sequences to search ESTs and perform target prediction using the program miRanda v1.0b [23] with the following parameter settings: Gap Open Penalty: −8.0; Gap Extend:−2.0; Score Threshold: 50.0; Energy Threshold: −20.0 kcal/mol; Scaling Parameter: 2.0. The target sequences were annotated using BLAST to the GO database (http://www.geneontology.org/), thus predicting which processes the miRNAs participate in.

2.5. Stem-Loop RT-PCR. Seven candidate miRNAs were chosen for verification of expression using stem-loop RT-PCR. Four PC miRNAs were randomly selected (PC-3p-186_7748, PC-5p-970_1302, PC-5p-108_13860, and PC-5p-14_74120),

Table 1: Criteria used for miRNAs annotation and hairpin structure determination.

miRNAs annotation
The miR_name is composed of the known miR name in a cluster, an underscore, and a matching annotation such as
L–n means that the miRNAs_seq (detected) is n base less than known rep_miRSeq in the left side;
R–n means that the miRNAs_seq (detected) is n base less than known rep_miRSeq in the right side;
L+n means that the miRNAs_seq (detected) is n base more than known rep_miRSeq in the left side;
R+n means that the miRNAs_seq (detected) is n base more than known rep_miRSeq in the right side;
2ss5TC13TA means 2 substitutin (ss), which are T $\rightarrow$ C at position 5 and T $\rightarrow$ A at position 1.
The miRNAs_seq (detected) is exactly the same as known rep_miRSeq; miRNAs_name is the name of representative miRNAs.
Hairpin determination
Definition of MFEI: MFEI = $-dG * 100/\text{mirLen}/\text{CG}\%$.
Criteria:
(1) free energy (dG in kCal/mol) ≤ -15,
(2) length of hairpin (up and down stem + terminal loop) ≥ 50,
(3) number of basepairs (bp) in stem region ≥ 16,
(4) length of terminal loop ≤ 20,
(5) number of basepairs (bp) in mature or mature $*$ region ≥ 12,
(6) percentage of small RNA in stem region (pm) $\geq 80\%$,
(7) number of allowed errors in mature region ≤ 7,
(8) number of allowed errors in one bulge in stem ≤ 12,
(9) number of allowed errors in one bulge in mature region ≤ 8,
(10) number of allowed biased errors in one bulge in mature region ≤ 4,
(11) number of allowed biased bulges in mature region ≤ 2,
(12) MFEI ≥ 0.7.

and 3 known miRNAs were also analyzed: fru-miR-204a, hsa-miR-142-3p_R-1, and fru-miR-126. Total RNA was extracted using Trizol reagent and reverse transcribed using a Reverse Transcription System (Promega). The 5 μL reactions contained 250 ng of RNA sample and 2 μM of stem-loop RT primer (0.5 μL). The reactions were incubated for 45 min at 42°C and 5 min at 95°C and then held at 4°C. 20 μL PCR reactions (2× Taq PCR Master Mix, Lifefeng) were run using 0.5 μL of cDNA product by incubating for 5 min at 95°C, followed by 40 cycles of 94°C for 30 sec, 60°C for 30 sec, and 72°C for 30 sec. The PCR products were detected with gel electrophoresis. The positive control was 18s rRNA [4]. All of the primers that were used for RT-PCR are listed in Table 2. The reverse primers for PC-3p-186_7748 and fru-miR-204a were miR-Reverse1, and fru-miR-126 miR-Reverse3, respectively, while the remaining 4 miRNAs used miR-Reverse2.

2.6. Real-Time PCR Analysis. According to our literature search, the small RNA sequencing results, miRNA target gene results, and GO analysis, three known miRNAs (hsa-miR-142-3p_R-1, xtr-miR-125b, and fru-miR-204) related to liver regeneration and liver diseases were assayed in normal and regenerating liver tissue at 0 h, 6 h, 12 h, and 24 h using real-time PCR. A FastStart Universal SYBR Green Master (ROX) fluorescence quantitative PCR reaction kit (Roche) was used, and a 20 μL reaction system was designed according to the kit instructions. Each sample was repeated 3 times, and the cycling conditions were as follows: 95°C for 2 min followed by 40 cycles of 95°C for 15 s and 60°C for 30 s. A melting curve analysis was used to ensure the specificity of real-time PCR primers.

2.7. DDRT-PCR. The total RNA of normal and regenerating liver tissues at 24 h was extracted using Trizol and was reverse transcribed using a Reverse Transcription System (Promega). The 25 μL reactions contained 1250 ng of RNA and 50 μM of anchor primer (2.5 μL). The reactions were incubated for 45 min at 42°C and 5 min at 95°C and then held at 4°C. 10 μL PCR reactions (2× Taq PCR Master Mix, Lifefeng) were performed using two anchor primers and twenty random primers. All of the primers used for DDRT-PCR are listed in Table 3. The 10 μL reactions contained 5 μL of 2× Taq PCR Master Mix, 0.1 μL 50 μmol/L anchor primer, 0.5 μL 10 μmol/L random primer, 0.5 μL cDNA, and 3.9 μL ddH_2O. The reactions were incubated for 5 min at 95°C, followed by 40 cycles of 95°C for 30 sec, 40°C for 2 min, and 72°C for 1 min. The PCR products were detected by nondenaturing polyacrylamide gel electrophoresis and silver staining.

2.8. Cloning and Sequencing. The target fragments were recovered using the crushing and soaking method. The screened differential genes were used as templates, and the corresponding anchor primers and random primers were

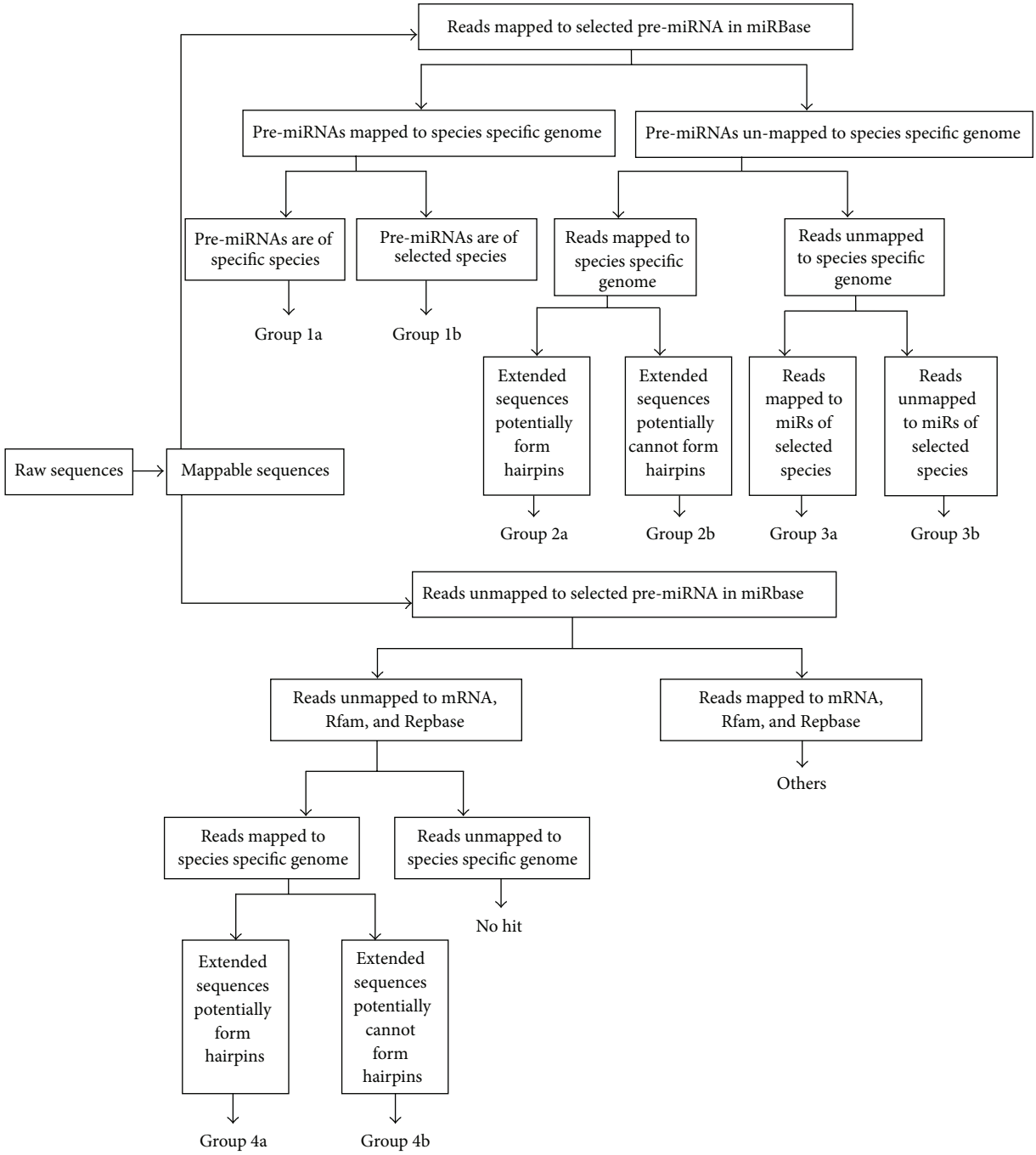


Figure 1: Procedure for mapping reads to database sequences. (1) specific species, that is, Callorhinchus milii; (2) selected species: *Fugu rubripes, Oryzias latipes, Xenopus laevis, Xenopus tropicalis, Homo sapiens,* and *Mus musculus.*

used as primers for PCR. The PCR reaction system and procedure were the same as the previous ones. After assaying for and recovering the target differential fragments, they were cloned into the pMD18-T simple vector to construct recombinant plasmids. The differential genes were tested through bacterial PCR. The 10 μL reactions contained 5 μL 2× Taq PCR Master Mix, 0.4 μL 10 μmol/L PCR Forward Primer, 0.4 μL 10 μmol/L PCR Reverse Primer, 0.5 μL cDNA and 3.9 μL ddH_2O. The reactions were incubated for 5 min at 95°C, followed by 40 cycles of 95°C for 30 sec, 40°C for 2 min, and 72°C for 1 min. DNA sequencing was performed by Shanghai Sunny Biotechnology Co., Ltd. The obtained target differential genes were matched to NCBI EST library sequences through BLAST.

3. Results

3.1. Screening miRNAs in Chiloscyllium plagiosum Liver. Raw reads in which the 3 adapters were not found and those with fewer than 15 bases after the 3 adapters were deleted;

Table 2: Primers used in RT-PCR experiment.

Gene	Primer
PC-3p-186_7748-RT	5′-GTCGTATCCAGTGCAGGGTCCGAG GTATTCGCACTGGATACGACATAGGC-3′
PC-5p-108_13860-RT	5′-CTCAACTGGTGTCGTGGAGTCGGC AATTCAGTTGAGAGCTTT-3′
PC-5p-14_74120-RT	5′-CTCAACTGGTGTCGTGGAGTCGGCA ATTCAGTTGAGATTCTC-3′
PC-5p-970_1302-RT	5′-CTCAACTGGTGTCGTGGAGTCGGCAA TTCAGTTGAGGGACAT-3′
hsa-miR-142-3p_R-1-RT	5′-CTCAACTGGTGTCGTGGAGTCGGCA ATTCAGTTGAGCCATAA-3′
fru-miR-126-RT	5′-CTCAACTGAATTGCCGACTCCACGACACCAGTTGAGGCATTA-3′
fru-miR-204-RT	5′-GTCGTATCCAGTGCAGGGTCCGAGG TATTCGCACTGGATACGACATGCCT-3′
PC-3p-186_7748-Forward	5′-GCACTTGTCCCAGCCTATGTC-3′
PC-5p-108_13860-Forward	5′-CCGTGTAAACATCTGTAACTGAAAG-3′
PC-5p-14_74120-Forward	5′-GCGAGCACTGTCTAACCTGAG-3′
PC-5p-970_1302-Forward	5′-GGTCCAACCTCTTATGTCCCTC-3′
hsa-miR-142-3p_R-1-Forward	5′-GGCTTTAGTGTTTCCTACTTTATGG-3′
fru-miR-126-Forward	5′-TCCTACCGTGAGTAATAATGCC-3′
fru-miR-204-Forward	5′-CGTGTTCCCTTTGTCATCCTATC-3′
18s rRNA-Forward	5′-GACTCAACACGGGAAACCTCA-3′
18s rRNA-Reverse	5′-CAGACAAATCGCTCCACCAA-3′
miR-Reverse1	5′-AGCAGGGTCCGAGGTATTC-3′
miR-Reverse2	5′-GTGTCGTGGAGTCGGCAAT-3′
miR-Reverse3	5′-AACTGAATTGCCGACTCCAC-3′

Table 3: The primers of DDRT-PCR.

Primer	Sequence	Primer	Sequence
Anchor primer 1	5′-AGCTTTTTTTTTTTVA-3′	Anchor primer 2	5′-AGCTTTTTTTTTTTVC-3′
Random primer 1	5′-GCTAACGATG-3′	Random primer 2	5′-TGGATTGGTC-3′
Random primer 3	5′-CTTTCTACCG-3′	Random primer 4	5′-TTTTGGCTCC-3′
Random primer 5	5′-GGAACCAATG-3′	Random primer 6	5′-AAACTCCGTC-3′
Random primer 7	5′-TCGATACAGG-3′	Random primer 8	5′-TGGTAAAGGG-3′
Random primer 9	5′-TCGGTCATAG-3′	Random primer 10	5′-CTGCTTGATG-3′
Random primer 11	5′-TACCTAAGCG-3′	Random primer 12	5′-CTGCTTGATG-3′
Random primer 13	5′-GTTTTCGCAG-3′	Random primer 14	5′-GATCAAGTCC-3′
Random primer 15	5′-GATCCAGTAC-3′	Random primer 16	5′-GCTCACGTAG-3′
Random primer 17	5′-GATCTGACAC-3′	Random primer 18	5′-GCTATCAGAC-3′
Random primer 19	5′-GATCATAGCG-3′	Random primer 20	5′-GATCAATCGC-3′

junk reads were removed. Following these changes, 15,361,801 and 10,146,583 reads remained in the *Chiloscyllium plagiosum* normal and regenerating livers at 0 h and 24 h after PH. The length distribution of the small RNA reads ranged from 15 to 30 nt (Figure 2). Through mapping analysis, 15,361,801 and 10,146,583 reads could be distributed into the groups that are presented in Table 4.

Two hundred and fifty-seven miRNAs were identified in *Chiloscyllium plagiosum*, including 91 known miRNAs and 166 *Chiloscyllium plagiosum* putative candidate (PC) miRNAs. In addition, 31 PC miRNAs were found only in 24 h (after PH) regenerative liver tissue, and 49 PC miRNAs were found only in normal liver tissue. A total of 76 unique miRNA reads were mapped to the pre-miRNAs of selected species in the miRBase, and these pre-miRNAs were further mapped to the *Callorhinchus milii* genome (group 1b). Additionally, 15 miRNA reads mapped to pre-miRNAs of the selected species but not to the *Callorhinchus milii* genome. These 15 miRNA reads were mappable to the *Callorhinchus milii* genome, and the extended sequences could potentially form hairpins (group 2a). Additional 166 PC miRNAs that were identified did not map to any

Table 4: Distribution of small RNA among LN and LR24.

Category	Total RNAs in LN	Percent	Total RNAs in LR24	Percent
Total small RNAs	15361801	100%	10146583	100%
Group 1a	0	0.00%	0	0.00%
Group 1b	1600835	10.40%	985411	9.70%
Group 2a	56020	0.40%	39787	0.40%
Group 2b	25388	0.20%	12569	0.10%
Group 3a	2300528	15.00%	1113914	11.00%
Group 3b	72334	0.50%	40956	0.40%
Group 4a	310150	2.00%	153216	1.50%
Group 4b	301238	2.00%	195545	1.90%
Group 5	6032607	39.30%	4341986	42.80%
Group 6	4662701	30.40%	3263199	32.20%

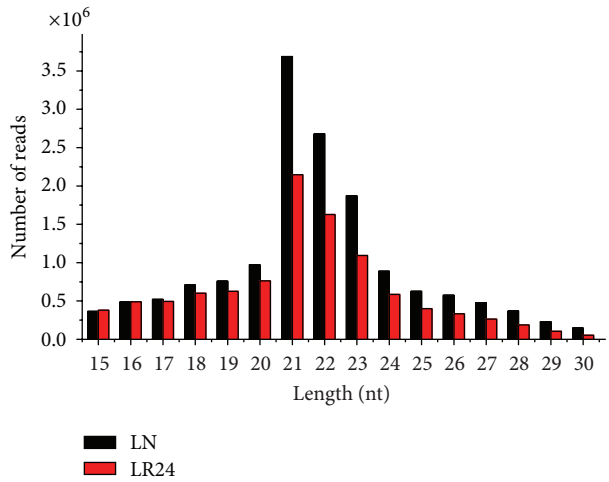


Figure 2: Length distribution of small 15-to-30 nucleotide RNAs in LN and LR24. The nucleotide (nt) lengths of the reads are indicated on the x-axis, and the number of total reads are indicated on the y-axis.

known pre-miRNAs, either in the mRNA, Rfam, or Repbase databases. These reads could map to the *Callorhinchus milii* genome, however, and exhibited the potential to form hairpin structures (group 4a). The predicted stem-loop secondary structures of 5 pre-miRNAs (fru-miR-126, hsa-miR-142-3p_R-1, ola-miR-1388-3p_R+5/ola-miR-1388-5p_R-1, PC-5p-14_74120/PC-3p-525510_1, and fru-miR-204a) from the Mfold database are presented in Figure 3. Information about all of these miRNAs is presented in supplementary Table 1; see in Supplementary Material available online at http://dx.doi.org/10.1155/2013/795676.

3.2. The Expression of Chiloscyllium plagiosum miRNAs. The average number of reads of known miRNAs in normal and regenerating liver tissues was 8173 and 8972, respectively. The average reads of PC miRNAs were 1687 and 1781, respectively. One hundred twenty-eight PC miRNA reads were less than 10 nt, which is consistent with a previous observation that nonconserved miRNAs are generally expressed at lower levels and represent tissue- or development-specific expression patterns [24]. Log_2 ratios of 116 miRNAs reads in regenerating and normal tissues were larger than 1, of which only 5 miRNAs reads were more than 10 nt, and only 8 known miRNAs were included. Therefore, both miRNA expression type and abundance have no obvious differences during liver regeneration from 0 h to 24 h.

Moreover, miRNAs are often present in the genome in clusters, where several miRNAs are aligned in the same orientation and transcribed as a polycistronic message, allowing them to act cooperatively. Based on miRBase's definition of a miRNA cluster (10,000 bp), we discovered a total of 34 clusters in the *Callorhinchus milii* genome, of which 16 clusters were generated from the 16 pre-miRNAs, and 2 mature miRNAs were cleaved from each pre-miRNA (Supplementary Table 2). The largest cluster (no. 27) contained six miRNAs.

3.3. Target Predictions for miRNAs and GO Analysis. miRNAs are involved in a wide variety of biological processes by binding target sites with seed sequences [25, 26]. Identifying the total number of target sites, especially for those low-abundance and species-specific miRNAs, is helpful for appreciating the breadth of miRNA functions [27]. We calculated the potential binding sites between miRNAs and the ESTs of *Chiloscyllium plagiosum* with miRanda to determine the potential genes targeted by miRNAs. Of 91 known miRNAs, 86 miRNAs had targeted genes; most miRNAs have more than one, and 144 miRNAs had multiple predicted target sites. Most predicted targeted genes may be regulated by more than one miRNA.

The functional classification for targeted genes was analyzed using Gene Ontology analysis. The targets that were functionally annotated in the GO database could be divided to 38 subclasses, presented in Figure 4. Forty-six miRNAs were related to cell proliferation and growth terms (Supplementary Table 3), including the following miRNAs: xtr-miR-34a_R-1 [28], fru-miR-122_R+1 [29], fru-miR-27b_R-1 [30], fru-miR-181b_R+1 [31], fru-miR-23b_R-1 [32], ola-miR-144_L-1R+2_1ss12TA [33], ola-miR-103_R+2 [29], and fru-miR-26 [34]. Based on comprehensive consideration of our literature search, the small RNA sequencing results, miRNA target gene results, and GO analysis, three known miRNAs (hsa-miR-142-3p_R-1, xtr-miR-125b, and fru-miR-204) related to liver

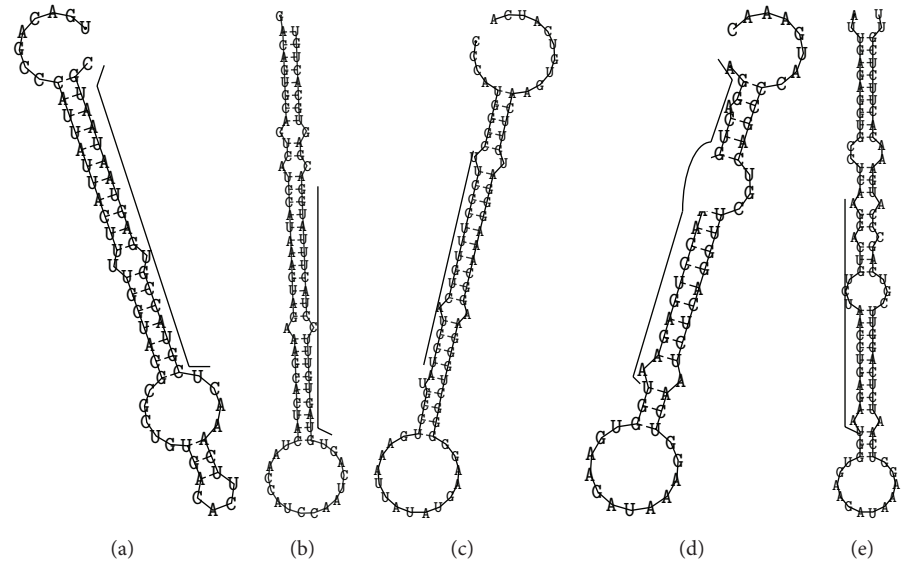


Figure 3: Secondary structures of 5 pre-miRNAs. (a) fru-miR-126, (b) hsa-miR-142-3p_R-1, (c) fru-miR-204a, (d) ola-miR-1388-3p_R+5/ola-miR-1388-5p_R-1, (e) PC-5p-14_74120/PC-3p-525510_1. The underlined letters refer to the candidate miRNA sequences. For (d) and (e), the underlined letters refer to the 5p miRNAs, and the upper-case letters refer to the 3p miRNA sequences.

regeneration and liver diseases were screened to the following experiments, and the potential binding sites between miRNAs and the ESTs of *Chiloscyllium plagiosum* with miRanda were presented in Figure 5.

3.4. Real-Time PCR Analysis of the miRNAs of Different Liver Regenerative Stages. Seven candidate miRNAs were selected for expression-level confirmation using stem-loop RT-PCR. Figure 6 shows that all of the 7 predicted miRNAs were expressed in the *Chiloscyllium plagiosum* liver. With 18s rRNA as positive control, the expression of three target miRNAs was normalized and used to analyze the expression of the miRNA at different liver regenerative stages. Figure 7 indicates that there is a large transcriptional difference between three known miRNAs at different stages of liver regeneration. Significant miRNA expression was observed in regenerating liver tissues at 6 h and 12 h. The expression level of xtr-miR-125b was the highest in regenerating liver tissues at 6 h and was 40 times higher than that in normal liver tissue. Levels of fru-miR-204, xtr-miR-125b, and has-miR-142-3p_R-1 in regenerating liver tissues at 24 h were 0.685, 0.86, and 1.586 times higher than those in normal liver tissue, respectively (Table 5). This result is consistent with the sequencing data.

3.5. Differential Gene Sequence Analysis. Forty pairs of PCR product results were analyzed by 5% polyacrylamide gel electrophoresis and silver staining, and the bands were clear. Six pairs of PCR products are presented in Figure 8, of which two obvious differential bands showed higher expression in regenerating liver. Overall, ten differential sequences were obtained through differential display analysis.

These ten differential sequences were run through BLAST. Three sequences exhibited high similarity to NCBI EST library sequences, and the other 7 sequences had no homology to any known genes. The homology of RF1-6l with 33 *Chiloscyllium griseum* cDNA clone sequences was up to 100%. The homology of NF2-2 with 2 cloudy catshark embryo cDNA library sequences was higher than 80%. The homology of RF2-6s with dogfish shark stem cell line sequences was 88%. The matching and differential expression results are presented in Table 6 and consist of 8 sequences ranging from 100 to 500 bp. Three sequences showed high similarity to NCBI EST library sequences through BLAST; 5 sequences were overexpressed, and another five were downregulated during liver regeneration.

4. Discussions

The emergence of high-throughput sequencing technology has greatly hastened the discovery of small expressed RNAs in newly analyzed and rare species. Because of the high sequence homology between miRNAs from related species and the stem-loop structure of their precursor sequences, miRNA precursor sequences can be used to perform homology screens covering the entire genome of species. These target gene sequences can be identified using RNA secondary structure analysis software (e.g., RNAfold, MirScan, and MFold) combined with dynamic analysis [35]. A combination of deep sequencing and bioinformatics analysis allows for the identification of new and rarely expressed nonconserved miRNAs, especially in less frequently studied species [36, 37].

Here, we report the first complete analysis of miRNA populations in the *Chiloscyllium plagiosum* liver using deep sequencing and bioinformatics analysis. Due to the high conservation of miRNAs between related species, we selected pre-miRNAs/miRNAs of *Danio rerio*, *Fugu rubripes*, *Oryzias latipes*, *Xenopus laevis*, *Xenopus tropicalis*, *Homo sapiens*, and *Mus musculus* to map the reads. In the present study, 257 miRNAs were identified, of which 166 were specific to *Chiloscyllium plagiosum*. These results indicate that the method was

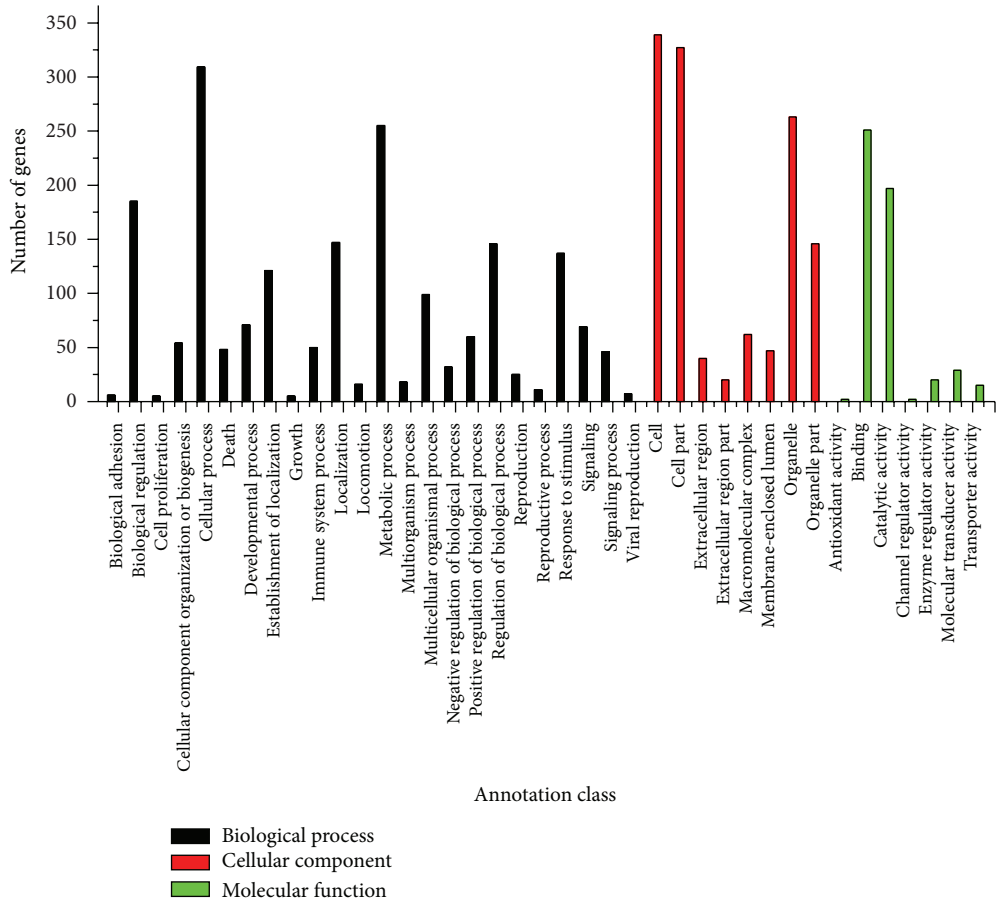


Figure 4: GO analysis of targeted genes. The classes are shown on the x-axis; the number of genes for each class is shown on the y-axis. The number of genes for the three classes—biological process, cellular component, and molecular function—is 687, 608, and 612, respectively.

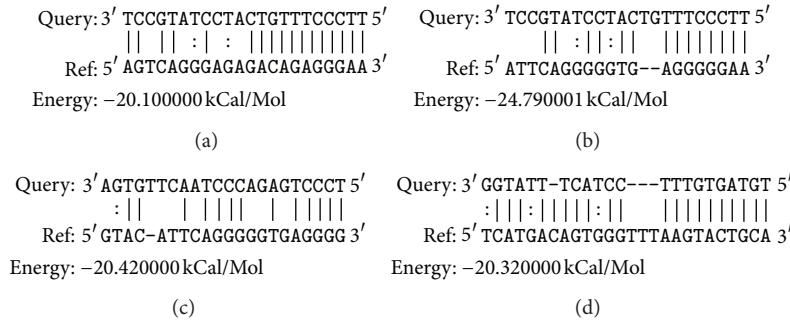


Figure 5: The potential binding sites between miRNAs and the ESTs of *Chiloscyllium plagiosum*. Query represents the miRNAs sequence, and Ref represents their predicted targets. (a) fru-miR-204; (b) fru-miR-204; (c) xtr-miR-125b; (d): has-miR-142-3p_R-1.

Table 5: Relative quantitative data analysis of miRNA in four periods regenerative liver tissues.

Gene	Average C_t value ± standard error				$2^{-\Delta\Delta C_t}$			
	LN	LR6	LR12	LR24	LN	LR6	LR12	LR24
18S rRNA	21.982 ± 0.116	21.982 ± 0.116	23.160 ± 0.016	23.267 ± 0.127	1.000	1.000	1.000	1.000
xtr-miR-125b	25.710 ± 0.141	25.710 ± 0.141	22.637 ± 0.354	26.400 ± 0.222	1.000	43.070	3.676	0.860
has-miR-142-3p_R-1	21.035 ± 0.265	21.035 ± 0.265	21.126 ± 0.239	21.395 ± 0.202	1.000	4.803	4.623	1.586
fru-miR-204a	18.316 ± 0.383	18.316 ± 0.383	18.598 ± 0.122	19.730 ± 0.135	1.000	4.209	2.226	0.685

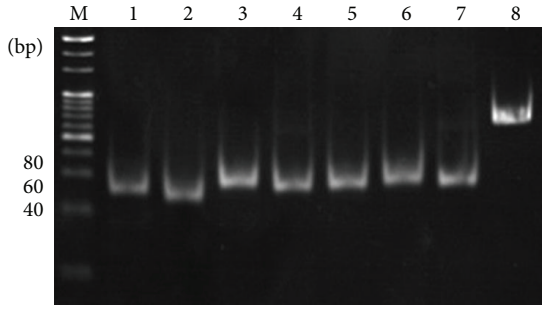


Figure 6: The identification of 7 miRNAs by stem-loop RT-PCR. The mRNA expression of 7 miRNAs was confirmed. The sizes of the PCR products were approximately 55 bp. 18s rRNA was used as a positive control. M: 20 bp DNA ladder maker; 1–12: PC-3p-186_7748, PC-5p-970_1302, PC-5p-108_13860, PC-5p-14_74120, hsa-miR-142-3p_R-1, fru-miR-204, fru-miR-126, 18s rRNA.

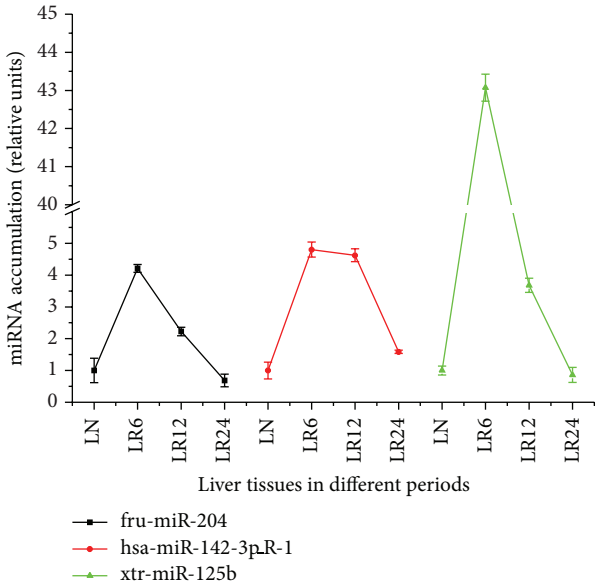


Figure 7: Relative expression changes of miRNAs in four periods of liver regeneration. The Real-time PCR results were processed by Microsoft Office Excel, and $2^{-\Delta\Delta C_t}$ was used to represent the miRNA differential expression. $\Delta C_t = C_t$ (target gene) $- C_t$ (reference gene) $\Delta\Delta C_t = C_t$ (experiment group) $- C_t$ (control group). the live issues in different periods are indicated on the x-axis, and the value of $2^{-\Delta\Delta C_t}$ showed in Table 5 are indicated on the y-axis.

Table 6: Analysis of differential expressed sequences.

Clone	Length (bp)	Query coverage	Max identity to the EST
RF1-6l	534	11%	100%
RF1-6s	258		novel
RF2-8	283		novel
NF2-2	173	94%	80+%
NF1-7	236		novel
NF1-19	326		novel
RF2-6s	244	78%	88%
RF2-6l	278		novel
NF2-19	77		novel
NF1-9	208		novel

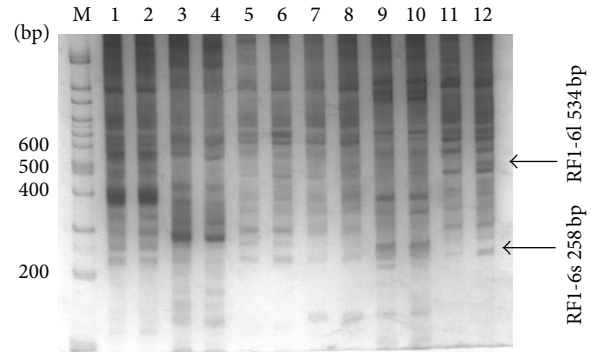


Figure 8: Results of DDRT-PCR. M: 100 bp DNA ladder marker; Lanes 1–12: NF1-1, RF1-1, NF1-2, RF1-2, NF1-3, RF1-3, NF1-4, RF1-4, NF1-5, RF1-5, NF1-6, RF1-6.

effective for identifying low-abundance and species-specific small RNAs. Comparative analysis of the *C. milii* genome with the whole-genome assemblies of *Danio rerio*, *Fugu rubripes*, and *Homo sapiens* suggested that (i) noncoding sequences in *Callorhinchus milii* are evolving more slowly than in teleost fishes and that (ii) the *Callorhinchus milii* genome has experienced fewer chromosomal rearrangements compared with teleost fish genomes. Although cartilaginous fish diverged from the human lineage before teleost fish, a higher proportion of regulatory elements are conserved between cartilaginous fish and humans than between teleost fish and humans [16, 38]. Based on conclusions that were drawn from miRNA analysis, our research conclusively demonstrated that cartilaginous and teleost fishes may not have as close a genetic relationship as previously thought.

In the present study, we observed that the frequency of most PC miRNAs was extremely low. This observation is in accordance with the low expression observed for nonconservative miRNAs and for those involved in organizational or special development periods. Moreover, miRNA genes often form clusters in the genome. Certain miRNAs in genes clusters share the same set of control sequences and are found on the same transcript, whereas some miRNAs in gene clusters are transcribed from independent transcripts, such as miR-433 and miR-127 [39]. An miRNA gene can produce two different miRNAs through duplex transcription and thus control different target genes [40]. A total of 34 miRNA gene clusters were identified in this study, of which 16 produced 32 miRNA groups through duplex transcription. Each of these duplex groups produced 2 miRNAs. This finding provides important clues for probing the function of these miRNAs in future research.

Through miRNA target prediction and GO analysis, we focused on the relation between miRNAs and their targets, thus focusing on mechanisms that involved cell proliferation, immune system processes, and other biological functions. However, as the genome of *Chiloscyllium plagiosum* is not fully sequenced and annotated, it was difficult to determine whether these miRNA targets have any functional bias.

It should be noted that many targets could not be annotated in the GO database due to their low homology to orthologs; thus, the miRNA-target network is likely more complex than the results presented here. MiRNA target genes were mostly predicted through bioinformatics software, and miRNA binding sites in most species are in 3′-untranslated regions of the target gene. However, it has been reported that siRNA-mediated translational inhibition can be induced by an incomplete complementary site in the ORF region of a mammalian reporter gene, which means that the target sequence of a given miRNA is not only in the 3′-untranslated region. Because the miRNA and genome of *Chiloscyllium plagiosum* have not been reported and the NCBI database includes just 33 mRNAs, the ESTs and these 33 mRNAs of *Chiloscyllium plagiosum* were used to predict miRNA target genes. The target genes should also be tested using a dual luciferase reporter gene assay system, western blot, or some other experimental method. Through target predictions for miRNAs and GO analysis, we can predict that the target genes are related to thirty-eight biological functions related to "molecular function," "cellular component," or "biological process." A total of 46 related miRNA were screened for cell proliferation and growth, and only eight miRNA have been reported in liver regeneration contexts. Our results will greatly promote research about liver regeneration mechanisms related to miRNA.

Liver regeneration involves regulation of cell proliferation, including such processes as G0/G1 and G1/S transitions of the cell cycle and cell division. Differential expression levels of three miRNAs related to cell proliferation and apoptosis were tested in regenerating liver tissues at 0 h, 6 h, 12 h, and 24 h through the real-time PCR. We determined that three target miRNAs (hsa-miR-142-3p_R-1, xtr-miR-125b, and fru-miR-204) exhibited great differential transcription in regenerating liver tissues at 0 h, 6 h, 12 h, and 24 h with higher expression in the early timepoints of liver regeneration. Expression of miRNAs showed a rising trend in regenerating liver tissues at 6 h, 12 h but in the later timepoints presented a downward trend, with normal levels in regenerating liver tissues at 24 h. It has been reported that apoptotic activity increases significantly in the early timepoints (0–6 h) and late timepoints (4–7 d) of liver regeneration. Therefore, we can speculate that the expression of these three miRNAs increased significantly in the start-up stage of liver regeneration. These miRNAs may be relevant to the inhibition of cell proliferation and DNA synthesis. After 24 h, expression decreased to normal. This may be relevant to the S stage of DNA synthesis, which peaks at this time. In addition, it has been reported that the expression of miRNA showed a rising trend in regenerating liver tissues of ratsat 3 h, 12 h, and 18 h after two-thirds PH, and expression of 70% of miRNAs showed a downward trend in regenerating rat livers 24 h after two-thirds PH. Expression of genes relevant to miRNA generation began to decrease 18–36 h after PH, most obviously 24 h after PH, including RNasen, Dgcr8, Dicer, Tarbp2, and ect [41]. Therefore, we can speculate that this upregulation followed by downregulation of expression may be induced by a negative feedback mechanism involving miRNA generation. It has been reported that miR-142-3p controls the target gene RAC1 and inhibits liver cancer cell proliferation and invasion [42], that miR-204a negatively controls the target gene Bcl-2 to promote apoptosis [43], and that miR-125b was upregulated in Taxol-resistant cells, causing a marked inhibition of Taxol-induced cytotoxicity and apoptosis and a subsequent increase in the resistance to Taxol in cancer cells [44]. Therefore, combining the reported references on miRNAs in liver regeneration and the related signal transduction network references may be helpful to screen and test the liver regeneration miRNAs and their target genes. Researching mechanisms such as when and how the interactions between miRNAs and target genes occur will deepen the understanding of miRNA regulation in the injured liver.

5. Conclusions

This study provides the first large-scale identification and characterization of *Chiloscyllium plagiosum* miRNAs, adds a significant number of novel miRNA sequences to the currently available database, and lays the foundation for further understanding of miRNA function in the regulation of *Chiloscyllium plagiosum* liver development. However, considerable work remains to confirm the identity of these miRNAs and their functional significance. Despite the conservation of miRNAs, it is likely that we overlooked several *Chiloscyllium plagiosum* specific miRNAs due to the unavailability of genome sequences for this species. Moreover, the expressed population of miRNAs can change in different tissues and at different development stages. The identified miRNAs may not represent all of the miRNAs that exist in *Chiloscyllium plagiosum*, and more research is required to acquire a full set of miRNAs for this species.

Acknowledgments

The authors thank LC-Bio (Hangzhou, China) for performing the Illumina sequencing. The authors also thank Qingtuo Guo and Lufeng Li for their assistance during the analysis and interpretation of data. They also thank Elsevier Language Editing for editing the language of this paper. This work was supported by financial grants from the National High Technology Research and Development Program (no. 2012ZX09102301-009), the National High Technology Research and Development Program (no. 2011AA100603), National Natural Science Foundation of China (No. 11105121), National Basic Research Program of China (no. 2012CB114600), Science Technology Department of Zhejiang Province (no. 2012C 22053), Zhejiang Natural Science Foundation (no. Y3110051), Science Technology Department of Zhejiang Province Project (no. 2012C22053), Financial grants from Zhejiang Sci-Tech University (no. 1016848-Y).

References

[1] S. Kumar and S. B. Hedges, "A molecular timescale for vertebrate evolution," *Nature*, vol. 392, no. 6679, pp. 917–920, 1998.

[2] B. Venkatesh, E. F. Kirkness, Y.-H. Loh et al., "Survey sequencing and comparative analysis of the elephant shark (callorhinchus milii) genome," *PLoS Biology*, vol. 5, no. 4, article e101, 2007.

[3] F. J. Huang and W. T. Wu, "Purification and characterization of a new peptide (S-8300) from shark liver," *Journal of Food Biochemistry*, vol. 34, no. 5, pp. 962–970, 2010.

[4] F. Zhou, Y. Wang, Y. Guan et al., "Construction and characterization of a cDNA library from shark regenerated hepatic tissue," *Fish and Shellfish Immunology*, vol. 30, no. 4-5, pp. 1170–1177, 2011.

[5] X. Zongfa, C. Guangming, H. Ying et al., "Study on immune activity of shark liver extracts," *Chinese Journal of Biochemical Pharmaceutics*, vol. 3, 1999.

[6] Q. Ma, Y. Q. Su, J. Wang, Z. M. Zhuang, and Q. S. Tang, "Molecular cloning and expression analysis of major histocompatibility complex class IIB gene of the Whitespotted bambooshark (*Chiloscyllium plagiosum*)," *Fish Physiology and Biochemistry*, vol. 39, no. 2, pp. 131–142, 2012.

[7] A. Grishok, A. E. Pasquinelli, D. Conte et al., "Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing," *Cell*, vol. 106, no. 1, pp. 23–34, 2001.

[8] G. Hutvágner and P. D. Zamore, "A microRNA in a multiple-turnover RNAi enzyme complex," *Science*, vol. 297, no. 5589, pp. 2056–2060, 2002.

[9] V. N. Kim, J. Han, and M. C. Siomi, "Biogenesis of small RNAs in animals," *Nature Reviews Molecular Cell Biology*, vol. 10, no. 2, pp. 126–139, 2009.

[10] Y. Miyazaki, H. Adachi, M. Katsuno et al., "Viral delivery of miR-196a ameliorates the SBMA phenotype via the silencing of CELF2," *Nature Medicine*, vol. 18, no. 7, pp. 1136–1141, 2012.

[11] Z. G. Cai, S. M. Zhang, Y. Zhang, Y.-Y. Zhou, H. B. Wu, and X. P. Xu, "MicroRNAs are dynamically regulated and play an important role in LPS-induced lung injury," *Canadian Journal of Physiology and Pharmacology*, vol. 90, no. 1, pp. 37–43, 2012.

[12] S. Subramanian and C. J. Steer, "MicroRNAs as gatekeepers of apoptosis," *Journal of Cellular Physiology*, vol. 223, no. 2, pp. 289–298, 2010.

[13] S. Bala and G. Szabo, "MicroRNA signature in alcoholic liver disease," *International Journal of Hepatology*, vol. 2012, Article ID 498232, 6 pages, 2012.

[14] X. Chen, M. Murad, Y. Y. Cui et al., "MiRNA regulation of liver growth after 50% partial hepatectomy and small size grafts in rats," *Transplantation*, vol. 91, no. 3, pp. 293–299, 2011.

[15] S. K. Venugopal, J. Jiang, T.-H. Kim et al., "Liver fibrosis causes downregulation of miRNA-150 and miRNA-194 in hepatic stellate cells, and their overexpression causes decreased stellate cell activation," *American Journal of Physiology*, vol. 298, no. 1, pp. G101–G106, 2010.

[16] B. Venkatesh, A. Tay, N. Dandona, J. G. Patil, and S. Brenner, "A compact cartilaginous fish model genome," *Current Biology*, vol. 15, no. 3, pp. R82–R83, 2005.

[17] M. Li, Y. Xia, Y. Gu et al., "MicroRNAome of porcine pre- and postnatal development," *PLoS ONE*, vol. 5, no. 7, article e11541, 2010.

[18] Z. Wei, X. Liu, T. Feng, and Y. Chang, "Novel and conserved micrornas in dalian purple urchin (*Strongylocentrotus nudus*) identified by next generation sequencing," *International Journal of Biological Sciences*, vol. 7, no. 2, pp. 180–192, 2011.

[19] S. Ambady, Z. Wu, and T. Dominko, "Identification of novel micrornas in *Xenopus laevis* metaphase ii arrested eggs," *Genesis*, vol. 50, no. 3, pp. 286–299, 2012.

[20] B. H. Zhang, X. P. Pan, S. B. Cox, G. P. Cobb, and T. A. Anderson, "Evidence that miRNAs are different from other RNAs," *Cellular and Molecular Life Sciences*, vol. 63, no. 2, pp. 246–254, 2006.

[21] B. Niu, L. Fu, S. Sun, and W. Li, "Artificial and natural duplicates in pyrosequencing reads of metagenomic data," *BMC Bioinformatics*, vol. 11, article 187, 2010.

[22] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.

[23] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks, "MicroRNA targets in *Drosophila*," *Genome Biology*, vol. 5, no. 1, 1 page, 2003.

[24] C. Z. Zhao, H. Xia, T. P. Frazier et al., "Deep sequencing identifies novel and conserved microRNAs in peanuts (*Arachis hypogaea L.*)," *BMC Plant Biology*, vol. 10, article 3, 2010.

[25] T. H. Beilharz, D. T. Humphreys, J. L. Clancy et al., "MicroRNA-mediated messenger RNA deadenylation contributes to translational repression in mammalian cells," *PLoS ONE*, vol. 4, no. 8, article e6783, 2009.

[26] P. Sethupathy, M. Megraw, and A. G. Hatzigeorgiou, "A guide through present computational approaches for the identification of mammalian microRNA targets," *Nature Methods*, vol. 3, no. 11, pp. 881–886, 2006.

[27] Y. Cai, X. Yu, Q. Zhou et al., "Novel microRNAs in silkworm (*Bombyx mori*)," *Functional and Integrative Genomics*, vol. 10, no. 3, pp. 405–415, 2010.

[28] H. Chen, Y. Sun, R. Dong et al., "Mir-34a is upregulated during liver regeneration in rats and is associated with the suppression of hepatocyte proliferation," *PLoS ONE*, vol. 6, no. 5, article e20238, 2011.

[29] R. P. Dippold, R. Vadigepalli, G. E. Gonye, B. Patra, and J. B. Hoek, "Chronic ethanol feeding alters miRNA expression dynamics during liver regeneration," *Alcoholism, Clinical and Experimental Research*, vol. 37, supplement 1, pp. E59–E69, 2013.

[30] M. H. Lu, C.-Z. Li, C. J. Hu et al., "MicroRNA-27b suppresses mouse MSC migration to the liver by targeting SDF-1α in vitro," *Biochemical and Biophysical Research Communications*, vol. 421, no. 2, pp. 389–395, 2012.

[31] F. Meng, H. Francis, S. Glaser et al., "Role of stem cell factor and granulocyte colony-stimulating factor in remodeling during liver regeneration," *Hepatology*, vol. 55, no. 1, pp. 209–221, 2012.

[32] B. Yuan, R. Dong, D. Shi et al., "Down-regulation of miR-23b may contribute to activation of the TGF-β1/Smad3 signalling pathway during the termination stage of liver regeneration," *FEBS Letters*, vol. 585, no. 6, pp. 927–934, 2011.

[33] I. Chaveles, A. Zaravinos, I. G. Habeos et al., "MicroRNA profiling in murine liver after partial hepatectomy," *International Journal of Molecular Medicine*, vol. 29, no. 5, pp. 747–755, 2012.

[34] Y. Zhu, Y. Lu, Q. Zhang et al., "MicroRNA-26a/b and their host genes cooperate to inhibit the G1/S transition by activating the pRb protein," *Nucleic Acids Research*, vol. 40, no. 10, pp. 4615–4625, 2012.

[35] B. Zhang, X. Pan, Q. Wang, G. P. Cobb, and T. A. Anderson, "Computational identification of microRNAs and their targets," *Computational Biology and Chemistry*, vol. 30, no. 6, pp. 395–407, 2006.

[36] C. Wang, J. Han, C. Liu et al., "Identification of microRNAs from Amur grapes (*vitis amurensis Rupr.*) by deep sequencing and analysis of microRNA variations with bioinformatics," *BMC Genomics*, vol. 13, article 122, 2012.

[37] D. Schotte, F. A. Moqadam, E. A. M. Lange-Turenhout et al., "Discovery of new microRNAs by small RNAome deep sequencing in childhood acute lymphoblastic leukemia," *Leukemia*, vol. 25, no. 9, pp. 1389–1399, 2011.

[38] B. Venkatesh, E. F. Kirkness, Y. H. Loh et al., "Ancient noncoding elements conserved in the human genome," *Science*, vol. 314, no. 5807, p. 1892, 2006.

[39] G. Song and L. Wang, "MiR-433 and miR-127 arise from independent overlapping primary transcripts encoded by the miR-433-127 locus," *PLoS ONE*, vol. 3, no. 10, article e3574, 2008.

[40] A. Stark, N. Bushati, C. H. Jan et al., "A single Hox locus in *Drosophila* produces functional microRNAs from opposite DNA strands," *Genes and Development*, vol. 22, no. 1, pp. 8–13, 2008.

[41] J. Shu, B. T. Kren, Z. Xia et al., "Genomewide microRNA down-regulation as a negative feedback mechanism in the early phases of liver regeneration," *Hepatology*, vol. 54, no. 2, pp. 609–619, 2011.

[42] L. Wu, C. Cai, X. Wang, M. Liu, X. Li, and H. Tang, "MicroRNA-142-3p, a new regulator of RAC1, suppresses the migration and invasion of hepatocellular carcinoma cells," *FEBS Letters*, vol. 585, no. 9, pp. 1322–1330, 2011.

[43] A. Sacconi, F. Biagioni, V. Canu et al., "miR-204 targets Bcl-2 expression and enhances responsiveness of gastric cancer," *Cell Death & Disease*, vol. 3, article e423, 2012.

[44] M. Zhou, Z. Liu, Y. Zhao et al., "MicroRNA-125b confers the resistance of breast cancer cells to paclitaxel through suppression of pro-apoptotic Bcl-2 antagonist killer 1 (Bak1) expression," *The Journal of Biological Chemistry*, vol. 285, no. 28, pp. 21496–21507, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 485034, 8 pages
http://dx.doi.org/10.1155/2013/485034

Research Article

Predicting Drugs Side Effects Based on Chemical-Chemical Interactions and Protein-Chemical Interactions

Lei Chen,[1] Tao Huang,[2,3,4] Jian Zhang,[5] Ming-Yue Zheng,[6] Kai-Yan Feng,[7] Yu-Dong Cai,[8,9] and Kuo-Chen Chou[9,10]

[1] *College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*
[2] *Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[3] *Shanghai Center for Bioinformation Technology, Shanghai 200235, China*
[4] *Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York City, NY 10029, USA*
[5] *Department of Ophthalmology, Shanghai First People's Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200080, China*
[6] *State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Shanghai 201203, China*
[7] *Beijing Genomics Institute, Shenzhen Beishan Industrial Zone, Shenzhen 518083, China*
[8] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[9] *Gordon Life Science Institute, Belmont, Massachusetts 02478, USA*
[10] *Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia*

Correspondence should be addressed to Yu-Dong Cai; cai_yud@126.com and Kuo-Chen Chou; kcchou@gordonlifescience.org

Received 18 June 2013; Accepted 30 July 2013

Academic Editor: Bing Niu

A drug side effect is an undesirable effect which occurs in addition to the intended therapeutic effect of the drug. The unexpected side effects that many patients suffer from are the major causes of large-scale drug withdrawal. To address the problem, it is highly demanded by pharmaceutical industries to develop computational methods for predicting the side effects of drugs. In this study, a novel computational method was developed to predict the side effects of drug compounds by hybridizing the chemical-chemical and protein-chemical interactions. Compared to most of the previous works, our method can rank the potential side effects for any query drug according to their predicted level of risk. A training dataset and test datasets were constructed from the benchmark dataset that contains 835 drug compounds to evaluate the method. By a jackknife test on the training dataset, the 1st order prediction accuracy was 86.30%, while it was 89.16% on the test dataset. It is expected that the new method may become a useful tool for drug design, and that the findings obtained by hybridizing various interactions in a network system may provide useful insights for conducting in-depth pharmacological research as well, particularly at the level of systems biomedicine.

1. Introduction

Many drugs approved by Food and Drug Administration (FDA) were recalled each year after some unexpected side effects were discovered; for example, in 2010, Reductil/Meridia, Mylotarg, and Avandia were withdrawn. According to the "Drug Recall" (http://www.drugrecalls.com/drugrecalls.html), about 20 million people had taken the drugs in 1997 and 1998 that were later withdrawn. The drug side effects may have seriously harmful consequences to human beings [1]. For instance, the antiobesity drug fenfluramine/phentermine, also known as fen-phen, may cause heart disease and hypertension. Developing and producing drugs that were later found having serious side effects would be a disaster to a pharmaceutical company. For instance, the withdrawal of the aforementioned antiobesity drug has cost Wyeth more than $21 billion in America alone [2]. Therefore, it will not only avoid causing harm to patients but also avoid wasting lots of money if we can discover the side effects of a drug compound in the early phase of drug discovery.

Many efforts have been made in this regard, such as utilizing the drug perturbed gene expression profiles or biological pathways, to predict the side effects of drugs [1, 3–7],

using chemical structures for the prediction of drugs side effects [8–10]. Although, most of the methods can only provide whether the query drug has some side effects, they cannot determine which side effects are most likely to happen or even the order information of the side effects. In this study, we proposed a novel computational method to predict the side effects of drugs based on chemical-chemical interaction and protein-chemical interaction. Compared to most of the previous studies, our method can provide the order information of the side effects, that is, prioritizing the side effects from the most likely one to the least likely one.

During the past decade, many compound databases have been constructed, such as KEGG (Kyoto Encyclopedia of Genes and Genomes) [11] and STITCH (Search Tool for Interactions of Chemicals) [12]. KEGG provides the information of chemical substances and reactions, while STITCH provides the interaction information of chemicals and proteins. Thus we can acquire the properties of many compounds and their other information from these databases. For those compounds not being covered by these databases, their properties can be inferred from the property-known compounds stored in the databases [13–16]. Likewise, the drugs side effects can also be inferred as elaborated below.

Recently, it was evidenced that interactive proteins are more likely to share common biological functions [17–20], and that interactive compounds are also more likely to share common biological functions [13, 16]. Since the side effects are part of biological functions of drugs, it would be feasible to use the chemical-chemical interactions to identify the drugs side effects. Unfortunately, some of the query drugs cannot be predicted for their side effects by this way because their interactive counterparts do not have any information of the side effects. To overcome such difficulty, we proposed to utilize the information of indirect interactions, including both the chemical-chemical interaction and the protein-chemical interaction, to identify the drugs side effects of which the direct chemical-chemical interaction data are not available. To evaluate the method, a benchmark dataset retrieved from SIDER [21] was constructed, which consisted of 835 drug compounds, and it was divided into one training dataset and one test dataset. By a jackknife test on the training dataset, the 1st order prediction accuracy was 86.30%, while it was 89.16% on the test dataset. To confirm the effectiveness of the method, another method based on chemical structure similarity obtained by SMILES string [22] was also conducted on the training and test datasets. Encouraged by the good performance of the method and superiority to the method based on chemical structure similarity, we hope that the proposed method can become a useful tool to predict drugs side effects and screen out drugs with undesired side effects.

2. Materials and Methods

2.1. Benchmark Dataset. The benchmark dataset used in the current study was downloaded from SIDER [21] at http://sideeffects.embl.de/, which integrated the side effects of 888 drugs from the US Food and Drug Administration (FDA) and other sources [21]. To obtain a high-quality, well-defined benchmark dataset, the data were collected strictly according to the following criteria: (i) only the 100 side effects with most drugs listed in SIDER and the corresponding drugs were included, and (ii) drugs without both chemical-chemical interactions and protein-chemical interactions were also excluded. Finally, we obtained a benchmark dataset **S** that contained 835 drugs belonging to 100 categories of side effects. The codes of the 835 drugs in each of the 100 side effect categories are given in Supplementary Material I available online at http://dx.doi.org/10.1155/2013/485034.

For the convenience of later formulation, let us use the symbols $C_1, C_2, C_3, \ldots, C_{100}$ to tag the 100 side effects, where C_1 represents "Nausea," C_2 "Headache," C_3 "Vomiting," and so forth, as described in the table in Supplementary Material II, in which the number of drugs with each of the 100 side effect tags is also given. Thus, the benchmark dataset **S** can be formulated as

$$\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_2 \cup \cdots \cup \mathbf{S}_{100}, \tag{1}$$

where $\mathbf{S}_i$ represents the subset that contains the drugs with the side effect C_i ($i = 1, 2, \ldots, 100$).

Since many drugs in **S** have multiple side effects that is, they may simultaneously occur in subsets with different side effect tags, it is instructive to introduce the concept of "virtual drug" sample, as illustrated as follows. A drug compound coexisting at two different side effect subsets will be counted as 2 virtual drugs even though they have an identical chemical structure, if coexisting at three different subsets, 3 virtual drugs; and so forth. Accordingly, the total structure-different drug compounds and the total number of the side-effect-different virtual drug compounds can be described by the following equation:

$$\begin{aligned} N(\text{str}) &= 835, \\ N(\text{vir}) &= \sum_{i=1}^{100} N(C_i) = 30{,}114, \end{aligned} \tag{2}$$

where N(str) is the number of the total structure-different drug compounds, N(vir) the number of the total side-effect-different virtual drug compounds in **S**, and $N(C_i)$ the number of drugs with the side-effect tag C_i. Substituting the numbers of $N(C_i)$ ($i = 1, 2, \ldots, 100$) in the table in Supplementary Material II into (2), we obtained N(vir) = 30,114 fully consistent with the results in (2) and the table in Supplementary Material II.

It can be seen from (2) that the total number of the side-effect-different virtual drug compounds is much greater than that of the total structure-different drug compounds. To provide an intuitive view about their distribution, a histogram of the number of drugs versus the number of side effects is given in Figure 1, from which we can see that, of the 835 drugs, only 6 have one side effect while the majority has more than 10 side effects. Thus, the prediction of drugs side effects is a multilabel classification problem. Like the case in dealing with compounds with multiple properties [13, 16], the proposed method would provide the order information of side effects from the most likely to the least likely.

To evaluate the methods as described below sufficiently, we randomly selected 10% (83) samples from **S** to compose

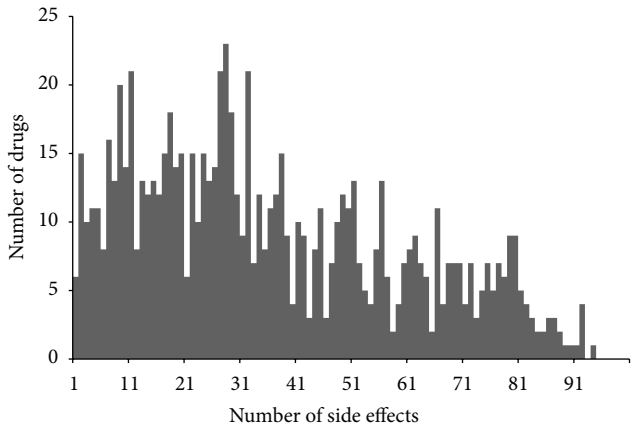


Figure 1: A histogram of the number of drugs versus the number of side effects.

the test dataset, denoted by $\mathbf{S}_{te}$, while the remaining 752 samples in $\mathbf{S}$ were used to construct the training dataset, denoted by $\mathbf{S}_{tr}$.

2.2. Chemical-Chemical Interactions and Protein-Chemical Interactions. It was evidenced that interactive proteins are more likely to share common biological functions than noninteractive ones [17–20]. Likewise, it has been indicated by some pioneer studies [13, 16] that interactive compounds follow the similar rules. Since side effect is one part of biological functions of drugs, using the properties of interactive compounds to identify drugs side effects is a feasible scheme.

To obtain the information of interactive compounds, we downloaded the data of chemical-chemical interactions from STITCH (http://stitch.embl.de/, chemical_chemical.links.detailed.v3.0.tsv.gz) [12], a well-known database containing known and predicted chemical-chemical interaction and protein-chemical interaction data from experiments, literature, or other reliable sources. In the datafile obtained each interaction unit contains two chemicals and five scores with titles "Similarity," "Experimental," "Database," "Textmining," and "Combined_score," respectively. Since the last score combines the information of other scores, we utilized the last score to indicate the interactivity of two chemicals in this study; that is, compounds in the interaction unit with "Combined_score" greater than zero were deemed to be interactive compounds. The interactive compounds thus considered here satisfy one of the following three properties: (I) they participate in the same reactions; (II) they share similar structures or activities; (III) they have literature associations. These three properties always indicate that the interactive compounds occupy the same biological pathways, suggesting they may induce similar side effects. It is confirmed that using chemical-chemical interactions retrieved from STITCH to identify drugs side effects is feasible. The "Combined_score" is termed as confidence score, because its value always indicates the likelihood that two interactive compounds can interact in a way that two compounds with high "Combined_score" mean that they can interact with high probability. For any two-drug compounds d_1 and d_2, their interaction confidence score was denoted by $Q^c(d_1, d_2)$. Particularly, if the interaction between d_1 and d_2 did not exist, their interaction confidence score was set to zero; that is, $Q^c(d_1, d_2) = 0$.

Since the data of chemical-chemical interactions in STITCH is not very complete at present; that is, some potential chemical-chemical interactions may not be reported in STITCH, predicted methods based on chemical-chemical interactions may have a limitation that samples without interactive counterparts in the training dataset cannot be processed. Thus, it is necessary to give some new schemes to measure the interactions that are not reported in STITCH. It is known that if two drug compounds can interact with a third compound or protein, these two drug compounds are likely to share some common functions. In view of this, we proposed a new scheme to measure the likelihood of interaction of two chemicals based on indirect chemical-chemical and protein-chemical interactions.

The data for the protein-chemical interactions were also downloaded from STITCH (http://stitch.embl.de/, protein_chemical.links.detailed.v3.0.tsv.gz). Each of the interaction units in the datafile obtained contains one compound, one protein, and four scores with titles "Experimental," "Database," "Textmining," and "Combined_score," respectively. With the similar argument, we used the value of "Combined_score," also termed as confidence score, to indicate the likelihood of the interaction's occurrence. For one protein p and one drug compound d, their interaction confidence score was denoted as $Q^p(p, d)$. If there was no interaction at all between the protein p and the drug d, it was also set to zero; that is, $Q^p(p, d) = 0$.

Now, we are ready to introduce the new scheme to measure the likelihood of interaction of two chemicals. For two compounds d_1 and d_2, suppose $I^c(d_1)$ denote a set of compounds that are directly interacting with the drug d_1 and $I^c(d_2)$ a set of compounds directly interacting with the drug d_2, formulated as

$$I^c(d_1) = \{d : Q^c(d, d_1) > 0\},$$
$$I^c(d_2) = \{d : Q^c(d, d_2) > 0\}. \quad (3)$$

In a similar way let $I^p(d_1)$ denote a set of proteins that are directly interacting with the drug d_1 and $I^p(d_2)$ a set of proteins directly interacting with the drug d_2, formulated as

$$I^p(d_1) = \{p : Q^p(p, d_1) > 0\},$$
$$I^p(d_2) = \{p : Q^p(p, d_2) > 0\}. \quad (4)$$

According to the set theory, the drug compounds that are interacting with both the drug d_1 and the drug d_2 should be the intersection of the set $I^c(d_1)$ and the set $I^c(d_2)$; that is, they will form a set given by

$$I^c(d_1, d_2) = I^c(d_1) \cap I^c(d_2). \quad (5)$$

Likewise, the human proteins that are interacting with both the drug d_1 and the drug d_2 should be the intersection of the

set $I^p(d_1)$ and the set $I^p(d_2)$; that is, they will form a set given by

$$I^p(d_1, d_2) = I^p(d_1) \cap I^p(d_2). \quad (6)$$

Thus, the likelihood of the interaction between d_1 and d_2 can be calculated via the following equation:

$$Q^h(d_1, d_2) = \left(\sum_{d' \in I^c(d_1, d_2)} \left(Q^c(d_1, d') + Q^c(d_2, d') \right) + \sum_{p' \in I^p(d_1, d_2)} \left(Q^p(p', d_1) + Q^p(p', d_2) \right) \right) \times \left(2 \left| I^c(d_1, d_2) \cup I^p(d_1, d_2) \right| \right)^{-1}, \quad (7)$$

where $\in$ is a symbol in the set theory meaning "member of."

2.3. Interaction-Based Method. It is instructive to recall that by using the information of protein-protein interactions, some methods have been developed to successfully predict the properties of proteins [17–20, 23]. Actually, the underlying idea of these methods was based on the assumption that interactive proteins are more likely to share common biological functions than noninteractive ones. Similarly, based on the argument in Section 2.2 and some previous studies [13, 16], interactive drugs are more likely to share similar side effects than noninteractive ones. Based on such an underlying idea, the following predicted method based on chemical-chemical and protein-chemical interactions was developed.

For convenience, some notations are necessary. Suppose there are n drugs in the training set $\mathbf{S}'$, say $d_1, d_2, \ldots, d_n$; the side effects of the drug d_i in the training dataset is described as

$$C(d_i) = [c_{i,1}, c_{i,2}, \ldots, c_{i,100}]^{\mathbf{T}} \quad (i = 1, 2, \ldots, n), \quad (8)$$

where $\mathbf{T}$ is the transpose operator and

$$c_{i,j} = \begin{cases} 1, & \text{if } d_i \text{ has the side effect } C_j, \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

Prediction Based on Chemical-Chemical Interactions. As elaborated previously, interactive drugs always share similar side effects. The likelihood that the query drug d has the side effect C_j can be calculated by

$$\prod^c (d \longrightarrow C_j) = \sum_{d_i \in \mathbf{S}'} Q^c(d, d_i) \cdot c_{i,j} \quad j = 1, 2, \ldots, 100. \quad (10)$$

According to (10), the greater score $\prod^c(d \rightarrow C_j)$ means that there are lots of interactive compounds of d that have the side effect C_j or some interactions between d and its interactive compounds with the side effect C_j are labeled by high confidence scores. Thus, the greater the score $\prod^c(d \rightarrow C_j)$ is, the more likely the drug compound d has the jth side effect, with $\prod^c(d \rightarrow C_j) = 0$ indicating that the probability for the drug d having the jth side effect is zero. Since a drug usually has multiple side effects (see Figure 1), the prediction should provide a series of candidate side effects ranging from the most likely one to the least likely one, rather than only giving the most likely one. Thus, for a query drug d, suppose we have

$$\prod^c (d \longrightarrow C_2) \geq \prod^c (d \longrightarrow C_4) \geq \cdots \geq \prod^c (d \longrightarrow C_{90}) > 0 \quad (11)$$

meaning that the highest likelihood of side effect for the drug d is C_2 or "Headache" (cf. table in Supplementary Material II), and the second highest is C_4 or "Rash", and so forth. In other words, C_2 is called the 1st order prediction, C_4 the 2nd order prediction, and so forth. Note that the outcome of (10) might be trivial; that is,

$$\prod^c \left(d \longrightarrow C_j \right) = 0 \quad \text{for } j = 1, 2, \ldots, 100 \quad (12)$$

implying that no meaningful or direct interactive drug compounds whatsoever can be found in the training dataset $\mathbf{S}'$ for the drug d. Under such a circumstance, an alternative approach should be used for predicting its side effects, as elaborated below.

Prediction Based on Hybrid Interactions. When the query drug d did not have any directly interactive drugs in the training dataset $\mathbf{S}'$ or the information of its directly interactive drugs was trivial, the data for the indirect chemical-chemical and protein-chemical interactions would be used to predict its side effects. The prediction method was formulated in a similar way as the above method. But now instead of (10), the likelihood that the query drug d has the side effect C_j should be calculated by

$$\prod^h \left(d \longrightarrow C_j \right) = \sum_{d_i \in \mathbf{S}'} Q^h(d, d_i) \cdot c_{i,j}. \quad (13)$$

By integrating the above two different approaches, the following steps were adopted to predict the side effects of the query drug d.

Step 1. The method based on the chemical-chemical interactions; that is, (10), was first utilized to identify its side effects.

Step 2. If the outcomes were trivial or no meaningful results were obtained as in the case of (12), the method based on the hybrid interactions, that is, (13), would be utilized to continue the prediction.

2.4. Similarity-Based Method. It is known that the compounds with similar structural properties always involve in similar biological activities [24]. The most well-known representing system to obtain the similarity information of two compounds is SMILES (Simplified Molecular Input Line Entry System) [22], which is a line notation for representing molecules and reactions using ASCII strings. Here, we also

used this system to obtain the representations of compounds, which were used to calculate the similarity score of two compounds and set up a new computational method to identify drugs side effect. The similarity score between two compounds with their SMILES representations can be obtained from Open Babel [25] that is an open chemical toolbox. For two-drug compounds d_1 and d_2, their similarity score obtained from Open Babel was denoted by $Q^s(d_1, d_2)$. Based on the fact that the compounds with similar structural properties always share the same biological activities, the likelihood that the query drug d has the side effect C_j can be calculated by

$$\prod^s \left(d \longrightarrow C_j\right) = \max_{d_i \in \mathbf{S}'} Q^s \left(d, d_i\right) \cdot c_{i,j} \tag{14}$$

meaning that the likelihood that the query drug d has the side effect C_j is formulated as the maximum similarity scores between d and those drugs with side effect C_j in the training dataset $\mathbf{S}'$. Obviously, the greater the score $\prod^s(d \rightarrow C_j)$, the more likely the drug compound d has the side effect C_j. Following the similar procedure of the method based on chemical-chemical interactions, we can also obtain the order information of the query drug d in terms of $\prod^s(d \rightarrow C_j)$ $(j = 1, 2, \ldots, 100)$.

2.5. Jackknife Test. In statistical prediction, Jackknife test [16] is often used to examine a predictor for its effectiveness in practical application. In the jackknife test, all the samples in the dataset will be singled out one-by-one and tested by the predictor trained by the remaining samples. During the process of jackknifing, both the training dataset and testing dataset are actually open, and each sample will be in turn moved between the two. The jackknife test can exclude the "memory" effect, and the arbitrariness problem can also be avoided. Thus, the outcome obtained by the jackknife test is always unique for a given benchmark dataset [26]. Accordingly, the jackknife test has been widely recognized and increasingly adopted to investigate the performance of various predictors [27–36]. Thus, the jackknife test was also adopted here to evaluate the anticipated accuracy of the current predicted methods.

2.6. Accuracy Measurement. For a query drug, we may identify a series of side effects with the current prediction method. For the jth order prediction, its prediction accuracy AC(j) can be calculated by

$$\mathrm{AC}\left(j\right) = \frac{P\left(j\right)}{N} \quad j = 1, 2, \ldots, 100, \tag{15}$$

where $P(j)$ denotes the number of drugs whose jth order prediction is one of the true side effects and N denotes the total number of structure-different drugs in the dataset. According to the prediction method with 100 orders of prediction results, high AC(j) with small j and low AC(j) with large j would indicate a good prediction [13, 16, 20]. Generally speaking, it also implies a good performance by the predictor if its 1st order prediction has a high success rate.

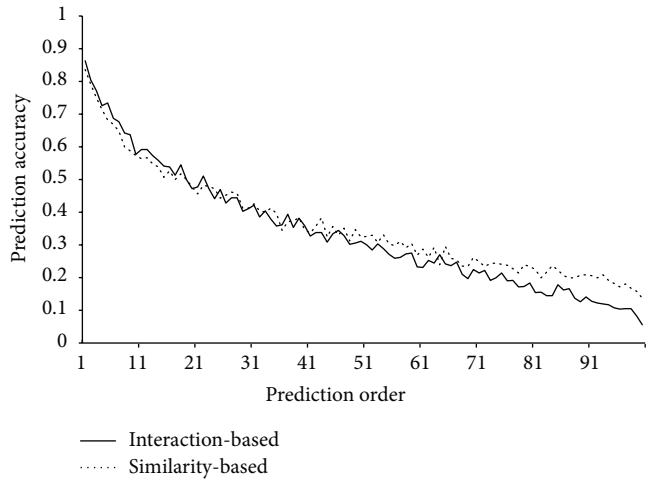


Figure 2: A plot of the prediction accuracy of two methods on the training dataset versus the order of prediction.

3. Results and Discussion

Of the 835 drugs in the benchmark dataset $\mathbf{S}$, 83 samples were randomly selected to compose test dataset $\mathbf{S}_{te}$, while the rest 752 samples composed the training dataset $\mathbf{S}_{tr}$. The predicted results of the interaction-based method and similarity-based method on the training and test datasets are as follows.

3.1. Performance of the Methods on the Training Dataset. For 752 drug compounds in the training dataset $\mathbf{S}_{tr}$, the interaction-based method and similarity-based method were conducted to make prediction with their performance evaluated by jackknife test. Listed in columns 2 and 4 of Table 1 are the first 20 prediction accuracies obtained by these two methods, from which we can see that the 1st order prediction accuracies of the interaction-based and similarity-based method were 86.30% and 83.64%, respectively, while the 2nd ones were 80.45% and 79.12%, respectively. The total 100 prediction accuracies obtained by these two methods are given in Supplementary Material III, and two curves with prediction accuracies as their Y-axis and prediction order as their X-axis are shown in Figure 2. It is observed that the prediction accuracies obtained by the interaction-based method descend generally with the increase of the order number, and the same situation also occurred for the prediction accuracies obtained by the similarity-based method. All of these imply that the two methods sorted the side effects of drug compounds in the training dataset quite well, and they are all quite effective in identifying drugs side effects.

3.2. Performance of the Methods on the Test Dataset. For the 83 drug compounds in the test dataset $\mathbf{S}_{te}$, the side effects of these samples were predicted by the interaction-based and similarity-based method based on the drug compounds in the training dataset $\mathbf{S}_{tr}$. After processing by (15), 100 prediction accuracies obtained by each method were obtained and were also given in Supplementary Material III. Listed in columns 3

Table 1: The first 20 prediction accuracies of the interaction-based and similarity-based methods in identifying the side effects of drugs in the training and test datasets.

Prediction order	Interaction-based		Similarity-based		Difference	
	Training dataset	Test dataset	Training dataset	Test dataset	Training dataset[a]	Test dataset[b]
1	86.30%	89.16%	83.64%	87.95%	2.66%	1.20%
2	80.45%	83.13%	79.12%	83.13%	1.33%	0.00%
3	77.13%	84.34%	75.00%	79.52%	2.13%	4.82%
4	72.61%	81.93%	71.41%	75.90%	1.20%	6.02%
5	73.40%	77.11%	68.22%	74.70%	5.19%	2.41%
6	68.75%	75.90%	66.89%	71.08%	1.86%	4.82%
7	67.69%	67.47%	64.76%	57.83%	2.93%	9.64%
8	64.23%	65.06%	59.97%	65.06%	4.26%	0.00%
9	63.70%	68.67%	58.78%	57.83%	4.92%	10.84%
10	57.71%	57.83%	57.31%	60.24%	0.40%	−2.41%
11	59.18%	60.24%	56.38%	67.47%	2.79%	−7.23%
12	59.18%	69.88%	56.65%	51.81%	2.53%	18.07%
13	57.31%	61.45%	54.79%	53.01%	2.53%	8.43%
14	55.85%	59.04%	53.86%	62.65%	1.99%	−3.61%
15	54.12%	54.22%	50.66%	57.83%	3.46%	−3.61%
16	53.86%	59.04%	52.66%	55.42%	1.20%	3.61%
17	51.33%	39.76%	50.00%	60.24%	1.33%	−20.48%
18	54.52%	62.65%	51.73%	53.01%	2.79%	9.64%
19	50.00%	56.63%	50.00%	38.55%	0.00%	18.07%
20	47.21%	44.58%	47.74%	51.81%	−0.53%	−7.23%

[a]Percentages in this column were calculated by percentages in column 2 minus percentages in column 4.
[b]Percentages in this column were calculated by percentages in column 3 minus percentages in column 5.

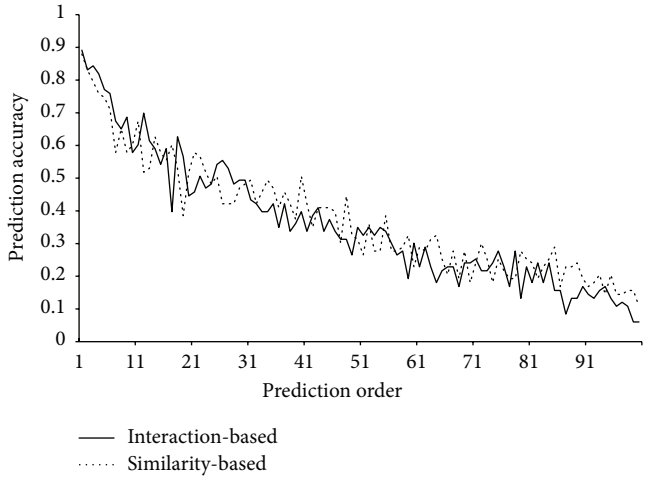


Figure 3: A plot of the prediction accuracy of two methods on the test dataset versus the order of prediction.

and 5 of Table 1 are the first 20 prediction accuracies obtained by these two methods, from which we can see that the 1st order prediction accuracies obtained by the interaction-based and similarity-based method were 89.16% and 87.95%, respectively. We also plotted two curves with prediction accuracies as their Y-axis and prediction order as their X-axis, which are shown in Figure 3. It is observed from Figure 3 that the accuracies also exhibit a trend of decrease with the increase of the order number. However, two curves in Figure 3 fluctuate more drastically and frequently than those of Figure 2, which may be caused by the low number of the samples in the test dataset. In any case, the interaction-based and similarity-based methods still sorted the side effects of samples in the test dataset reasonably well, implying again that these two methods are quite effective in identifying drugs side effects.

3.3. Comparison of the Interaction-Based and Similarity-Based Method. For 752 samples in the training dataset $\mathbf{S}_{\text{tr}}$ and 83 samples in the test dataset $\mathbf{S}_{\text{te}}$, the interaction-based and similarity-based methods were all used to identify their side effects. Listed in columns 6 and 7 of Table 1 are the differences of the first 20 prediction accuracies obtained by these two methods, from which we can see that the 1st order prediction accuracies obtained by the interaction-based method on the training and test datasets were 2.66% and 1.20% higher than those of similarity-based method. Furthermore, most prediction accuracies in Table 1 obtained by the interaction-based method are higher than the corresponding accuracies obtained by the similarity-based method, indicating that interaction-based method is more effective in identifying drugs side effects. It is also confirmed from Figures 2 and 3 that the curve obtained by the interaction-based method is always above the curve obtained by the similarity-based method when the prediction order is low. However, with

the increase of order number, the curve obtained by the similarity-based method keeps up with and exceeds the curve obtained by the interaction-based method, which may be caused by the following two reasons: (I) the high prediction accuracies, obtained by the interaction-based method, with low order number cause the low number of correctly predicted samples with high prediction order; (II) the system of using chemical similarity between two chemicals is more complete than that in STITCH at present, which leads to the fact that the similarity-based method can always identify more side effects than the interaction-based method. It is expected that the interaction-based method can be improved as more and more chemical-chemical and protein-chemical interactions become available in STITCH.

3.4. Discussion. It is a multitarget learning problem to predict the side effects of drugs, just like the case in dealing with a protein system with multiple subcellular location sites [37]. For each of the drugs investigated, we need to consider how many different side effects it may have and what are the probabilities these side effects may occur. To deal with this complicated statistical systems like that, we adopted the strategy of the multiple prediction orders, ranging from the most likely side effect prediction order to the least one, that is, giving the information to users, which side effect is most likely, which one is the second likely one, and so forth. Compared to most of the previous studies on the prediction of drugs side effects, our method can provide more information. The multiple prediction orders method can also be utilized to deal with other multi-target learning problems, such as subcellular location prediction [37] and functions of proteins [20].

In addition to the multi-target issue, we also faced the problem of coverage scope. Of the 835 drug compounds in the benchmark dataset, some of them have the information of chemical-chemical interaction, while for the rest such information is missing. To establish a predictor that can be used to predict the side effects of drugs under both the circumstances, the approach of the direct chemical-chemical interaction and the approach of the indirect chemical-chemical interaction were introduced. For the drug compounds belonging to the 1st circumstance, the predictions were conducted based on the direct chemical-chemical interactions (cf. (10)); for the rest drug compounds belonging to the 2nd circumstance, the predictions were conducted based on the hybrid interactions (cf. (13)). Thus, the side effects of all the 835 drugs could be predicted.

Finally, the good performance of the interaction-based method on the training and test datasets suggests that predictions based on the indirect interactions was also quite good, indicating that the entire interaction network—involving all the drug compounds and their direct or indirect interactions, as well as their interactions with human proteins—determines the side effects of drug compounds.

4. Conclusions

In this study, we proposed a novel prediction method to identify drugs side effects. For any query drug d, its side effects were determined by the following strategy: (1) if there exist interactive compounds of d in the training set, only chemical-chemical interactions were used to identify its side effects; (2) otherwise, both chemical-chemical interactions and protein-chemical interactions were employed to make prediction. Good performance of the method on the training and test datasets indicates that our method is quite effective in identifying drugs side effects. We hope that the method would assist in the prediction of drugs side effects during drug development and screening out drug candidates with undesired side effects.

Authors' Contribution

Lei Chen and Tao Huang contributed equally to this work.

Acknowledgments

This contribution is supported by National Basic Research Program of China (2011CB510101, 2011CB510102), National Natural Science Foundation of China (61202021, 61105097, 31371335), Innovation Program of Shanghai Municipal Education Commission (12YZ120, 12ZZ087), the grant of "The First-class Discipline of Universities in Shanghai," Shanghai Educational Development Foundation (12CG55), and Science and Technology Program of Shanghai Maritime University (no. 20120105).

References

[1] T. Huang, W. Cui, L. Hu, K. Feng, Y. Li, and Y. Cai, "Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles," *PLoS ONE*, vol. 4, no. 12, Article ID e8126, 2009.

[2] S. Saul, *Fen-Phen Case Lawyers Say They'll Reject Wyeth Offer*, New York Times, New York, NY, USA, 2005.

[3] P. E. Blower, C. Yang, M. A. Fligner et al., "Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data," *Pharmacogenomics Journal*, vol. 2, no. 4, pp. 259–271, 2002.

[4] K. J. Bussey, K. Chin, S. Lababidi et al., "Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel," *Molecular Cancer Therapeutics*, vol. 5, no. 4, pp. 853–867, 2006.

[5] R. H. Shoemaker, "The NCI60 human tumour cell line anticancer drug screen," *Nature Reviews Cancer*, vol. 6, no. 10, pp. 813–823, 2006.

[6] M. Fukuzaki, M. Seki, H. Kashima, and J. Sese, "Side effect prediction using cooperative pathways," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '09)*, pp. 142–147, November 2009.

[7] L. Xie, J. Li, L. Xie, and P. E. Bourne, "Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors," *PLoS Computational Biology*, vol. 5, no. 5, Article ID e1000387, 2009.

[8] J. Scheiber, J. L. Jenkins, S. C. K. Sukuru et al., "Mapping adverse drug reactions in chemical space," *Journal of Medicinal Chemistry*, vol. 52, no. 9, pp. 3103–3107, 2009.

[9] E. Pauwels, V. Stoven, and Y. Yamanishi, "Predicting drug side-effect profiles: a chemical fragment-based approach," *BMC Bioinformatics*, vol. 12, article 169, 2011.

[10] N. Atias and R. Sharan, "An algorithmic framework for predicting side effects of drugs," *Journal of Computational Biology*, vol. 18, no. 3, pp. 207–218, 2011.

[11] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[12] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: interaction networks of chemicals and proteins," *Nucleic Acids Research*, vol. 36, no. 1, pp. D684–D688, 2008.

[13] L. Hu, C. Chen, T. Huang, Y. Cai, and K. C. Chou, "Predicting biological functions of compounds based on chemical-chemical interactions," *PLoS ONE*, vol. 6, no. 12, Article ID e29491, 2011.

[14] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.

[15] L. Chen, Z. He, T. Huang, and Y. Cai, "Using compound similarity and functional domain composition for prediction of drug-target interaction networks," *Medicinal Chemistry*, vol. 6, no. 6, pp. 388–395, 2010.

[16] L. Chen, W. Zeng, Y. Cai, K. Feng, and K. C. Chou, "Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities," *PLoS ONE*, vol. 7, no. 4, Article ID e35254, 2012.

[17] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular Systems Biology*, vol. 3, article 88, 2007.

[18] P. Bogdanov and A. K. Singh, "Molecular function prediction using neighborhood features," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 2, pp. 208–217, 2010.

[19] Y. A. I. Kourmpetis, A. D. J. van Dijk, M. C. A. M. Bink, R. C. H. J. van Ham, and C. J. F. ter Braak, "Bayesian markov random field analysis for protein function prediction based on network data," *PLoS ONE*, vol. 5, no. 2, Article ID e9293, 2010.

[20] L. Hu, T. Huang, X. Shi, W. Lu, Y. Cai, and K. C. Chou, "Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties," *PLoS ONE*, vol. 6, no. 1, Article ID e14556, 2011.

[21] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs," *Molecular Systems Biology*, vol. 6, article 343, 2010.

[22] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, pp. 31–36, 1988.

[23] K. Ng, J. Ciou, and C. Huang, "Prediction of protein functions based on function-function correlation relations," *Computers in Biology and Medicine*, vol. 40, no. 3, pp. 300–305, 2010.

[24] M. Dunkel, S. Günther, J. Ahmed, B. Wittig, and R. Preissner, "SuperPred: drug classification and target prediction," *Nucleic Acids Research*, vol. 36, pp. W55–W59, 2008.

[25] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: an open chemical toolbox," *Journal of Cheminformatics*, vol. 3, article 33, 2011.

[26] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review)," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[27] C. Chen, Z. Shen, and X. Zou, "Dual-layer wavelet SVM for predicting protein structural class via the general form of Chou's pseudo amino acid composition," *Protein and Peptide Letters*, vol. 19, no. 4, pp. 422–429, 2012.

[28] M. Hayat and A. Khan, "Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC," *Protein and Peptide Letters*, vol. 19, no. 4, pp. 411–421, 2012.

[29] L. Nanni, A. Lumini, D. Gupta, and A. Garg, "Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 2, pp. 467–475, 2012.

[30] H. Mohabatkar, M. M. Beigi, and A. Esmaeili, "Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine," *Journal of Theoretical Biology*, vol. 281, no. 1, pp. 18–23, 2011.

[31] G. Fan and Q. Li, "Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 304, pp. 88–95, 2012.

[32] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[33] D. Zou, Z. He, J. He, and Y. Xia, "Supersecondary structure prediction using Chou's pseudo amino acid composition," *Journal of Computational Chemistry*, vol. 32, no. 2, pp. 271–278, 2011.

[34] D. N. Georgiou, T. E. Karakasidis, J. J. Nieto, and A. Torres, "Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 257, no. 1, pp. 17–26, 2009.

[35] X. Zhao, X. Li, Z. Ma, and M. Yin, "Identify DNA-binding proteins with optimal Chou's amino acid composition," *Protein and Peptide Letters*, vol. 19, no. 4, pp. 398–405, 2012.

[36] L. Chen, W.-M. Zeng, Y.-D. Cai, and T. Huang, "Prediction of metabolic pathway using graph property, chemical functional group and chemical structural set," *Current Bioinformatics*, vol. 8, pp. 200–207, 2013.

[37] P. Du, T. Li, and X. Wang, "Recent progress in predicting protein sub-subcellular locations," *Expert Review of Proteomics*, vol. 8, no. 3, pp. 391–404, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 401649, 12 pages
http://dx.doi.org/10.1155/2013/401649

Research Article

Network-Based Inference Framework for Identifying Cancer Genes from Gene Expression Data

Bo Yang, Junying Zhang, Yaling Yin, and Yuanyuan Zhang

School of Computer Science and Technology, Xidian University, Xi'an 710071, China

Correspondence should be addressed to Junying Zhang; jyzhang@mail.xidian.edu.cn

Received 16 May 2013; Revised 15 July 2013; Accepted 17 July 2013

Academic Editor: Tao Huang

Great efforts have been devoted to alleviate uncertainty of detected cancer genes as accurate identification of oncogenes is of tremendous significance and helps unravel the biological behavior of tumors. In this paper, we present a differential network-based framework to detect biologically meaningful cancer-related genes. Firstly, a gene regulatory network construction algorithm is proposed, in which a boosting regression based on likelihood score and informative prior is employed for improving accuracy of identification. Secondly, with the algorithm, two gene regulatory networks are constructed from case and control samples independently. Thirdly, by subtracting the two networks, a differential-network model is obtained and then used to rank differentially expressed hub genes for identification of cancer biomarkers. Compared with two existing gene-based methods (t-test and lasso), the method has a significant improvement in accuracy both on synthetic datasets and two real breast cancer datasets. Furthermore, identified six genes (*TSPYL5, CD55, CCNE2, DCK, BBC3,* and *MUC1*) susceptible to breast cancer were verified through the literature mining, GO analysis, and pathway functional enrichment analysis. Among these oncogenes, *TSPYL5* and *CCNE2* have been already known as prognostic biomarkers in breast cancer, *CD55* has been suspected of playing an important role in breast cancer prognosis from literature evidence, and other three genes are newly discovered breast cancer biomarkers. More generally, the differential-network schema can be extended to other complex diseases for detection of disease associated-genes.

1. Introduction

Treating cancer is quite difficult because more and more evidence has revealed that cancer is a kind of complex genetic disease that involves in multiple genes, proteins, pathways, and regulatory interconnections. In order to provide useful information for cancer treatment, several landmark studies [1–3] were performed to uncover oncogenes or biomarkers of cancer development, progression, or recurrence.

Gene-based approaches have emerged in recent years to identify sets of tumor-related genes, such as the "top-down" approach as defined in [4] or "minimal biological input" in 76-gene Rotterdam signature [5]. These methods usually utilize microarray gene expression profiling technique and differential expression analysis to identify the cancer-associated genes whose expression levels change significantly among patients suffering from cancer. Though they have been applied to identification of biomarkers relevant to cancer developing or progressing, the gene-based approaches suffer frequently from uncertainty of tremendous candidate genes, which limits our comprehension to the way that tumor appears and grows.

To recognize complex interaction patterns, pathways, and overrepresented biological processes, gene set enrichment analysis (GSEA) [6] has been exploited repeatedly in the gene-based approaches. The GSEA focus on groups of genes that share common biological functions or signaling pathways defined, respectively, by gene ontology (GO) [7] or KEGG [8], and so forth. Recent works also demonstrated that the detected biomarkers based on GO analysis and pathway information are more reproducible than individual marker genes [9]. Those biomarkers can also improve classification accuracy by 8% compared to the original 70 genes [1].

Increasing evidence suggests that cancer related genes are usually organized as pathways or gene networks which consist of a group of interacting genes at molecular level.

Moreover, gene signatures discovered from previous studies often enrich in common cancer-related pathways and similar biological processes. The opinion seems to be advocated and accepted by many researchers that only those which can significantly enrich in tumor-induced signaling pathways or relative biological processes are helpful and valuable for molecular diagnostics [10].

Several network-based methods have been proposed to identify novel oncogenes, subnetworks, or pathways involved in tumor progression. Chuang et al. [11] applied a protein network-based approach to identify biomarkers by extracting subnetworks from protein interaction databases. They also demonstrated that biomarkers detected with the network-based method are more reproducible than individual marker genes selected without network information. Wu et al. [12] integrated different types of networks and known gene-phenotype association information to compute similarity score and predict disease genes. Fröhlich [13] constructed a consensus signature by mapping different gene signatures on a protein interaction network, in which a clustering algorithm was performed based on shortest path distances of different genes in a protein-protein interaction network. In addition, Chen et al. [14] developed a network-constrained support vector machine approach for cancer biomarker identification. The method results in an improved prediction performance with network biomarkers by integrating gene expression data and protein-protein interaction data.

Differential network analysis plays a key role for elucidating fundamental biological responses as well as discovering important differences between the different biological states [15]. In contrast with conventional gene-based methods, by performing the differential network analysis, more characteristic genes or subnetworks known to be related to disease development are identified. Valcárcel et al. [16] inferred a differential network from males with normal fasting glucose (NFG) and impaired fasting glucose (IFG), in which shrinkage estimates of the partial correlation are executed for network construction, and then the differences were explored by utilizing statistical tests between the two defined groups (NFG and IFG). Gambardella et al. [17] developed a powerful procedure named DINA to identify tissue-specific pathways using a slightly modified information entropy measure. Although it can discover differences across a set of networks, DINA is not able to detect distinct network topologies that have equal density. Iancu et al. [18] revealed gene coexpression patterns and detected modules using a custom differential network analysis procedure including correlation coefficient, clustering, and permutation test. In addition, West et al. [19] presented differential network entropy and demonstrated that gene expression differences between normal and cancer tissue are anticorrelated with local network entropy changes. These findings may have potential implications for identifying novel oncogenes.

In this paper, we present a novel differential-network based inference framework, called network-based statistical analysis method (netSAM) to detect oncogenes. Using differential network modeling and functional enrichment analysis rather than purely the differential expression analysis of a single gene or pathway, netSAM overcomes some limitations of the gene-based methods, such as uncertainty of identification or unfitness for generalization. The applicability and effectiveness of the netSAM algorithm are demonstrated on simulated and real data through numerous experiments. Our results show that the netSAM outperforms two gene-based methods (t-test and lasso) in accuracy, precision, and overlap ratio, and so forth. Furthermore, we applied netSAM to identify breast cancer genes from two benchmark datasets (Wang et al. and Van De Vijver et al.) and obtained a cancer-associated gene signature consisting of 6 genes (*TSPYL5, CD55, CCNE2, DCK, BBC3,* and *MUC1*), which have been proven biologically reasonable via GO and pathway analyses. The literature mining reveals that the resulting signature possesses higher prediction capability compared to previous work, and it would be useful in both predicting metastasis of breast cancer and facilitating treatment decision.

Our contributions in this paper are composed of three aspects. First, a novel gene regulatory network construction algorithm is proposed, and its inference ability is demonstrated accurately and efficiently. Second important contribution is a scale-free property-based informative prior score. Third, another important contribution of the proposed method is the differential-network schema for the identification of oncogenes. This framework can be extended easily to other complex diseases.

The remainder of the paper is organized as follows. In Section 2, we provide all details of the netSAM. Section 3 presents experimental results and analysis. Conclusions and future works can be found in Section 4.

2. Materials and Methods

2.1. Differential Network-Based Inference Framework. We propose a new differential network-based scheme netSAM to evaluate the relative importance of genes based on the linkage characteristics of the entire network. Firstly, the netSAM explores the transcriptional regulatory mechanism underlying distinct cancer phenotypes by filtering genes that are differentially expressed as well as by inferring differential network from "case" and "control" samples. Secondly, the netSAM selects the top-scoring interacting genes, which appear to construct the cancer-related subnetwork, as candidate genes of cancer susceptibility. In this process, we assume that the higher score a gene has, the more likely it is a cancer-associated gene. Finally, we investigate the functional enrichment of top-ranked genes and evaluate reliability of the biomarkers. The overall work flow for the present study is described as follows.

Compared to the gene-based methods, the advantages or features of the netSAM include (a) identifying oncogenes by constructing the differential network rather than differentially expressed analysis, (b) focusing on the "hub" genes which provide insights into the functional modules or pathways, and (c) uncovering gene regulatory relationships via network inference as well as characteristic of the scale-free network.

In general, the differential network-based detection of cancer genes includes five steps as described in Figure 1.

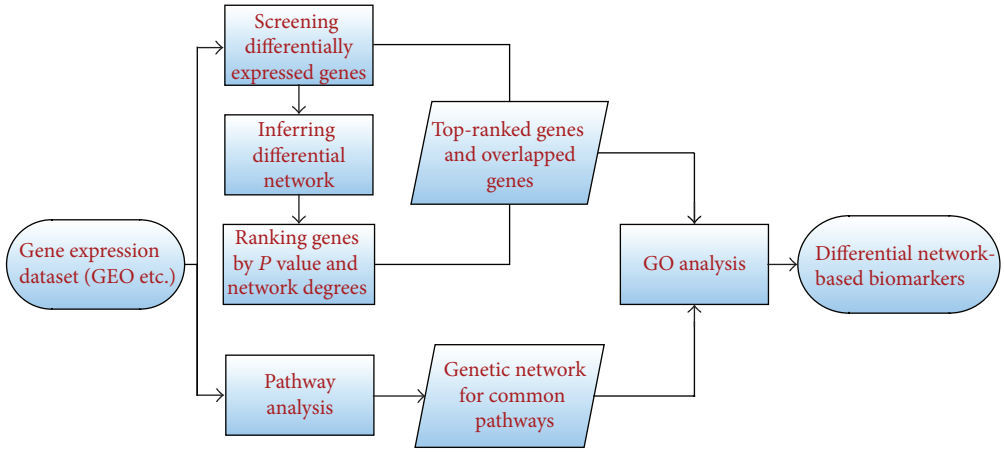


Figure 1: The flowchart of the scheme: the differential network-based identification of cancer biomarkers.

Step 1. Extracting differentially expressed (DE) genes. In order to remove the features (genes) that show no or minimal discriminatory ability, gene expression data are firstly processed with log2 transformation and then differentially expressed genes are determined by two criteria, fold change of expression level as well as P value of Student's t-test. In this step, genes whose value of fold-change larger than 2, and meanwhile P value less than 0.01 are considered as DE genes.

Step 2. Inferring "case" and "control" networks. Based on the reduced features, the individual network is inferred respectively, upon "case" and "control" samples including three steps as the following.

(i) Computing all regression coefficients β that represent interactions between genes, where $\widehat{\beta}_{ij} = \mathbf{x}_j^T \mathbf{x}_i$ $(i \neq j, 1 \leq i, j \leq p)$, $\mathbf{x}_j^T$ means the transpose of $\mathbf{x}_j$ and $\mathbf{x}_j$ is the expression profile of gene j.

(ii) Calculating a posteriori score based on likelihood score and informative prior. To make the degree distributions of the underlying network satisfy the scale-free property, the netSAM employs power-law distribution and linear correlation to construct a prior probability distribution. Next, performing boosting updates to obtain an optimal estimation of coefficients β.

(iii) Constructing the gene regulatory network $\mathbf{G}$ and the adjacency matrix can be formulated as $\mathbf{G}_{ij} = \begin{cases} 0, & \text{if } \operatorname{sgn}(\widehat{\beta}_{ji}\widehat{\beta}_{ij})=0 \\ 1, & \text{otherwise} \end{cases}$, where $\operatorname{sgn}(\cdot)$ denotes the sign function.

In the network $\mathbf{G}$, the weights of the edges are set to 1 when there is a connection between two genes and 0 otherwise.

Details of Step 2 are given in Algorithm 1 of appendix. Reasons for choosing the boosting regression include (a) the adaptability to achieve the optimal balancing of variance and error, (b) the ability to easily identify genes, and (c) a high computational accuracy and a low calculation time.

Step 3. Constructing the differential network. Upon the two networks obtained from "case" and "control" samples, a differential network is established through comparing difference of interactions and subtracting "case" network from "control" one.

Comparison of the "case" and "control" genetic networks will reveal many discrepancies, for example, some interactions are unique and only exist in either networks. As pointed in [15], through network subtraction, the trivial interactions can be removed and detection of differentially represented pathways can be performed. Differentially genetic interactions can also be used to identify novel cancer metastasis-dependent pathways. Thus, network comparison reflects a landscape of differential genetic interactions especially to the genetic disease response.

Step 4. Identifying differential network hubs. After building the differential network, the genes are ranked to identify network hubs according to P value and degree, that is, number of interactions. With established thresholds for differential interactions (degree $\geqslant 5$, P value $\leqslant 0.01$), the hubs of the differential network are identified.

The network hubs, that is, genes with many interactions, regulate a variety of cellular functions and are essential for gene-induced lethality or sickness. As previous investigations have shown [17, 20], many hubs in differential interaction network serve as key components of cancer-associated pathways and can be used to discover cancer-induced genes. Our experimental results on breast cancer data also demonstrated that two such differential network hubs, *DCK* and *BBC3*, link to MAPK signaling pathway and metastasis, as reported in [21, 22].

Step 5. Prioritizing the network hubs. To identify the most potential cancer genes among the genetic interaction hubs, prioritization of the differential-network hubs is performed.

For the network hubs discovered from the previous Step 4, two ranking methods are performed to select the most promising genes that are associated with cancer. Firstly, GeneRank [23] is employed as computational strategies for gene

INPUT: $n \times p$ dimensional data matrix $\mathbf{X} = (x_1, \ldots, x_i, \ldots, x_p)$, ($n$ denotes the number of samples and p represents the number of genes, $1 < i < p$).
OUTPUT: An adjacency matrix of graph $\mathbf{G}$.

(1) Initialize iterating counter $k = 0$, coefficient of the regression $\widehat{\beta}_{ij} = 0$ for all $1 \le i, j \le p$, $\lambda = 2$, and step size $\varepsilon = 0.01$, $S = 10 \cdot p$.

(2) Standardize column vector $\mathbf{x}_i$ to zero mean and unit norm for gene i $(1 < i < p)$. Set residual $\mathbf{R}_i^{(k)} \leftarrow \mathbf{x}_i$.

(3) Fit $p \times (p-1)$ regressions $\widehat{\beta}_{ij} = \mathbf{x}_j^T \mathbf{R}_i^{(k)}$ $(i \ne j,\ 1 \le i,\ j \le p)$, where $\mathbf{x}_j^T$ means the transpose of $\mathbf{x}_j$.

(4) Calculate a posteriori score based on likelihood score and informative prior as following
$\log\left(\text{posteriori score}_{ij}\right) = -\log\left(\text{RSS}_{ij}\right) + \log\left(C^2\right)$
$C = r\left(\log \pi(d),\ \log(d)\right)$,
$\pi(d) = d^{-\lambda}\ (\lambda > 0), \quad d = \sum_{1 \le m \le p} \left|\widehat{\beta}_{mj}\right|$,
where RSS refers to (3) and $r(\cdot, \cdot)$ denotes Pearson's correlation coefficient.

(5) Find the edge having the best score of posteriori probability
$\left(\widehat{i}, \widehat{j}\right) = \underset{i,j}{\text{argmax}} \log\left(\text{posteriori score}_{ij}\right)$

(6) Perform the boosting update
$\widehat{\beta}_{\widehat{i},\widehat{j}}^{(k+1)} \leftarrow \widehat{\beta}_{\widehat{i},\widehat{j}}^{(k)} + \delta_{\widehat{i},\widehat{j}}^{(k)}$, and
$\delta_{\widehat{i},\widehat{j}}^{(k)} = \varepsilon \cdot \widehat{\beta}_{\widehat{i},\widehat{j}} \cdot \text{sgn}\left(\mathbf{x}_{\widehat{j}}^T \mathbf{R}_{\widehat{i}}^{(k)}\right)$, where $\varepsilon > 0$
and set $\mathbf{R}_{\widehat{i}}^{(k+1)} \leftarrow \mathbf{R}_{\widehat{i}}^{(k)} - \delta_{\widehat{i},\widehat{j}}^{(k)} \mathbf{x}_{\widehat{j}}$ and increment counter $k \leftarrow k + 1$

(7) Repeat Steps 3 to 6 until $k = S$.

(8) Calculate $\widehat{\rho}_{ij}$ from coefficient matrix $\widehat{\boldsymbol{\beta}}^{(S)}$,
if $\text{sgn}\left(\widehat{\beta}_{ij}^{(S)}\right) = \text{sgn}\left(\widehat{\beta}_{ji}^{(S)}\right)$,
$\widehat{\rho}_{ij} = \min\left\{1, \sqrt{\widehat{\beta}_{ij}^{(S)} \widehat{\beta}_{ji}^{(S)}}\right\}$
otherwise, $\widehat{\rho}_{ij} = 0$,
where sgn $(\cdot)$ denotes the sign function.

(9) Return an adjacency matrix of network

$$\mathbf{G} = \left\{e_{ij} \mid e_{ij} \in \{0, 1\}\right\}, \quad e_{ij} = \begin{cases} 0, & \text{if } \widehat{\rho}_{ij} = 0 \\ 1, & \text{otherwise} \end{cases}$$

Algorithm 1: Posteriori score-based boosting regression algorithm for inferring networks as Step 2 of netSAM.

prioritization in the netSAM. The reason to apply GeneRank is that the algorithm does not require a predefined threshold of important genes and can provide a reordering of genes in terms of their importance and connectivity in the entire network. In GeneRank, a node represents a gene, and an edge is described with expression profile correlation coefficients. Additionally, it requires a connectivity matrix of the network, a vector of differential expression level, and a controlling parameter as its inputs. Specially, to acquire the ranking result based on the connectivity of the network as well as the differential expression level of genes, the controlling parameter is set to 0.5. Secondly, the all hubs are ranked additionally according to their degrees (i.e., number of edges) in the differential network. As the degree can indicate connectivity information and importance of each gene in the whole network, a new ranking result, which is different from that of GeneRank, will be generated by utilizing the sort of degrees. Thirdly, the common genes are selected from the top-ranked hubs of the two previous ranking lists. Thus, a candidate gene-set comprised of essential regulatory hubs of the differential network is obtained. In virtue of the candidate gene-set, a signature consisting of cancer-associated genes is determined finally.

In summary, since the differential network spans the validated differentially expressed genes, the cancer-related genes discovered upon it can provide stronger predictive power than traditional gene-based method. Based on differential network inference framework, the netSAM can be extended easily to a majority of currently known genetic diseases.

2.2. Bayesian Criterion and the Posteriori Score. Our method assumes that the individual network **G** can be scored according to its posterior probability given that data is known. The main idea is to choose the edge with the largest score which takes into account the likelihood as well as the prior information of scale-free network. The notion of the most probable network structure is made formal by the Bayesian score criterion, which is simply the posterior probability of **G** given **X**:

$$P(G \mid \mathbf{X}) = \frac{P(G, \mathbf{X})}{P(\mathbf{X})} \propto P(\mathbf{X} \mid G) \cdot P(G), \quad (1)$$

$$\log P(G \mid \mathbf{X}) \cong \log P(\mathbf{X} \mid G) + \log P(G). \quad (2)$$

Here, **X** is matrix of gene expression data, $P(\mathbf{X} \mid G)$ means the (marginal) likelihood probability, and $P(G)$ means a prior distribution over network structure **G**.

Based on the above discussion, the combined score measure consists of two parts: one is the approximate likelihood, and the other is the network prior information.

Residual sum of squares (RSS) is a measure of the discrepancy between the data and an estimation model. A small RSS indicates a tight fit of the model to the data. Residual sum of squares can be represented as follows:

$$\mathrm{RSS} = \sum_{k=1}^{n} \left(\mathbf{x}_{ki} - \widehat{\mathbf{x}}_{ki}\right)^2 = \sum_{k=1}^{n} \left(x_{ki} - \sum_{j=1, j \neq i}^{p} \widehat{\beta}_{ij} x_{kj} \right)^2, \quad (3)$$

where n denotes the number of samples and p represents the number of features (genes).

Accordingly, we define the approximate log-likelihood score for connectivity strength across genes as follows:

$$\log(\text{likelihood}) = -\log(\mathrm{RSS}). \quad (4)$$

To capture the mechanisms underlying biology systems and complex networks, a scale-free network prior is applied. Scale-free property means that the frequency distribution $\pi(d)$ of the connectivity in a network follows a so-called power law: $\pi(d) \sim d^{-\lambda}$, where d equals the number of node degrees in the network. The prior information over the network **G** describing scale-free property can be encoded as the following:

$$\begin{aligned} \log(\text{priori}) &= \log\left(C^2\right), \\ C &= r\left(\log \pi(d), \log(d)\right), \\ \pi(d) &= d^{-\lambda} \quad (\lambda > 0), \quad d = \sum_{1 \leq m \leq p} \left|\widehat{\beta}_{mj}\right|. \end{aligned} \quad (5)$$

To quantify the association or connection between genes, we define the parameter $\widehat{\beta}$ as regression coefficient obtained using boosting regression algorithm. Based on likelihood and informative prior, we thus can rewrite (2) as the following:

$$\log(\text{posteriori score}) = -\log(\mathrm{RSS}) + \log\left(C^2\right). \quad (6)$$

We applied score function mentioned above to select the fitted edge of network **G** by computing the largest posterior score for all possible gene interaction. Solving this problem leads to the following optimal estimate problem:

$$\widehat{\beta}_i = \underset{\beta_i}{\operatorname{argmax}}\ \text{posteriori score}(\mathbf{G}; \mathbf{X}). \quad (7)$$

Here, β_i denotes all the coefficient of gene i $(1 < i < p)$ regressed upon other genes through boosting method.

2.3. Functional Enrichment Analysis of Candidate Genes. Gene set enrichment analysis (GSEA) [6] is a computational tool that investigates whether a predefined gene set shows statistical significance. A gene set that contains terms of biological process of gene ontology is constructed, and then overrepresented GO categories are investigated in the detected cancer gene signature by conducting GO analysis using the BiNGO plug-in of Cytoscape [24]. Gene ontology functional enrichment analysis is employed, in which the hypergeometric test is used for functional overrepresentation and false discovery rate for the multiple hypotheses testing correction. Only the corrected P values less than 0.05 are considered significant.

Besides, associations between differential genetic interactions and known pathways are investigated. As shown in the differential network, differential genetic interactions are much more likely to occur among pairs of genes connecting two different subnetworks than among pairs of genes within the same subnetwork. On the basis of these findings, a map of genes and their differential genetic interactions is constructed, in which some of hubs have not been previously linked to cancer development. To validate the newly identified oncogenes, a pathway analysis is performed using DAVID and the parameters are set as default numbers. The significantly enriched functional modules based on KEGG [8] pathway are investigated.

In brief, GO and pathway analyses indicate the effectiveness of the netSAM, which highlights potential application of the method that may be prominent when developing targeted therapeutics. It is also reasonable to believe that the genes detected by the netSAM are highly relevant to cancer either by sharing common cancer-related signaling pathways or by GO functional terms.

3. Results and Discussion

In this section, results of experiments with synthetic and real-world data sets are included. We performed a numerical comparison with two existing algorithms (t-test [25] and lasso [26, 27]), including GO and pathway analyses. While they provide efficient inference for medium-scale data, t-test and lasso typically cannot fully capture the relational complexity for large-scale datasets. Experiments demonstrated the reliability and the effectiveness of the netSAM algorithm. Furthermore, our algorithm occupied a higher position in the accuracy/efficiency trade-off. In addition, validation of the biological reasonability of detected genes as biomarkers was done through analysis of functional enrichment and a vast amount of independent literature.

3.1. Simulated Data Experiments. In order to estimate the accuracy of netSAM algorithm and compare its performance

with two commonly used gene-based algorithms, that is, t-test and lasso, we generated synthetic data sets by using the SynTReN [28], which simulates benchmark microarray datasets with the known underlying biological networks for the purpose of developing and testing new network inference algorithms. Through SynTReN, we simulated a biological network with a known topological structure as well as the corresponding gene expression data. Although numerous tuning parameters can be changed to generate datasets of different sizes and complexity in the software, we kept the default tuning parameters controlling the complexity aspects and only changed the ones controlling noise and the size of dataset being generated.

We generated 100 microarray datasets that consist of 200 genes and 100 sample points (noise $\sigma = 0.5$); the resulting graphs had approximately 500 connections. For each generated data set, the network structure learned from each method was then compared with the true underlying structure. We ran each experiment 10 times and averaged the results.

3.2. Comparison of the Accuracy and Robustness with t-Test and Lasso. Using the synthetic dataset described above, we evaluated the accuracy and robustness of different identification approaches via receiver operating characteristic (ROC), area under curve (AUC), positive predictive value (PPV), and false discovery rate (FDR). ROC, AUC, and PPV will have a value of 1 if the method can perfectly identify the connections in the genetic network.

Seen from Figure 2, the netSAM algorithm gets comparatively lower FDR and higher PPV for more edges than t-test and lasso. Additionally, robustness, AUC against SNR (signal-to-noise ratio), of biomarker identification over three algorithms is shown in Figure 2(d). In the figure, the average AUC of netSAM is about 0.8, which means that netSAM can select more suited gene biomarkers than t-test and lasso. On the contrary, lasso obtains the worst performance over four measures against the other two algorithms. It should be emphasized that these measures depict the inference ability of three algorithms on the same underlying network.

3.3. Identification of Breast Cancer-Associated Genes Using NetSAM. In real data experiment, we applied netSAM to breast cancer gene expression microarray dataset previously reported by Wang et al. [29] and Van De Vijver et al. [1]. Only those patients with estrogen receptor positive breast cancer are used as "case" samples, and the remaining estrogen receptor positive samples are assigned to "control" group. Both case and control samples are included in our experiments. After that, the netSAM is applied to the two datasets separately to get two breast cancer gene-set candidates. Finally, they are ranked and intersected for detection of breast cancer genes.

Wang et al. dataset was downloaded from NCBI GEO [30] database GSE2034 [29]. It employs the expression of 22,000 transcripts from total RNA of frozen tumor samples from 286 lymph node-negative primary breast cancer samples that contained 77 estrogen-receptor negative (ER−) and 209 estrogen-receptor positive (ER+) samples, and gene expression profiles were analyzed with Affymetrix Human Genome U133A Array (HG-U133A). Van De Vijver et al. [1] gene expression dataset consists of 295 samples, including 151 lymph node-negative disease and 144 lymph node-positive disease. There are approximately 25,000 human genes which were transcribed and labeled to microarrays for each sample.

Estrogen receptors (ERs) are a group of proteins found inside cells. Once activated, the ER is able to bind to DNA to regulate the activity of different genes. Estrogen receptor positive tumors are the most important subtype of breast cancer. A significant majority (about 70%) of women who died with breast cancer have estrogen receptor positive (ER+) tumors. In these cases, estrogen receptors are overexpressed and referred to as "ER-positive." While molecular biology has broadened our understanding of breast cancer, we still lack sufficient knowledge of estrogen receptor positive tumors. Aiming at promoting comprehension on estrogen signaling and regulation mechanism contributing to tumorigenesis, we, therefore, focused on patients with estrogen receptor positive breast cancer. In the experiments, we chose 80 samples in Wang et al. and 78 ones in Van De Vijver et al. among the estrogen-receptor positive patients. These selected patients had been diagnosed with metastasis during their follow-up visits within 5 years of surgery and were labeled as "case" group in our study. The remaining 129 and 217 samples, respectively, in the two studies, were then assigned to "control" group.

Using netSAM, 761 and 938 differential genetic interactions were identified totally on the two datasets, respectively, among which 342 and 461 interactions were "positive," which indicated inducible epistasis, whereas 419 and 477 were "negative," which indicated suppression. Moreover, we detected 119 hub genes on Wang et al. dataset and 162 on Van De Vijver et al. dataset. A subset of 76 genes was found common between the two candidate gene-sets (119 and 162 genes, resp.). Results of GO and pathway enrichment analyses for the 76 intersection genes are shown in Sections 3.5 and 3.6.

To obtain a breast cancer gene signature, we firstly selected the top 10 ranked genes, respectively, from the two candidate gene-sets (119 and 162). Then, an intersection set was generated between two top 10 ranked gene sets. Finally, six intersection genes were regarded as the breast cancer susceptibility genes, that is, the signature consisting of *TSPYL5, CD55, CCNE2, DCK, BBC3,* and *MUC1*.

In addition, the top 50 ranked genes identified by netSAM from Wang et al. dataset are shown in Figure 3. Seen from Figure 3, not only the known breast cancer metastasis genes (*BRCA1, TP53,* and *ERBB2*) but also the novel cancer susceptibility genes such as *TSPYL5, CD55, CCNE2, DCK, BBC3,* and *MUC1* were identified. These recognized genes interact with many other genes to coregulate the progression and evolvement of breast cancer. The node size is relevant to the breast cancer susceptibility which represents the possibility of gene relating to cancer. Figure 3 was created using Cytoscape [24].

3.4. Overlap Analysis between Identified Signature and Literature Reference Gene Set. In this section, we compared the netSAM with gene-based approaches (t-test and lasso)

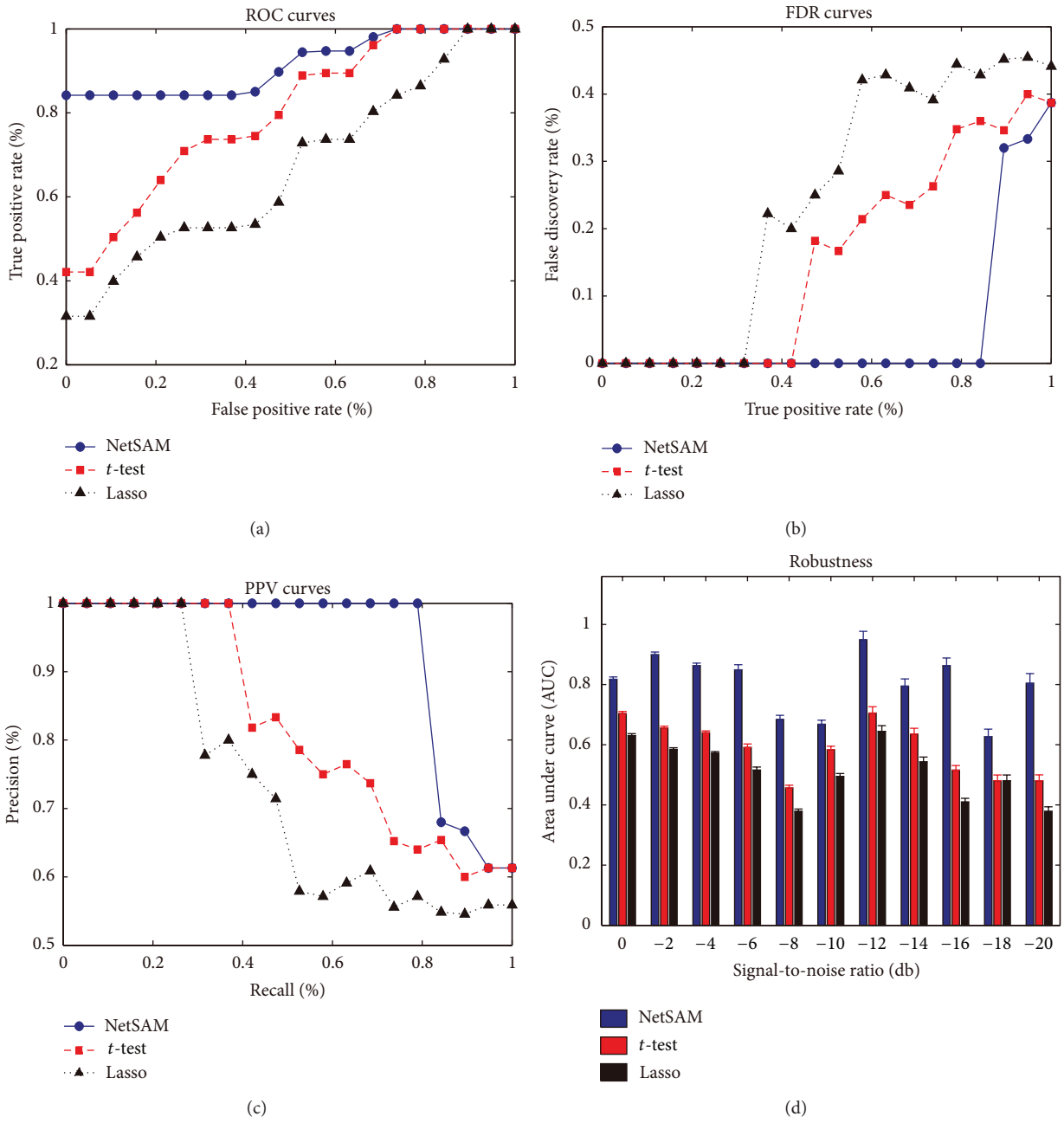


FIGURE 2: Comparison of accuracy and robustness between netSAM, t-test, and lasso on 100 synthetic datasets. (a) ROC curves: true positive rates against false positive rates. (b) FDR curves: error discovery rates against true positive rates. (c) PPV curves: precision versus recall value. (d) Robustness values (AUC versus SNR) are calculated based on five-fold cross-validation, where standard deviations are shown in error bars.

on the breast cancer datasets to further examine which method would obtain better signature. To compare overlapped genes through literature mining, we also compiled a list of cancer-associated genes, BCGS (breast cancer literature reference gene set), by collecting genes known to be associated with breast cancer from literature curation and web-based resources. BCGS includes 452 representative cancer-associated genes. The gene symbols were searched as well as extracted from the 1098 PubMed literatures using keyword (breast cancer* gene AND Humans [mesh] OR "Breast Neoplasm" [mesh] AND "Neoplasm Metastasis" [mesh] biological process [go]) in PubMed [31]. These genes form the basis of our "cancer-associated genes" dataset. We then utilized overlap ratio between literature-published gene set BCGS and our candidate genes as evidence of feasibility and the effectiveness of the netSAM.

When two distinct sets share at least one element in common, they are "intersecting" or "overlapping." In the genomic scenario, we utilized an overlap measure to examine the overlapping capability between the curated gene set BCGS

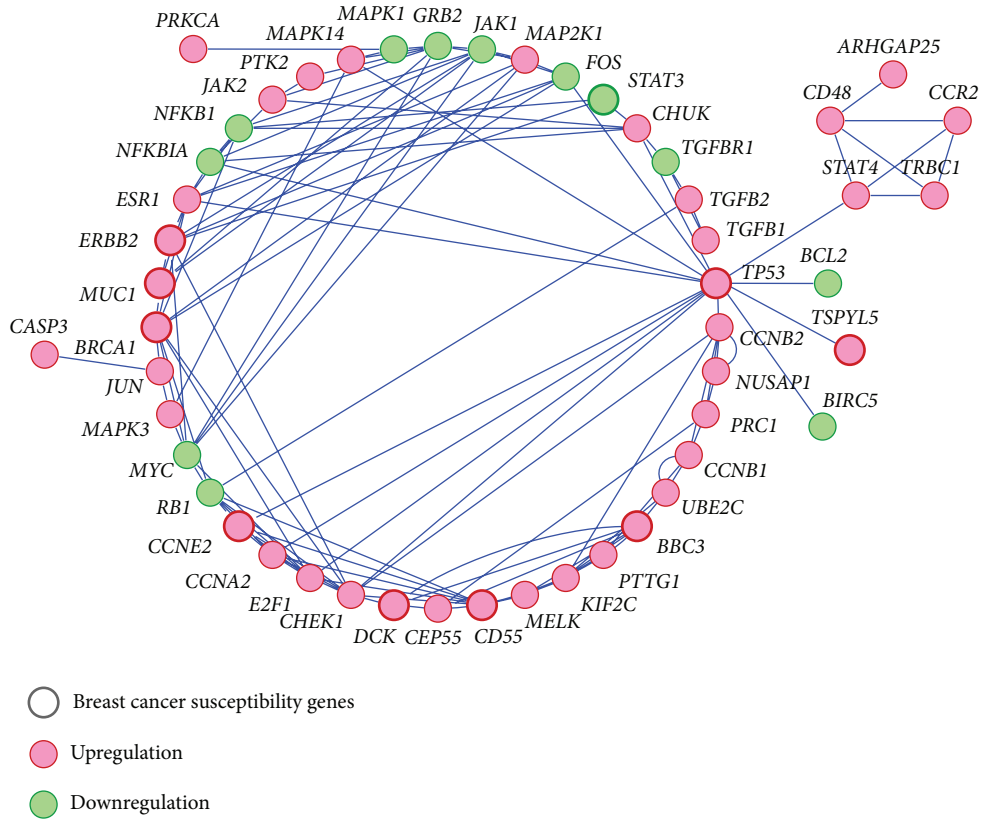


FIGURE 3: The breast cancer-related genetic subnetwork consisting of the top 50 ranked genes identified through netSAM method from Wang et al. breast cancer dataset. Genes are represented as circles, a significant coregulation between two genes as a line.

and the cancer gene set identified with different detection algorithms. Specifically speaking, the overlap ratio is defined as the number of intersection genes divided by the number of identified genes.

To validate predictive power of netSAM, the overlap ratio and trend analysis of overlap are performed. The comparison results among netSAM, t-test, and lasso are displayed in Figure 4 based on Wang et al. and Van De Vijver et al. breast cancer datasets. The comparison of overlap ratio indicates that netSAM can identify some novel cancer-causing genes that are not found by t-test and lasso. Only a few of known breast cancer genes were identified correctly by t-test and lasso. Seen from Figure 4, the netSAM can identify more overlapped genes than those of the other two methods, which indicates that the netSAM obtains a better reproducibility across different data sets in terms of biomarker identification. Furthermore, Figure 4 also shows that a number of candidate genes (about 60%) identified by netSAM significantly overlap with known breast cancer genes in BCGS. Accordingly, we can conclude that netSAM is a more effective approach for identifying biomarkers.

Although BCGS consists of 452 genes based on the results of searching the related articles referenced in PubMed [31], until now, however, most of genes still have not been proved to be breast cancer susceptibility genes with absolute certainty. Thus, when these genes are used as true breast cancer genes to test the performance of our method, it would potentially cause some bias.

3.5. GO Analysis. Most cancers, including breast cancer, are complex disorders that are generally caused by multiple genes and their complex interactions. By mapping the 76 intersection genes identified by the netSAM to the gene ontology (GO) [7] terms, we found 11 GO functional categories, given in Table 1. The obtained GO terms are consistent to those in curated literature [32], which suggested that the above categories largely captures the functional facets of the breast cancer-specific gene network. Several cellular processes such as metabolism, cell proliferation and replication, apoptosis, inflammation, and cell cycle are known to be pivotal for tumorigenesis. The result of GO analysis indicates that our discovered signature has an enrichment score (ES) of 0.79, which means that identified oncogenes contain the majority of genes contributing to the enrichment score.

The full detail of Gene Ontology enrichment analysis is shown in Table 1. Tumor genes identified by netSAM are enriched in important biological processes catalogued in the Gene Ontology. From Table 1, it can be seen that the detected oncogenes are significantly enriched in GO terms of apoptosis, metabolism, immune response, and cell cycle. Inflammatory response is overrepresented and can be considered as potential candidate because chronic inflammation is

Table 1: Significantly enriched GO terms of biological process via BiNGO functional annotation analysis for the 76 intersection genes.

GO term	Hypergeometric test P value	Benjamini correction P-value	Frequency of mapped genes (%)	Fisher P-value
Immune system process	$1.5280E-14$	$1.7847E-11$	33.3	$2.3E-12$
Cell cycle	$3.5350E-12$	$2.0645E-9$	20.4	$1.3E-12$
Immune response	$6.2486E-12$	$2.4328E-9$	24.7	$1.3E-9$
Cell division	$1.5915E-11$	$4.4740E-9$	18.2	$1.3E-11$
Nuclear division	$2.2983E-11$	$4.4740E-9$	16.1	$7.2E-12$
Apoptotic process	$2.2983E-11$	$4.4740E-9$	16.1	$7.2E-12$
Metabolism	$3.9513E-11$	$5.7689E-9$	16.1	$1.3E-11$
Cell proliferation	$1.0537E-10$	$1.2307E-8$	22.5	$3.4E-11$
Inflammatory response	$5.4845E-8$	$4.2706E-6$	41.9	$1.4E-10$
Response to stimulus	$6.6080E-5$	$1.9433E-3$	44.0	$5.6E-10$
System development	$5.1327E-4$	$8.4436E-3$	31.1	$2.3E-11$

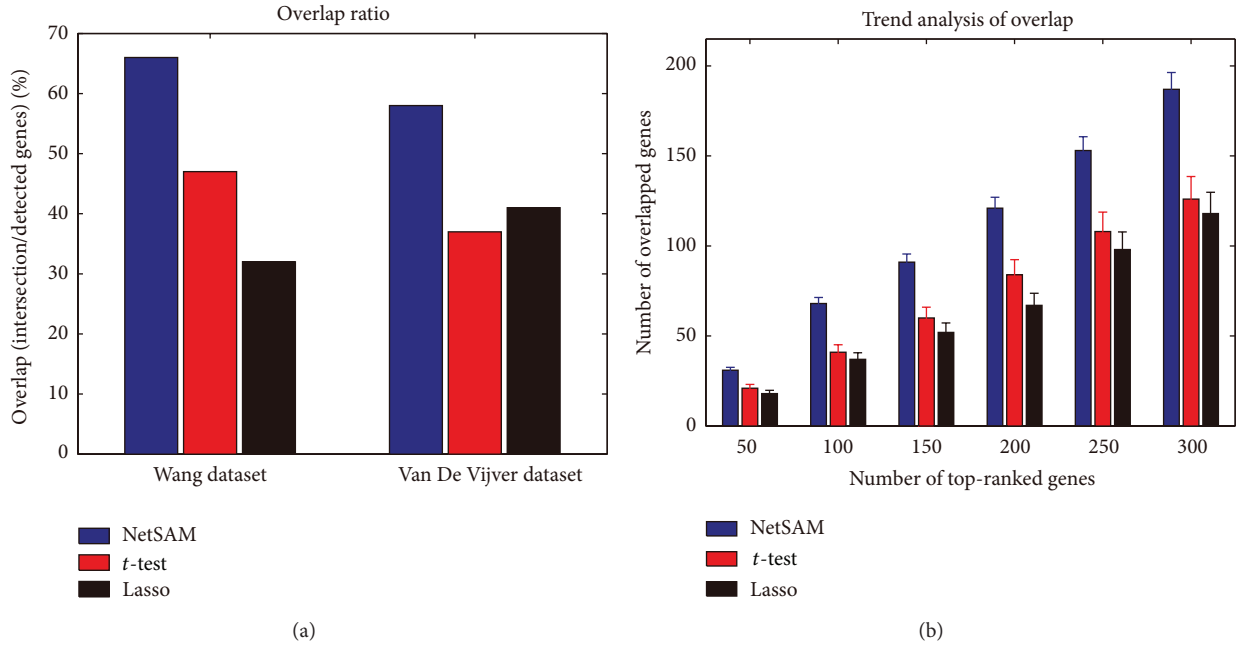


Figure 4: (a) Overlaps of the identified genes using netSAM, t-test, and lasso based on Wang et al. and Van De Vijver et al. breast cancer datasets. (b) Trend of overlap: number of overlapped genes versus top-ranked genes (error bars denote standard deviation estimated over 100 tests).

widely believed to be a predisposing factor for cancer. These results suggested that the above categories largely captured the functional facets of the breast cancer specific gene.

3.6. KEGG Pathway Functional Analysis. Gene set enrichment analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) [8] pathway was conducted to find additional supporting evidence as described in Table 2. The enriched pathways were found. In the enriched pathways, TGF-beta, p53, and Notch and JAK-STAT signaling pathways are frequently reported to be related to breast tumor metastasis [33]. Notch signaling pathway may play essential role in the cross-talk between metastasis and relapse free. Recently, it has been found that *p53* activates the MAPK pathways through a feedback loop in human cancer. Moreover, we found that the detected genes were enriched for many known pathways, such as Apoptosis and Cell cycle. DAVID [34] genetic disease class category analysis indicated that the Benjamin P value of Apoptosis and Cell cycle is $1.1E-6$ and $3.3E-4$, respectively. Six hub genes (*TSPYL5, CD55, CCNE2, DCK, BBC3,* and *MUC1*) were all proved cancer-related hub genes. From Table 2, one can conclude that identified six genes significantly enriched in ECM, P53, and cell cycle pathway.

The signaling pathways depicted in Figure 5 include MAPK and JAK-STAT signaling pathways, which were highlighted in the top-ranked cancer-related genetic network identified by netSAM method from Wang et al. breast cancer dataset.

Table 2: KEGG pathway functional analysis via DAVID for the 76 intersection genes.

KEGG pathway	Count	Frequency (%)	P value	Benjamin
Viral myocarditis	10	10.4	$1.6E-8$	$1.0E-6$
Apoptosis	8	8.3	$3.3E-8$	$1.1E-6$
Type I diabetes mellitus	8	8.3	$1.0E-7$	$1.7E-6$
Autoimmune thyroid disease	8	8.3	$4.2E-7$	$5.3E-6$
Cell cycle	9	9.4	$3.1E-5$	$3.3E-4$
TGF-beta signaling pathway	8	8.3	$1.7E-4$	$1.2E-3$
Notch signaling pathway	6	6.2	$3.9E-3$	$2.4E-2$
ECM-receptor interaction	5	5.2	$8.3E-3$	$4.8E-2$
JAK-STAT signaling pathway	7	7.3	$1.2E-2$	$6.2E-2$
P53 signaling pathway	4	4.2	$4.9E-2$	$2.1E-1$
Immune network	3	3.1	$8.0E-2$	$3.0E-1$

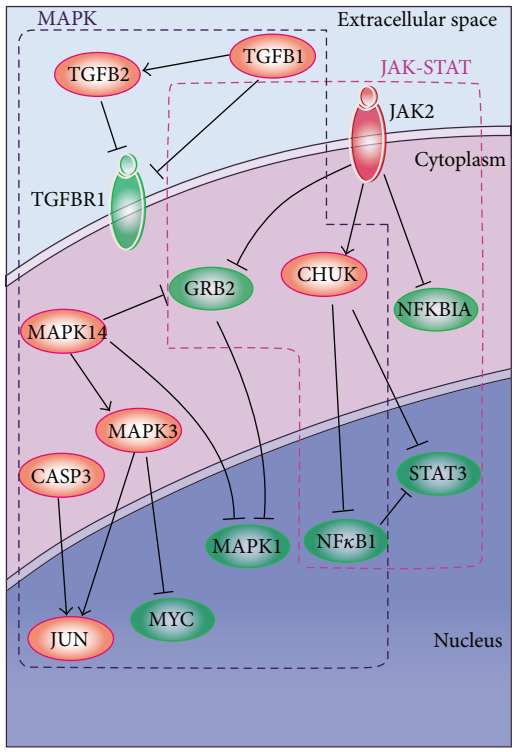


Figure 5: Signaling pathways highlighted in the identified cancer-related genetic network by netSAM on Wang et al. dataset, including MAPK and JAK-STAT pathways.

4. Conclusions

In this paper, we proposed netSAM to identify breast cancer-related genes from two benchmark breast cancer datasets (Wang et al. and Van De Vijver et al). Using netSAM, we identified six novel genes (*TSPYL5, CD55, CCNE2, DCK, BBC3,* and *MUC1*) as cancer biomarkers for predicting survival and metastasis in patients with breast cancer. Each of 6 genes in our signature not only has links to potential cancer relapse through the literature, they also have been shown in most cases to be directly linked to prognostic outcome, metastasis, and apoptosis. Furthermore, the six novel genes identified in our experiments are overlapped with the breast cancer gene set BCGS of literature curation. Further functional enrichment analysis and independent literature evidence also confirm that our identified potential cancer-causing genes are biologically reasonable, indicating the effectiveness of our method. Moreover, nearly 60% of the 119 oncogenes found by netSAM were certified as breast cancer susceptibility genes or known cancer-associated genes through literature mining. Our results indicate that the resulting signature possesses higher prediction precisions compared to previous work in the area and might be useful in predicting metastasis of breast cancer and facilitating treatment decisions.

TSPYL5 (TSPY-like 5), also known as *KIAA1750*, is involved in nucleosome assembly, a process which can alter the regulatory mechanisms of a cell [35], which is likely to occur in cancer. *TSPYL5* has been previously used as a prognostic biomarker in breast cancer [36]. In addition, it has been noted to play a role in the circulation of luteinizing hormone (LH), which is known to prompt tumor growth in breasts. Moreover, the individual gene (*TSPYL5*) is present in the 17 genes selected by Alexe et al. [3]. *CD55* has been used previously as a prognostic biomarker in gastric cancer. *CD55* has been shown to be important in breast cancer prognosis [37].

CCNE2 encodes a protein similar to cyclin that serves as regulators of cyclin dependent kinase (CDK). A significant increase in the expression level of this gene was observed in tumor-derived cells. *CCNE2* has also been conformed to qualify as independent prognostic markers for lymph node-negative breast tumor patients and reported to have a predictive value in ER positive cases among breast cancer patients [9].

The *DCK* (deoxycytidine kinase) gene is required for the phosphorylation of several deoxyribonucleosides and their nucleoside analogs. It has been used to study resistance to chemotherapy in myeloid leukemia (AML) and breast cancer patients [21]. In addition, this particular gene may catalyze the metabolic activation of gemcitabine, a drug that has been used to treat several different types of cancer. However, the exact function of this gene is still unknown.

The *BBC3* gene, also known as *PUMA*, is located on human chromosome 19q13.3-q13.4 and is homologous with

a *BCL2* family member. *BBC3* has a distinguished function in regulating other genes [38]. Many tumor genes are correlated with *BBC3*. The biological role for *BBC3* is to induce apoptosis via the mitochondrial apoptotic pathway. Furthermore, *BBC3* is also transcriptionally activated by the tumor suppressor *p53*, which is a key regulator of apoptosis and tumor genesis in breast cancer [22].

MUC1 gene encodes a highly glycosylated protein located on the apical surface of mammary epithelia that is aberrantly overexpressed in approximately 90% of human breast cancers [39]. However, its role in cancer metastasis is yet less well understood. *MUC1* protein overexpression has been associated with cell adhesion inhibition as well as increased metastatic and invasive potential of tumor cells. This overexpression allows *MUC1* to interact with members of the *ERBB* family of receptor tyrosine kinases [40].

In the proposed netSAM procedure, a series of statistical methods and techniques were employed. Despite the difference in methodology, our analysis confirmed some of previous findings. For example, we also found the correlation of *ERBB2* and *MUC1* with breast cancer prognosis. Besides, when we applied traditional gene-based methods (*t*-test and lasso) to the gene expression datasets, we found that only a small part of the known tumor genes was identified as breast cancer-related genes.

In conclusion, oncogenes found by netSAM can be used to stratify patients for treatment of the disease as well as extend perception to the disease mechanism for breast cancer, supply potential information in clinical decision-making, and help to reduce costs of therapy. However, these genes could not yet be fully justified with the current clinical knowledge, and further experimental validation is urgent. Differential genetic interaction networks have been proved very powerful for mapping the pathways that modulate/mediate essential cell functions. Our work demonstrated that a differential network-based inference method can provide a powerful tool for identifying associated genes in human disease.

Future work includes exploring other procedures for further improving accuracy and efficiency of detection, for example, using protein interaction network information. It is also believed that the incorporation of additional biological data and information would acquire better biomarkers for disease gene discovery.

Appendix

See, Algorithm 1.

Acknowledgments

The authors would like to acknowledge the help of the editors and reviewers for many constructive comments and useful suggestions that greatly helped to improve the paper. This research was supported in part by grants from the National Natural Science Foundation of China (Grant nos. 61070137, 60371044, and 60933009) and the Science and Technology Research Development Program in Shaanxi province of China (Program no. 2009 K01-56).

References

[1] M. J. Van De Vijver, Y. D. He, L. J. van 't Veer et al., "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[2] C. Sotiriou, S.-Y. Neo, L. M. McShane et al., "Breast cancer classification and prognosis based on gene expression profiles from a population-based study," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 18, pp. 10393–10398, 2003.

[3] G. Alexe, S. Alexe, D. E. Axelrod et al., "Breast cancer prognosis by combinatorial analysis of gene expression data," *Breast Cancer Research*, vol. 8, no. 4, article R41, 2006.

[4] C. Sotiriou and M. J. Piccart, "Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?" *Nature Reviews Cancer*, vol. 7, no. 7, pp. 545–553, 2007.

[5] C. Desmedt, F. Piette, S. Loi et al., "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series," *Clinical Cancer Research*, vol. 13, no. 11, pp. 3207–3214, 2007.

[6] A. Subramanian, H. Kuehn, J. Gould, P. Tamayo, and J. P. Mesirov, "GSEA-P: a desktop application for gene set enrichment analysis," *Bioinformatics*, vol. 23, no. 23, pp. 3251–3253, 2007.

[7] M. Ashburner, C. A. Ball, J.A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[8] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.

[9] C. Sotiriou and L. Pusztai, "Gene-expression signatures in breast cancer," *New England Journal of Medicine*, vol. 360, no. 8, pp. 752–800, 2009.

[10] K. Wu, L. House, W. Liu, and W. C. S. Cho, "Personalized targeted therapy for lung cancer," *International Journal of Molecular Sciences*, vol. 13, no. 9, pp. 11471–11496, 2012.

[11] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, 2007.

[12] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Molecular Systems Biology*, vol. 4, article 189, 2008.

[13] H. Fröhlich, "Network based consensus gene signatures for biomarker discovery in breast cancer," *PloS ONE*, vol. 6, no. 10, Article ID e25364, 2011.

[14] L. Chen, J. Xuan, R. B. Riggins, R. Clarke, and Y. Wang, "Identifying cancer biomarkers by network-constrained support vector machines," *BMC Systems Biology*, vol. 5, article 161, 2011.

[15] T. Ideker and N. J. Krogan, "Differential network biology," *Molecular Systems Biology*, vol. 8, article 565, 2012.

[16] B. Valcárcel, P. Würtz, N. K. Seich al Basatena et al., "A differential network approach to exploring differences between biological states: an application to prediabetes," *PLoS ONE*, vol. 6, no. 9, Article ID e24702, 2011.

[17] G. Gambardella, M. N. Moretti, R. de Cegli et al., "Differential network analysis for the identification of condition-specific pathway activity and regulation," *Bioinformatics*, vol. 29, no. 14, pp. 1776–1785, 2013.

[18] O. D. Iancu, D. Oberbeck, P. Darakjian et al., "Differential network analysis reveals genetic effects on catalepsy modules," *PLoS ONE*, vol. 8, no. 3, Article ID e58951, 2013.

[19] J. West, G. Bianconi, S. Severini, and A. E. Teschendorff, "Differential network entropy reveals cancer system hallmarks," *Scientific Reports*, vol. 2, article 802, 2012.

[20] C. Y. Shen, Y. Huang, Y. Liu et al., "A modulated empirical bayes model for identifying topological and temporal estrogen receptor alpha regulatory networks in breast cancer," *BMC Systems Biology*, vol. 5, article 67, 2011.

[21] C. L. Costantino, A. K. Witkiewicz, Y. Kuwano et al., "The role of HuR in gemcitabine efficacy in pancreatic cancer: HuR up-regulates the expression of the gemcitabine metabolizing enzyme deoxycytidine kinase," *Cancer Research*, vol. 69, no. 11, pp. 4567–4572, 2009.

[22] T. G. Cotter, "Apoptosis and cancer: the genesis of a research field," *Nature Reviews Cancer*, vol. 9, no. 7, pp. 501–507, 2009.

[23] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert, "GeneRank: using search engine technology for the analysis of microarray experiments," *BMC Bioinformatics*, vol. 6, article 233, 2005.

[24] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics*, vol. 27, no. 3, pp. 431–432, 2011.

[25] J. Skog, T. Würdinger, S. van Rijn et al., "Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers," *Nature Cell Biology*, vol. 10, no. 12, pp. 1470–1476, 2008.

[26] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society B*, vol. 73, no. 3, pp. 273–282, 2011.

[27] G. Geeven, R. E. van Kesteren, A. B. Smit, and M. C. M. de Gunst, "Identification of context-specific gene regulatory networks with GEMULA—gene expression modeling using LAsso," *Bioinformatics*, vol. 28, no. 2, pp. 214–221, 2012.

[28] T. Van den Bulcke, K. Van Leemput, B. Naudts et al., "SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms," *BMC Bioinformatics*, vol. 7, article 43, 2006.

[29] Y. Wang, J. G. M. Klijn, Y. Zhang et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[30] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[31] J. McEntyre and D. Lipman, "PubMed: bridging the information gap," *Canadian Medical Association Journal*, vol. 164, no. 9, pp. 1317–1319, 2001.

[32] W. Lv and T. Yang, "Identification of possible biomarkers for breast cancer from free fatty acid profiles determined by GC-MS and multivariate statistical analysis," *Clinical Biochemistry*, vol. 45, no. 1-2, pp. 127–133, 2012.

[33] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control," *Nature Medicine*, vol. 10, no. 8, pp. 789–799, 2004.

[34] G. Dennis Jr., B. T. Sherman, D. A. Hosack et al., "DAVID: database for annotation, visualization, and integrated discovery," *Genome Biology*, vol. 4, no. 5, p. P3, 2003.

[35] S. Henikoff, "Nucleosome destabilization in the epigenetic regulation of gene expression," *Nature Reviews Genetics*, vol. 9, no. 1, pp. 15–26, 2008.

[36] A. M. Sieuwerts, M. P. Look, M. E. Meijer-Van Gelder et al., "Which cyclin E prevails as prognostic marker for breast cancer? Results from a retrospective study involving 635 lymph node-negative breast cancer patients," *Clinical Cancer Research*, vol. 12, no. 11 I, pp. 3319–3328, 2006.

[37] Z. Madjd, L. G. Durrant, R. Bradley, I. Spendlove, I. O. Ellis, and S. E. Pinder, "Loss of *CD55* is associated with aggressive breast tumors," *Clinical Cancer Research*, vol. 10, no. 8, pp. 2797–2803, 2004.

[38] L. Raj, T. Ide, A. U. Gurkar et al., "Selective killing of cancer cells by a small molecule targeting the stress response to ros," *Nature*, vol. 475, no. 7355, pp. 231–234, 2011.

[39] N. K. Ibrahim, K. O. Yariz, I. Bondarenko et al., "Randomized phase II trial of letrozole plus Anti-MUC1 antibody AS1402 in hormone receptor-positive locally advanced or metastatic breast cancer," *Clinical Cancer Research*, vol. 17, no. 21, pp. 6822–6830, 2011.

[40] D. W. Kufe, "Muc1-c oncoprotein as a target in breast cancer: activation of signaling pathways and therapeutic approaches," *Oncogene*, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 905429, 9 pages
http://dx.doi.org/10.1155/2013/905429

Research Article

SeedSeq: Off-Target Transcriptome Database

Shaoli Das,[1] Suman Ghosal,[1] Jayprokas Chakrabarti,[1,2] and Karol Kozak[3,4]

[1] *Indian Association for the Cultivation of Science, Kolkata, West Bengal 700032, India*
[2] *Gyanxet, BF 286 Salt Lake, Kolkata, West Bengal 700064, India*
[3] *Medical Faculty, Technical University of Dresden, 1307 Dresden, Germany*
[4] *Universitätsklinikum Dresden, Blasewitzer Straße 43, 01307 Dresden, Germany*

Correspondence should be addressed to Karol Kozak; karol.kozak@lmc.biol.ethz.ch

Received 21 April 2013; Revised 27 June 2013; Accepted 1 July 2013

Academic Editor: Yudong Cai

Detection of potential cross-reaction between a short oligonucleotide sequence and a longer (unintended) sequence is crucial for many biological applications, such as high content screening (HCS), microarray nucleotide probes, or short interfering RNAs (siRNAs). However, owing to a tolerance for mismatches and gaps in base-pairing with target transcripts, siRNAs could have up to hundreds of potential target sequences in a genome, and some small RNAs in mammalian systems have been shown to affect the levels of many messenger RNAs (off-targets) besides their intended target transcripts (on-targets). The reference sequence (RefSeq) collection aims to provide a comprehensive, integrated, nonredundant, well-annotated set of sequences, including mRNA transcripts. We performed a detailed off-target analysis of three most commonly used kinome siRNA libraries based on the latest RefSeq version. To simplify the access to off-target transcripts, we created a SeedSeq database, a new unique format to store off-target information.

1. Introduction

Recently, RNA interference (RNAi), a natural mechanism for gene silencing, [1, 2] has made its way as a widely used method in molecular and cell biology in both academics and industry. Pharmaceutical and biotech companies have set up libraries for large-scale screens employing thousands of short-interfering RNAs (siRNAs) or short hairpin RNA (shRNA) encoding vectors to identify new factors involved in the molecular pathways of diseases [3]. The design of RNAi reagents is the key to obtaining reliable screening results in large-scale RNAi studies. Several recent studies demonstrated that the degradation of intended transcripts by siRNA (so-called "on-targets") and unintended effects arising from inadvertent targets (so-called "off-targets") depend on the sequence of the RNAi reagent and have to be computationally analyzed [4, 5]. For knock-down/screening purposes, different companies offer sets of siRNAs targeting the whole genome (or a subset of it) for various organisms. Typically, they offer at least three different siRNAs, for each target gene. These siRNAs can either be used as single siRNAs or can be mixed and used as a pool of siRNAs. The main reason for offering several siRNAs per target is the varying knock-down efficiency of the individual oligos and the occurrence of off-target effects. In our study, we focus on sequence-dependent off-target effects that can be attributed to the binding of the siRNA to other mRNA transcripts than their target mRNA [6, 7]. Partial complementarity between siRNA and mRNA seems to be sufficient to reduce the number of silenced mRNA [6]. Based on this tolerance for mismatches and gaps in base pairing with targets, siRNAs could have up to hundreds of potential target sequences in the genome. Currently, the degree of complementarity between the two sequences needed for silencing is not well defined. Sequence dependent off-target effects are caused in many possible ways (Table 1). First of all, it has been reported that off-target effects occur with a high probability, if the siRNA shows ~90% complementarity (17 nucleotides out of 19) to an off-target gene [8–10]. However, a 21-nucleotide double-stranded RNA sharing only partial complementarity with an mRNA is still competent to cause gene silencing via translational repression [8, 11]. It seems that as few as 11 contiguous complementary nucleotides or a total of 15 are sufficient to reduce the level of mRNA transcripts [12]. The complementarity of the siRNA

Table 1: Cause for sequence-dependent off-target effects.

#	Causes for off-target effects	References
1	Nearly exact complementarity	[8, 10]
2	15 nt in total or even 11 continuous nt match	[8]
3	Seed region complementarity	[9, 12, 15]
4	miRNA function (seed region-3′′UTR conserved region complementarity)	[13]
5	Multiple occurrences of the seed region in an mRNA sequence	[13, 15]
6	Complementary region at the cleavage site, center of the siRNA	[11, 16]
7	Tolerance of G:U wobble	[16]
8	Seed complementation frequency	[12]
9	High G/C content in the seed region	[15]

seed region (the first 2–8 bases of the antisense siRNA-strand) plays a major role in the occurrences of off-target effects [9] (see Figure 1). Further analyses showed a high tolerance for mismatches outside the seed region, whereas differences within this 5′′ end of the siRNA are barely tolerated [12–14].

The center region of the siRNA is important to stabilize the siRNA-mRNA-duplex and to enhance mRNA degradation [11]. Alemán and colleagues analyzed this central region, which comprises the cleavage site of the mRNA (position 8–10 of the antisense strand; see Figure 1). They deciphered that mismatches in this region of the siRNA seem to be critical [16] and result in no cleavage. Additionally, they also tested the aspect of a G:U wobble and discovered that the G:U base pair is recognized like an authentic Watson-Crick base pair in the antisense RNA-mRNA duplex. This wobble base pairing expands the range of potential targets for a specific siRNA.

Design and validation of siRNAs are based on sequence-dependent analysis (so-called "on-target analysis"). In design process, using the sequence information, all siRNA constructs are computationally mapped onto RNA transcript sequence RefSeq-RNA using homology search algorithms. RefSeq database is a collection of taxonomically diverse, nonredundant, and richly annotated sequences representing naturally occurring molecules of DNA, RNA, and protein. Included are sequences from plasmids, organelles, viruses, archaea, bacteria, and eukaryotes. Each RefSeq is constructed wholly from sequence data submitted to the International Nucleotide Sequence Database Collaboration (INSDC). RefSeqs provide a foundation for uniting sequence data with genetic and functional information. They are generated to provide reference standards for multiple purposes ranging from genome annotation to reporting locations of sequence variation in RNAi experiments.

In order to predict off-target effects and annotate transcripts with potential off-targeting (by oligos from available siRNA libraries) information based on latest RefSeq version, a number of sequence similarity search methods or algorithms can be applied. For example, the Basic Local Alignment Search Tool (BLAST) [17] is adopted to find nearly exact homologies. Although BLAST is an excellent tool for broad sequence alignments, it falls short in its ability to accurately predict small local homologies. Other bioinformatics tools, [15, 16, 18] which do not have this shortcoming, try to predict interactions between siRNAs and mRNAs. But unfortunately,

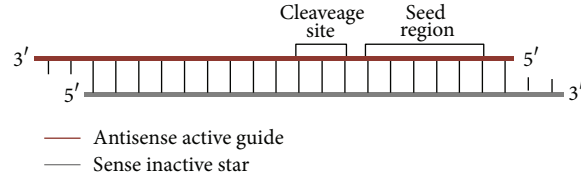


Figure 1: Structure of an siRNA: 21 bp RNA duplex with 2 nucleotides 3′′ overhanging on each strand; the two strands are called antisense or active or guide strand and sense or inactive or star strand, respectively; the first 2–8 bases of the antisense strand are called seed region, and at bases 8–10 of the antisense siRNA strand is the cleavage site.

these sequence-based prediction tools frequently do not consider specific off-target parameters like target site location, 3′′UTR conserved regions, and design specificity. Also there is no standard format which allows for distribution of off-target analysis results. Here, we describe a novel method and a database, supporting the analysis of potential off-target transcripts. We will demonstrate our approach based on three available siRNA libraries. Our analysis enabled us to determine potential off-target transcripts and to create new database format called "SeedSeq." SeedSeq aim is to provide a data source for better design, validation of siRNA libraries, and experiments. SeedSeq is available similar to RefSeq in standard gene bank [19] format file, providing easy access for bioinformatics community. SeedSeq version 1 is limited to 3 kinome libraries. Following versions will consider genome wide off-target analysis results.

2. Results

As we described above microRNA (miRNA)-mediated gene modulation has shown that complementary base pairing between the seed region and sequences in the 3′UTR of mRNA is associated with miRNA-mediated gene knockdown [20]. As siRNAs and miRNAs are believed to share some portion of the RNAi machinery, we investigated whether complementarity between the seed region of the siRNA and any region of the transcript was associated with off-targeting. To accomplish this, we predicted off-targets for 3 kinome siRNAs libraries (Ambion designed in 2006, Thermo Scientific Dharmacon designed in 2009; Qiagen designed in

Table 2: Selected siRNAs and their off-target overlap with on-targets of miRNAs from tarbase.

On-target gene symbol	siRNA antisense strand	siRNA id	Off-target gene symbol	Target of endogenous miRNA
WNK3	AAAUACUGACAAACGUGAGGC	s35278	ZEB1/TCF8	hsa-miR-200b
STRADB	AAAUACUGAUAUCCAAUGGGC	s30875	ZEB1/TCF8	hsa-miR-200b
MAPK4	UAAUGCUGAUCAACGAUCCUU	SI00606011	BACH1	hsa-miR-155
MAPK4	UAAUGCUGAUC AACGAUCCUU	SI00606011	TP53INP1	hsa-miR-155
WNK3	AAAUACUGACAAACGUGAGGC	s35278	RERE	hsa-miR-429
PLXNA3	AAAGUGCUUCCAUUGAUGAUG	s30979	Cyclin D2	hsa-miR-302b
PLXNA3	AAAGUGCUUCCAUUGAUGAUG	s30979	MBNL2	hsa-miR-302d
PLXNA3	AAAGUGCUUCCAUUGAUGAUG	s30979	VEGF	hsa-miR-372
PLXNA3	AAAGUGCUUCCAUUGAUGAUG	s30979	APP	hsa-miR-520c-3p
TYRO3	UUGGCACUAAAGGUCACCGUU	SI00288344	Mitf	hsa-miR-96
ITPKB	UUGGUCCAUAGUCUCCCUCUG	SI04435592	KCNQ1	hsa-miR-133a
TNK1	UAAUGCUCCAGGAUGCGCCAG	SI03649674	MEIS1	hsa-miR-155
PACSIN1	UUGGUCCCUCAGAUGGGCCUG	SI00127918	Pitx3	hsa-miR-133b

2008). We validated prediction of off-targets using the above described method with two rounds of validation. Firstly, we validated our predictions on the siRNA transfection data from Jackson et al. [9]. Secondly, our prediction was validated against validated target sets of miRNAs that share the same seed region sequence with the siRNAs in our dataset.

From the Jackson dataset, we used the expression data after 24 hours of transfection in all the cases, and the transcripts with a negative expression change ($P < 0.01$) was regarded as off-targets. We predicted siRNA off-targets for all the $8 + 19 = 27$ siRNA sequences in the dataset with/without filtering for conserved target sites. Without using conservation check, 65% of the predicted off-targets had a significantly negative expression change relative to the mock transfection control while using the filtering for conserved target sites marginally improved the true off-target prediction rate (66.85%). This again shows that target site conservation is not a significant criterion for siRNA mediated off-targeting. The true positive versus false positive off-target prediction by our algorithm for transcripts targeted by 19 siRNAs in Jackson dataset is presented in Figure 2.

For another validation, the validated set of miRNA targets are collected from Tarbase [21] and miRecords database [22]. We considered miRNAs from these databases which contain same seed sequence as siRNA and were trying to find an overlap between predicted off-targets of the siRNA and validated targets of miRNAs.

As a result of this validation, we found 11 siRNAs (Table 2) for which the off-target transcripts are validated targets of miRNAs with same seed sequence (4 from TarBase version 5 and 7 from miRecords), when the off-targets were predicted without conservation check. Here, we introduce a result from siRNA off-target analysis which takes into account existing kinome wide RNAi libraries using computational pipeline described above. Our analysis of three kinome-wide RNAi libraries for human revealed differences in genome coverage and off-target predicted quality. The differences most likely depend on two factors: the quality of the underlying genome release and the factors known to influence the reagent quality at the time of the library design. It has been reported that

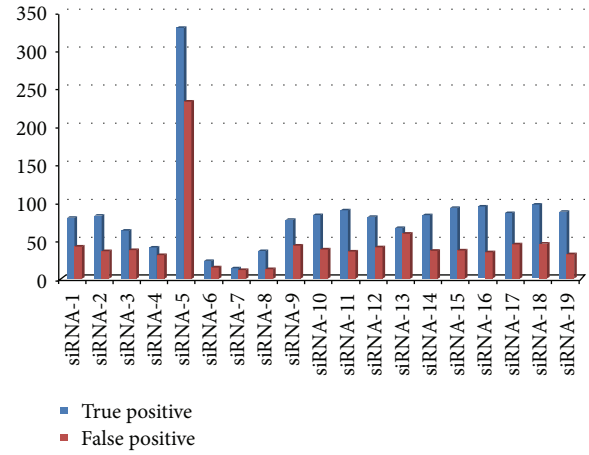


Figure 2: The number of true positive versus false positive off-target predictions by our algorithm for 19 siRNAs from the Jackson set. The off-target transcripts showing significant ($P > 0.01$) expression change after 24 hours of siRNA transfection were treated as true positives while transcripts not showing significant alteration in expression were treated as false positives.

the amount of off-targets target is different for each gene and heterogeneously distributed (Figures 3 and 4). Our report shows that the amount of off-targets ranges from 0 to 240.

We next sought to compare 3 libraries based on which transcripts are the most sensitive transcripts targeted by siRNAs from those libraries. From 3 libraries we were able to collect the most off-target sensitive transcripts (Figure 5). Very interesting phenomenon is that both transcripts NM_001030055 and NM_001173 of gene ARHGAP5 are the most sensitive off-target transcripts in both libraries, Ambion and Dharmacon, but are targeted by different siRNAs. Frequent false positives siRNAs complicate the analysis of genome-wide RNAi screens that is why it is important to identify the candidate off-targeted transcripts in primary screening data. Several transcripts can be particularly susceptible to off-target silencing [13, 15, 18]. Such "off-targeted"

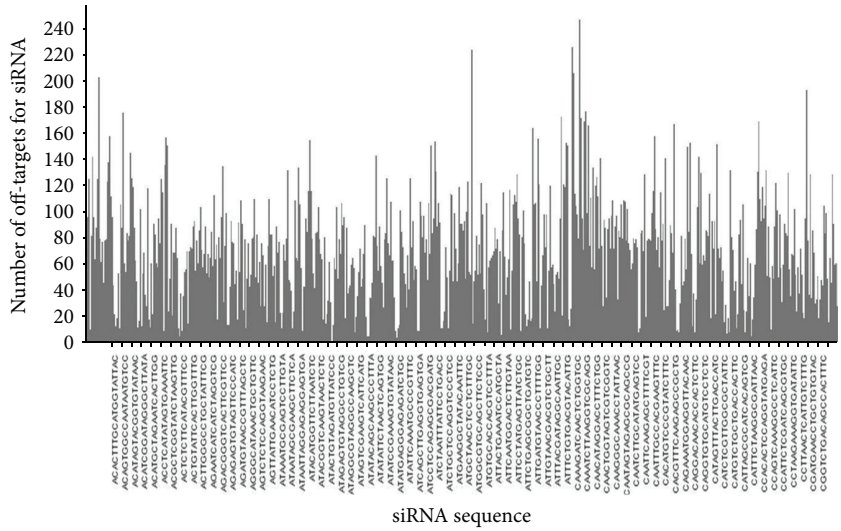


Figure 3: Distribution of siRNA off-targets across Dharmacon-Kinome library.

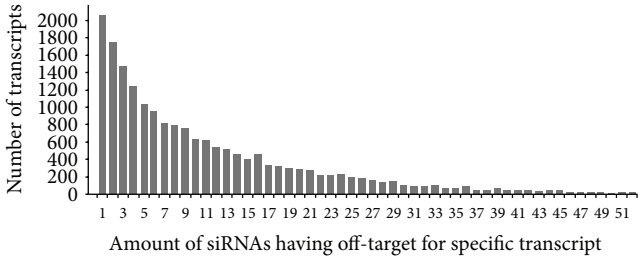


Figure 4: Distribution of number of transcripts versus number of siRNAs having those transcripts as off-targets based on analysis across 3 kinome libraries. For example transcript NM_020931 of gene KIAA1586 (Gene ID: 57691) is off-target of 7 ambion siRNAs: s224548, s14545, s29915, s3587, s8669, s25384, and s36140.

transcripts are detected after much effort has been expended to validate genes of interest.

Our next goals were (1) to compare siRNAs with maximum number of off-targets across 3 libraries and (2) to compare transcript across 3 libraries. The distribution of siRNAs off-targets per library is not flat; some siRNAs have a large number of off-targets while many contain only a few. The distribution of the efficacies of the siRNAs is shown in Figure 6. Among the 3 libraries, Qiagen siRNAs showed lowest number of off-targets. There are only 3 siRNAs from Qiagen in the list of siRNAs from all 3 libraries with highest number of off-target transcripts. Some siRNAs particularly have many off-targets. siRNA from Qiagen designed for CDKN2C is an outlier among all siRNAs having 1800 potential off-targets.

We selected the genes having the most off-targeted transcripts and made a comparison across the 3 libraries (Figure 7). Figure 8 shows that it is possible to have gene transcripts which are unintentionally targeted by around 70 siRNAs parallel in two libraries (e.g., gene USP9X has two transcripts which are off-targets of ~70 siRNAs Ambion, ~70 siRNAs Dharmacon). We also selected gene CACNA1 whose all transcripts are highly off-target sensitive.

3. Discussion

Off-target transcripts analysis results can enhance the validation rate in RNAi screens. To our knowledge, today SeedSeq is the new format and unique collection of human mRNA transcripts and their siRNA reagents off-target predictions available in GenBank format. It was designed to assist experimentalists in determining which transcripts are the sensitive for being off-target. Our analysis revealed a record of prominent off-targeted transcripts for several available siRNAs libraries. It is not clear why several transcripts behave similar to miRNA-like off-target effects. SeedSeq may be used in the siRNA design algorithms. Usage of our method as an additional analysis component of RNAi cell based screens should enable researchers to counter-screen for downregulation of sensitive transcripts and reduce the false positive siRNAs during the validation process. Detection of transcripts sensitive to off-target effects will also enable a better understanding of the rules like miRNA-like off-targeting and improve the design quality of siRNA reagents.

4. Material and Methods

4.1. SeedSeq Format. The Seed Sequence (SeedSeq) database is an open access (RNAiAtlas http://rnaiatlas.ethz.ch/), annotated and curated collection of publicly available mRNA transcripts and their siRNAs targeting them as off-target. This database is built based on off-target analysis described below, and, like RefSeq, provides only a single record for each mRNA transcript of natural biological molecule (i.e., DNA, RNA) at the current stage for 3 kinome siRNA libraries. It is a unique resource because it provides a curated sequence database linked records from siRNA to target transcripts.

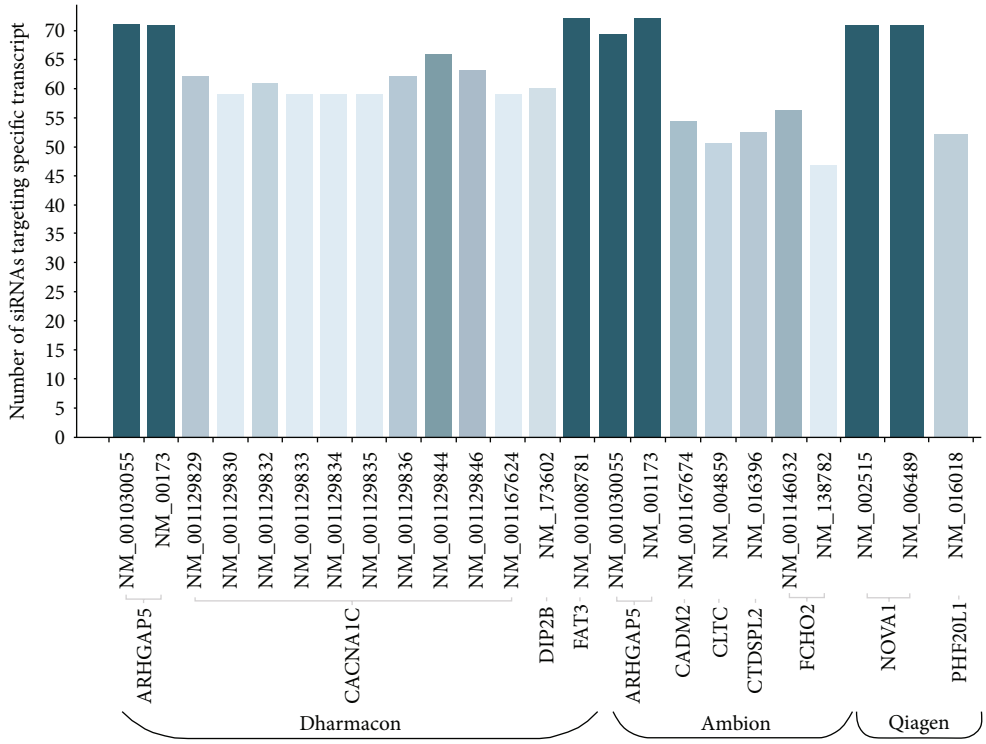


FIGURE 5: Highest number of siRNAs having off-targets for specific gene transcripts in 3 inome libraries. Transcripts of gene ARHGAP5 are the most sensitive in Ambion and Dharmacon for being an off-target transcript.

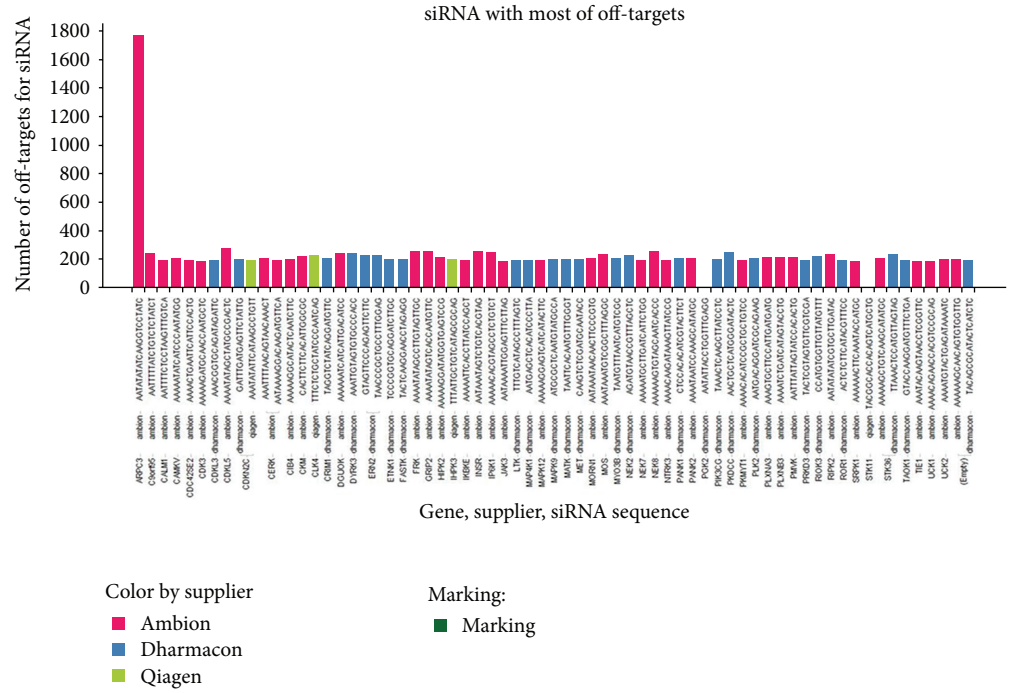


FIGURE 6: siRNAs having the highest number of off-targets across 3 libraries. siRNA designed for gene ARPC3 from Ambion is an outlier having 1800 off-targets. siRNAs designed by Dharmacon and Qiagen for gene CDKN2C indicate equal amounts of off-targets.

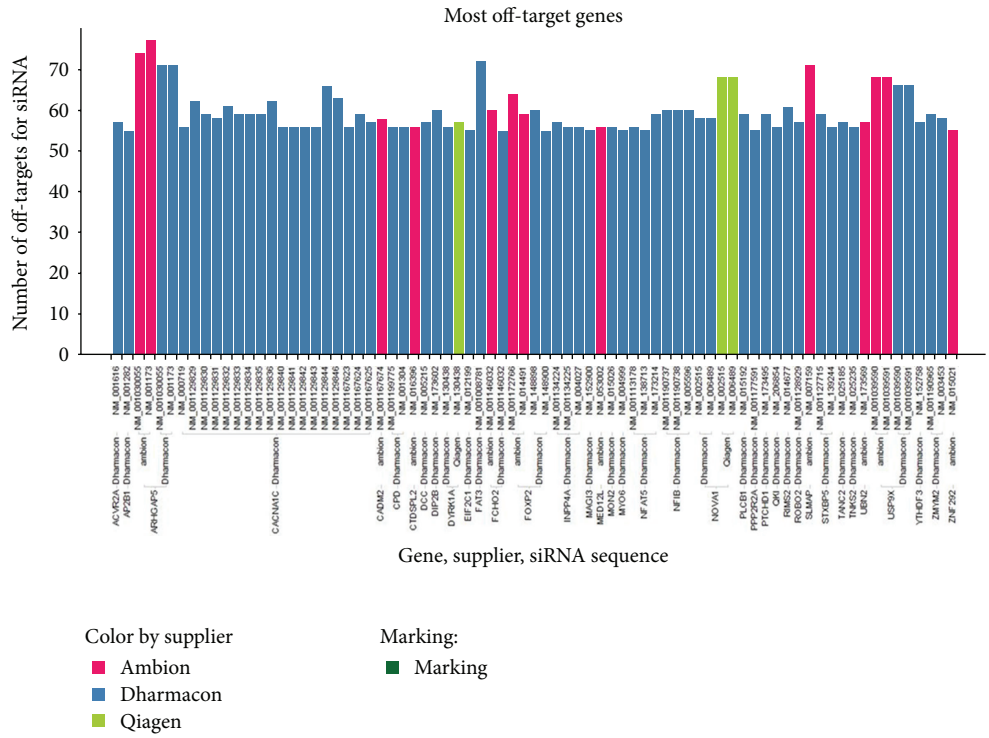


Figure 7: Transcripts distribution for being the most off-targeted genes across 3 libraries. For example transcripts USP9X are off-targets of oligos from two libraries parallely. Those transcripts indicate high sensitivity for being off-targets. All transcripts of gene CACNA1 are shown to be off-target sensitive.

For each target gene, SeedSeq aims to provide the gene transcripts and the information about siRNAs (siRNA supplier identifier, siRNAID) which potentially target those transcripts. SeedSeq is provided in GenBank format and is currently limited to kinases. Transcripts sequence records are presented in a standard format and subjected to computational validation. SeedSeq similar to RefSeq contains different transcripts categories: NM-mRNA, NR-ncRNA, XM-predicted mRNA model, and XR-predicted ncRNA model. SeedSeq is accessible via BLAST, Entrez readers, and RNAiAtlas site. SeedSeq records appear similar in format to RefSeq and GenBank records. A sequence in GenBank sequence format is a rich format for storing sequences and associated annotations. It shares a feature table vocabulary and format with the EMBL and DDJB formats. An example sequence in GenBank format is

LOCUS GXP_170357 743 bp DNA

DEFINITION loc=GXL_141619|sym=TPH2|geneid=121278|acc=GXP

_170357| taxid=9606|spec=Homo sapiens|chr=12|ctg=NC_000012|str=(+)| start=70618393|end=70619135|len=743| tss=501,632|

homgroup=4612|promset=1|descr=tryptophan hydroxylase 2|

comm=GXT_2756574/AK094614/632/gold; GXT_2799672/NM_173353/501/bronze

ACCESSION GXP_170357

BASE COUNT 216 a 180 c 147 g 200 t

ORIGIN

1 TTGATTACCT TATTTGATCA TTACACATTG TACGCTTGTG TCAAAATATC ACATGTGCCT

61 TATAAATGTG TACAACTATT AGTTATCCAT AAAAATTAAA AATTAAAAAA TCCGTAAAAT

121 GGTTTAAGCA TTCAGCAGTG CTGATCTTTC TTAAATTATT TTTCTAATTT TGGAAAGAAA

181 GCACAAAATC TTTGAATTCA CAATTGCTTA AAGACTGAGG TTAACTTGCC AGTGGCAGGC

241 TTGAGAGATG AGAGAACTAA CGTCAGAGGA TAGATGGTTT CTTGTACAAA TAACACCCCC

301 TTATGTATTG TTCTCCACCA CCCCCGCCCA AAAAGCTACT CGACCTATGA AACAAATCAC

361 ACTATGAGCA CAGATAACCC CAGGCTTCAG GTCTGTAATC TGACTGTGGC CATCGGCAAC

```
LOCUS       NM_016075               4427 bp    mRNA    linear   PRI 04-MAR-2010
DEFINITION  Homo sapiens vacuolar protein sorting 36 homolog (S. cerevisiae)
            (VPS36), mRNA.
ACCESSION   NM_016075
VERSION     NM_016075.2  GI:71051597
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
  JOURNAL   Nat. Struct. Mol. Biol. 13 (11), 1029-1030 (2006)
   PUBMED   17057716
  REMARK    GeneRIF: Crystallographic and biochemical analyses reveal that the
            GLUE domain of the human ESCRT-II EAP45 (also called VPS36) subunit
            is a split pleckstrin-homology domain that binds ubiquitin along
            one edge of the beta-sandwich.
FEATURES             Location/Qualifiers
     variation       538..559
                     /gene="VPS36"
                     /gene_synonym="C13orf9; CGI-145; DKFZp781E0871; EAP45"
                     /replace=""
                     /replace="t"
                     /db_xref="dbSNP:35567300"
     exon            563..595
                     /gene="VPS36"
                     /gene_synonym="C13orf9; CGI-145; DKFZp781E0871; EAP45"
                     /inference="alignment:Splign"
                     /number=7
     STS             4182..4302
                     /gene="VPS36"
                     /gene_synonym="C13orf9; CGI-145; DKFZp781E0871; EAP45"
                     /standard_name="RH183"
                     /db_xref="unists:2411"
     target_siRNAs   s223236,s14825,s17290,J-004732-06,J-003137-09,J-027174-05,
                     Comma separeted list of siRNAs:
                     s- Ambion siRNAs, J- Dharmacon siRNAs
```

Figure 8: SeedSeq record example.

421 CAGAAATGAG TTTCTTTCTA ATC-AGTCTTG CATCAGTCTC CAGTCATTCA TATAAAGGAG

481 CCCGGGGATG GGAGGATTCG CAT-TGCTCTT CAGCACCAGG GTTCTGGACA GCGCCCCAAG

541 CAGGCAGCTG ATCGCACGCC CCT-TCCTCTC AATCTCCGCC AGCGCTGCTA CTGCCCCTCT

601 AGTACCCCCT GCTGCAGAGA AAG-AATATTA CACCGGGATC CATGCAGCCA GCAATGATGA

661 TGTTTTCCAG TAAATACTGG GCA-CGGAGAG GGTTTTCCCT GGATTCAGCA GTGCCCGAAG

721 AGCATCAGCT ACTTGGCAGC TCA.

The format also allows for sequence names and comments to precede the sequences. Attributes in FEATURE section novel to SeedSeq records include unique information about siRNAs which potentially target this record transcript. Off-target information is described as feature annotations. This annotation is provided by off-target analysis. SeedSeq record may be an essentially unchanged, validated copy of the original RefSeq record extended with off-target features or includes additional information supplied by siRNA on-target analysis. The GenBank format originates from the GenBank software package but has now become a standard in the field of bioinformatics. The simplicity of GenBank format makes it easy to manipulate and parse sequences using text-processing tools and languages like Java, Python, Ruby, and Perl. A FEATURES section (Figure 8) is designed to be associated with a sequence and can have a location on that sequence. It is a way of describing the characteristics of a specific part of a sequence. siRNA off-target information in SeedSeq is saved in FEATURE section under tag "target_siRNAs." FEATURES are in programming languages represented by SeqFeature objects. SeqFeature objects can also have one or more annotations associated with them. SeedSeq file can be accessed similarly as GenBank, SwissProt, or EMBL file using standard GenBank format readers. An example is a SeqIOTool25 class which contains methods for reading GenBank, SwissProt, and EMBL files.

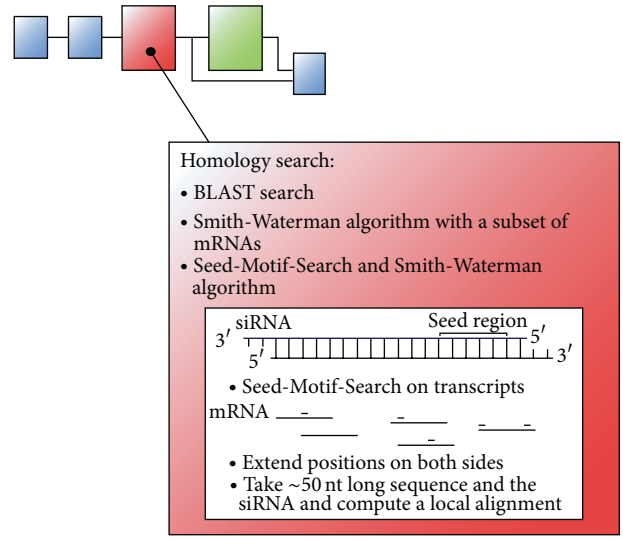


Figure 9: General structure of the concept for analysing screening results of off-target effects. Three variants of complementarity search for finding potential off-target effect depending on the type of off-targets.

4.2. Off-Target Analysis Concept. Available sequence analysis tools fail to reliably predict off-target transcripts for siRNA sequences. Building upon current understanding for the occurrence of off-target effects, a new modular analytic process is applied to create SeedSeq database. This process can be specifically adapted to a variety of options in results interpretation to identify potential off-target transcript for every siRNA of interest.

4.3. The Analytic Process. Potential off-target effects are predicted based on sequence complementarity regions between siRNAs and mRNAs. For flexibility and extensibility reasons, the process is composed of a set of steps, which must be performed in sequence to get to an effective analysis (see Figure 9).

The first step is to find homologies between siRNAs and all mRNAs available in RefSeq database. This concept contains many variants for such a complementarity search using different algorithms to perform a sequence alignment between siRNA and mRNA. A detailed description of the different complementarity search strategies is given in the next subsection.

The resulting list of a complementarity search can be too long to find the important results just by visual inspection. Therefore, the next step is to filter this list to reduce its size to meaningful results.

4.3.1. Complementarity Search. In this analysis step, it can be determined if there exists a complementary region between the selected siRNA sequences and the mRNA transcripts.

Many different sequence alignment algorithms are available to perform such a complementarity search, but they are not optimal for the purpose of this process step by default. Therefore, alternative strategy for the use of these algorithms has been developed to find nearly exact complementary regions as well as small local complementarities (see also Figure 9).

4.3.2. Smith-Waterman Algorithm. The Smith-Waterman algorithm is an accurate algorithm used to build local alignments between two sequences. Since its use with all mRNAs from the RefSeq database is not practicable, a feasible alternative is to limit the number of mRNAs to approximately 200. On a Windows 7 system (two Intel Xeon Quad-core 2.00 GHz CPUs, 16 GB RAM), the analysis of a 20-gene (4 oligonucleotides) library constructs took about 1 hour. By reducing the number of sequences, it is possible to perform a local alignment for all the siRNAs.

4.3.3. Seed-Motif-Search Combined with the Smith-Waterman Algorithm. Because of the mentioned runtime problem when performing a local alignment with the Smith-Waterman algorithm, a feasible variant to search for complementarity is introduced here. In this variant, an initial step reduces the length of the mRNA sequences to enable the use of a local alignment algorithm. This reduction is made because the seed region of the siRNA seems to play a significant role in causing off-target effects. At the beginning, all occurrences of the seed motif of every siRNA are localized in the genes. After detecting this small region, a sequence of ~50 nt around this seed motif is cut out in the mRNA. Thus, as a result of this first step, a huge number of sequences of ~50 nt in length are obtained containing the seed region of each siRNA. Due to the small length of the sequences, it is now possible to perform a local alignment with the Smith-Waterman algorithm.

4.3.4. miRNA-Like Off-Target Prediction. Prediction of miRNA like off-targets involves finding the seed complementarity of the siRNA with the 3″UTR of a nontarget mRNA. But considering only seed region complementarity identifies a large number of off-targets that could not be actually targeted. For example, for haxamer siRNA seed sequences (2–7th base from 5″ end of antisense strand), there may be thousands (or more) predicted matches in the 3′UTRs of human mRNA transcripts [10]. So, a more rational approach is needed for prediction of siRNA off-targets that needs understanding of the miRNA target recognition procedure. The Whitehead Institute siRNA selection tool uses target site conservation information for filtering the most probable off-targets. For endogenous miRNA target prediction, use of target site conservation among closely related species proved to be effective for recognition of functional target sites. For siRNA off-target prediction, restricting the seed-matched site search to the sites conserved within orthologous locations of closely related species (human, mouse, rat, and dog) greatly reduces the number of predicted off-targets and also possibly increase the chance of predicting functional target sites. For prediction of miRNA-like off-targets, we used an algorithm that combines the conventional 6-mer or 7-mer seed motif search with 3′ compensatory rule of miRNA target prediction. To check if target site conservation is a considerable factor in predicting siRNA off-targets, we searched for sites conserved among human, chimp, mouse, rat and dog.

Seed Motif Search. We used Smith-Waterman algorithm for alignment of the siRNA sense strand with the target mRNA. For the optimal alignment output, we considered the targets that have (1) perfect match with the nucleotides 2–8 (for 7-mer (m8) seed) from the 3″ end of the siRNA sense strand and (2) one mismatch in the above said seed region but perfect match with the nucleotides 13–19 from the 3″ end of the siRNA sense strand (3″ compensatory rule).

Generation of Conservation Data. Genome wide conservation data generated using multi 46-way alignment (for 46 vertebrate species) was downloaded from UCSC genome browser. Genomic regions (within human genome) conserved within human, chimp, mouse, rat, and dog are then extracted and mapped within the coordinates of human mRNAs (downloaded from UCSC genome browser) to get the location of the conserved regions within human mRNAs. Conserved regions of length 8 bases or more are only considered.

4.3.5. Dataset for Validation of Predicted Off-Targets. Dataset for off-target transcript expression change after siRNA transfection was collected from the study of Jackson et al. [7] The Jackson dataset consists of mRNA expression change data after siRNA transfection on HeLa cells for 8 different siRNAs designed for 8 different genes and 19 sequence variants of a single siRNA designed for the gene MAPK14 (GSE5814). The data included genes that displayed a significant ($P < 0.01$) difference in expression level relative to mock transfection control.

Experimentally validated miRNA targets dataset in human was collected from TarBase and miRecords. TarBase and mRecords both store manually curated collection of experimentally tested miRNA targets in human, mouse, fruitfly, worm and zebrafish. The miRNA-targets those are tested positive or negative are marked to distinguish between them. Each positive target site is described by the miRNA that binds it, the gene in which it occurs, the nature of the experiments that were conducted to test it, the sufficiency of the site to induce translational repression and/or cleavage, and the paper from which all these data were extracted.

List of Abbreviations

siRNA: Small interfering RNA
HCS: High content screening
RNAi: RNA interference
shRNA: Short hairpin RNA
INSDC: International Nucleotide Sequence Database Collaboration
BLAST: Basic Local Alignment Search Tool
miRNA: MicroRNA.

Disclosure

The SeedSeq database version 1 is available in GenBank format from RNAiAtlas (http://rnaiatlas.ethz.ch/index/seed-seq).

References

[1] X. D. Zhang, "Genome-wide screens for effective siRNAs through assessing the size of siRNA effects," *BMC Research Notes*, vol. 1, article 33, 2008.

[2] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello, "Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans," *Nature*, vol. 391, no. 6669, pp. 806–811, 1998.

[3] G. J. Hannon, *RNAi: A Guide to Gene Silencing*, Cold Spring Harbor Laboratory Press, New York, NY, USA, 2003.

[4] J. Kurreck, "RNA interference: perspectives and caveats," *Journal of RNAi and Gene Silencing*, vol. 1, pp. 50–51, 2005.

[5] L. Pelkmans, E. Fava, H. Grabner et al., "Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis," *Nature*, vol. 436, no. 7047, pp. 78–86, 2005.

[6] Y. Fedorov, A. King, E. Anderson et al., "Different delivery methods—different expression profiles," *Nature Methods*, vol. 2, article 241, 2005.

[7] A. L. Jackson, J. Burchard, D. Leake et al., "Position-specific chemical modification of siRNAs reduces "off-target" transcript silencing," *RNA*, vol. 12, no. 7, pp. 1197–1205, 2006.

[8] A. L. Jackson, S. R. Bartz, J. Schelter et al., "Expression profiling reveals off-target gene regulation by RNAi," *Nature Biotechnology*, vol. 21, no. 6, pp. 635–637, 2003.

[9] A. L. Jackson, J. Burchard, J. Schelter et al., "Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity," *RNA*, vol. 12, no. 7, pp. 1179–1187, 2006.

[10] A. Birmingham, E. M. Anderson, A. Reynolds et al., "3″ utr seed matches, but not overall identity, are associated with rnai off-targets," *Nature Methods*, vol. 3, pp. 199–204, 2006.

[11] S. Saxena, Z. O. Jónsson, and A. Dutta, "Small RNAs with imperfect match to endogenous mRNA repress translation. Implications for off-target activity of small inhibitory RNA in mammalian cells," *The Journal of Biological Chemistry*, vol. 278, no. 45, pp. 44312–44319, 2003.

[12] M. Amarzguioui, T. Holen, E. Babaie, and H. Prydz, "Tolerance for mutations and chemical modifications in a siRNA," *Nucleic Acids Research*, vol. 31, no. 2, pp. 589–595, 2003.

[13] J. G. Doench, C. P. Petersen, and P. A. Sharp, "siRNAs can function as miRNAs," *Genes and Development*, vol. 17, no. 4, pp. 438–442, 2003.

[14] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[15] X. Lin, X. Ruan, M. G. Anderson et al., "siRNA-mediated off-target gene silencing triggered by a 7 nt complementation," *Nucleic Acids Research*, vol. 33, no. 14, pp. 4527–4535, 2005.

[16] L. M. Alemán, J. Doench, and P. A. Sharp, "Comparison of siRNA-induced off-target RNA and protein effects," *RNA*, vol. 13, no. 3, pp. 385–395, 2007.

[17] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[18] N. Schultz, D. R. Marenstein, D. A. de Angelis et al., "Off-target effects dominate a large-scale RNAi screen for modulators of the TGF-β pathway and reveal microRNA regulation of TGFBR2," *Silence*, vol. 2, article 3, 2011.

[19] H. Miller, C. N. Norton, and I. N. Sarkar, "GenBank and PubMed: how connected are they?" *BMC Research Notes*, vol. 2, article 101, 2009.

[20] L. P. Lim, N. C. Lau, P. Garrett-Engele et al., "Microarray analysis shows that some microRNAs downregulate large numbers of-target mRNAs," *Nature*, vol. 433, no. 7027, pp. 769–773, 2005.

[21] T. I. Vergoulis, P. Vlachos, G. Alexiou et al., "TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support," *Nucleic Acids Research*, vol. 40, no. D1, pp. D222–D229, 2012.

[22] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li, "miRecords: an integrated resource for microRNA-target interactions," *Nucleic Acids Research*, vol. 37, no. 1, pp. D105–D110, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 263829, 7 pages
http://dx.doi.org/10.1155/2013/263829

Research Article

SubMito-PSPCP: Predicting Protein Submitochondrial Locations by Hybridizing Positional Specific Physicochemical Properties with Pseudoamino Acid Compositions

Pufeng Du[1,2] and Yuan Yu[1,2]

[1] *School of Computer Science and Technology, Tianjin University, Tianjin 300072, China*
[2] *Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin 300072, China*

Correspondence should be addressed to Pufeng Du; pufengdu@gmail.com

Received 12 May 2013; Revised 10 July 2013; Accepted 20 July 2013

Academic Editor: Lei Chen

Knowing the submitochondrial location of a mitochondrial protein is an important step in understanding its function. We developed a new method for predicting protein submitochondrial locations by introducing a new concept: positional specific physicochemical properties. With the framework of general form pseudoamino acid compositions, our method used only about 100 features to represent protein sequences, which is much simpler than the existing methods. On the dataset of SubMito, our method achieved over 93% overall accuracy, with 98.60% for inner membrane, 93.90% for matrix, and 70.70% for outer membrane, which are comparable to all state-of-the-art methods. As our method can be used as a general method to upgrade all pseudoamino-acid-composition-based methods, it should be very useful in future studies. We implement our method as an online service: SubMito-PSPCP.

1. Introduction

Mitochondrion is a type of membrane-enclosed subcellular organelle that can be found in most eukaryotic cells [1]. It is involved in many biological processes, such as energy metabolism, programmed cell death, and ionic homeostasis [2]. Every mitochondrion can be divided into four subcompartments, including inner membrane, outer membrane, intermembrane space, and the matrix. The proteins in mitochondria can vary in different tissues and organisms. For example, human mitochondria may contain about 600 different proteins [3], while over 900 proteins were found in mouse mitochondria [4]. Mitochondria have been reported to be related in several human diseases and may play an important role in the aging process [5].

Computational identification of protein subcellular locations has become a challenge in the last decade. Recently, the research in this area focused on four different topics: (1) the prediction of multisites protein subcellular localization [6–9]; (2) the prediction of protein sub-subcellular locations [10], including the prediction of protein subnuclear locations, submitochondrial locations, and subchloroplast locations; (3) the prediction of topology-specific protein subcellular locations [11, 12]; and (4) the prediction of conditional mislocated protein subcellular locations [13]. Several promising results have been achieved in these four topics. Li et al. did a serial of interesting work to predict multisites protein subcellular localization by introducing the multilabel classification methods [14–16]. Lin et al. presented a serial of impressive results in predicting protein submitochondrial and subchloroplast locations [17, 18]. They also achieved great success in applying computational approaches in identifying Golgi-resident protein types as well as mycobacterial membrane protein types [19, 20].

Over the last few years, several studies focused on reporting computational methods to predict protein submitochondrial locations. Du and Li started this topic by proposing the SubMito system and the first benchmarking dataset [21]. Nanni and Lumini introduced a genetic-algorithm-based method to select sequence-based protein descriptors [22]. Shi et al. introduced the wavelet-SVM method to improve the prediction performance [23]. Fan and Li proposed a hybrid

method using six different types of descriptors with incremental diversity algorithm as a feature selection procedure [24]. Zakeri et al. employed anther hybrid method to incorporate sequence-based descriptors, functional domain descriptors, and secondary structure information [25]. Lin et al. proposed to use the overrepresented tetrapeptides to predict the protein submitochondrial locations [17]. All of these methods improved the prediction accuracy on the same benchmarking dataset [26, 27].

With the increment of prediction accuracy, the complexity of algorithms and the dimensions of the feature vector to represent the protein sequence are also increasing. Du and Li started this topic by using 1080 dimensional vectors. Nanni and Lumini created 15 artificial features by combining several hundreds of different features. Shi et al. employed the discrete wavelet transformation and summary statistics to reduce the dimensions of features. Fan and Li introduced thousands of original features and used incremental diversity algorithm to reduce them to 613 dimensions. Zakeri et al. combined over a thousand different features in their method. Lin et al. used 160,000 original features and reduced them to 1302 dimensions using a confidence parameter. Except, SubMito, all the state-of-the-art methods were using different machine-learning-based algorithms to reduce the feature dimensions. It seems that the key to improve the prediction performance is to choose the right dimension reduction algorithms.

Although the dimension reduction algorithms are consolidated based on statistics and are supported well by the underlying mathematical theories, it is usually difficult to reason the selected dimensions in a biological sense. We admit that the dimension reduction algorithms are effective and useful. It should be regarded as a powerful tool to improve the prediction performance of bioinformatics predictors. However, in this paper, we would present a method that can produce comparable prediction performance with only about 100 dimensions of features and without using any dimensional reduction algorithm.

2. Materials and Methods

2.1. Datasets. There are several datasets existing for predicting submitochondrial locations. These datasets are always extracted from UniProt database with several filtering procedures. Since the methods, which were proposed along with these datasets, may have different requirement to the dataset, there are differences in the filtering procedures. In order to reflect the most recent advances in the available data as well as demonstrating the prediction power of the current method, two datasets were adopted in the current study. One dataset was directly extracted from the most recent version of UniProt database, and the other is the SubMito dataset that was published by Du and Li.

The procedures for filtering the raw data from UniProt database are described as follows: First, the reviewed sequences in the UniProt database, which are annotated with subcellular location "mitochondrion," were retrieved using the UniProt online query and retrieval system. Secondly, the sequences were screened to ensure every sequence has a uniquely annotated submitochondrial location among the four locations: mitochondrial inner membrane, mitochondrial outer membrane, mitochondrial matrix, and mitochondrial intermembrane space. Due to the limited number of multi-sites submitochondrial proteins, we do not consider them in the current study. Thirdly, the sequences which are fragment of other proteins are excluded. The remaining sequences are processed using the CD-HIT program to remove the highly homologues sequences. The identity cutoff was set to 40% in the CD-HIT program. Finally, the submitochondrial locations, which contain less than 15 sequences, were discarded. The remaining 983 sequences compose the dataset of this study. Among the 983 sequences, there are 661 sequences from inner membrane, 177 sequences from matrix, and 145 sequences from outer membrane. We use this dataset as the basis to train and test our method. This dataset was denoted as the SML3-983 dataset in the current study.

Table 1: Summary of the dataset.

Submitochondrial locations	Number of proteins	
	SML3-317	SML3-983
Inner membrane	131	661
Outer membrane	41	145
Matrix	145	177
Total	317	983

The dataset of SubMito was also adopted as the basis for comparing the performance of our method to other existing methods, as all existing methods reported jackknife test performance on this dataset. The SubMito dataset contains 317 protein sequences from 3 submitochondrial locations, including 131 sequences from inner membrane, 41 sequences from outer membrane and 145 sequences from matrix. The pairwise sequence similarity in the dataset is lower than 40%. This dataset was denoted by the SML3-317 dataset in the current study. The summary of both datasets is shown in Table 1.

2.2. Sequence Representations. In order to improve the performance in predicting protein subcellular localizations, one of the keys is to represent the protein sequences with an effective discrete numerical form, which is able to reflect the intrinsic correlation with their localizations [28]. The PseAACs (pseudoamino acid compositions) have been commonly used to represent protein sequences in predicting their subcellular locations [29]. It is also extended recently to represent nucleotide sequences as well [30]. The basic idea of the PseAAC is to extract the sequence order information with the autocorrelation coefficients of the protein sequence if every residue on the protein sequence can be represented with a number [31]. The physicochemical properties of amino acids, like hydrophobicity and hydrophilicity values, were used for this purpose [32].

Biology is a natural science with historical dimensions. In the evolution history, the mutations in DNA level may produce the changes of single residues or insertion or deletion of several residues on the protein sequences. However, the function and the localization of the protein may remain unchanged. Therefore, we should investigate a group of evolutionary related protein sequences rather than a single

protein sequence, which will make it easy to determine which residues are relatively more important in preserving the function and the localization of the protein. In recent years, the PsePSSM (pseudopositional specific scoring matrix), which applies the pseudoamino acid composition concept on the PSSM (positional specific scoring matrix), was widely applied in representing protein sequences [33–36].

Next, we propose a method that replaces the physicochemical properties in the PseAAC with the PSPCP (positional specific physicochemical properties), which can be derived from the PSSM and the existing physicochemical properties.

Let $P = R_1R_2 \cdots R_L$ be a protein sequence with length L, where R_1,R_2,. . .,R_L are the L residues on the protein sequence. By searching P against the SwissProt database using PSI-BLAST program [37] with three iterations and 0.001 as the e-value threshold, a PSSM can be produced as follows:

$$E(P) = \begin{bmatrix} E_{1\to 1} & E_{1\to 2} & \cdots & E_{1\to 20} \\ E_{2\to 1} & E_{2\to 2} & \cdots & E_{2\to 20} \\ \vdots & \vdots & \ddots & \vdots \\ E_{L\to 1} & E_{L\to 1} & \cdots & E_{L\to 20} \end{bmatrix}, \quad (1)$$

where $E_{i\to j}$ is a score generated by the PSI-BLAST. This score described the propensity of the ith residue on the protein sequence that is being changed to the jth type of amino acid during the evolutionary process.

Because of the PSSM generation process in PSI-BLAST, this number can be either positive or negative. It can also vary in a large range. In order to make every element in (1) within the range [0, 1], a conversion was performed to create a standardized matrix as follows:

$$A(P) = \begin{bmatrix} A_{1\to 1} & A_{1\to 2} & \cdots & A_{1\to 20} \\ A_{2\to 1} & A_{2\to 2} & \cdots & A_{2\to 20} \\ \vdots & \vdots & \ddots & \vdots \\ A_{L\to 1} & A_{L\to 1} & \cdots & A_{L\to 20} \end{bmatrix}, \quad (2)$$

where

$$A_{i\to j} = \frac{\exp\left(E_{i\to j}\right)}{\sum_{j=1}^{20}\exp\left(E_{i\to j}\right)}, \quad i = 1, 2, \ldots, L;\ j = 1, 2, \ldots, 20. \quad (3)$$

Let $H(r, j)$ be the rth physicochemical property of the jth type of residue. We now use the rth physicochemical property to derive a PSPCP for R_i on the protein sequence P, as given by

$$d_{i,r}(P) = \sum_{j=1}^{20} A_{i\to j} h(r, j), \quad (4)$$

where $d_{i,r}(P)$ is the PSPCP derived from the rth physicochemical property for R_i and $h(r, j)$ is the normalized rth physicochemical property of the jth type of residues. It can be computed as follows:

$$h(r, j) = \frac{H(r, j) - m(r)}{s(r)}, \quad (5)$$

where

$$m(r) = \frac{1}{20}\sum_{j=1}^{20} H(r, j),$$
$$s(r) = \sqrt{\frac{1}{20}\sum_{j=1}^{20}\left(H(r, j) - m(r)\right)^2}. \quad (6)$$

We now use the PSPCP to replace the physicochemical properties in the amphiphilic pseudoamino acid compositions (AmPseAAC) [31]. We compute the following twenty descriptors to replace the amino acid compositions in the AmPseAAC:

$$f_j(P) = \frac{1}{L}\sum_{i=1}^{L} A_{i\to j}, \quad j = 1, 2, \ldots, 20. \quad (7)$$

The pseudofactor that describes the kth tier sequence-order effect with the PSPCP, which is derived from the rth physiochemical property, can be formulated as (8)

$$u_{k,r}(P) = \frac{1}{L-k}\sum_{i=1}^{L-k} d_{i,r}(P)\, d_{i+k,r}(P). \quad (8)$$

Given the parameters, w and λ, and R types of physicochemical properties, we create $20 + \lambda R$ descriptors for protein P as follows:

$$q_n(P) = \begin{cases} \dfrac{f_n(P)}{\sum_{n=1}^{20} f_n(P) + w\sum_{r=1}^{R}\sum_{k=1}^{\lambda} u_{k,r}(P)}, \\ \qquad 1 \le n \le 20, \\ \dfrac{w u_{k,r}(P)}{\sum_{n=1}^{20} f_n(P) + w\sum_{r=1}^{R}\sum_{k=1}^{\lambda} u_{k,r}(P)}, \\ \qquad n = 20 + (r-1)\lambda + k, \\ \qquad 1 \le k \le \lambda,\ 1 \le r \le R, \end{cases} \quad (9)$$

where w should be in the range $(0, 1)$ and λ can be a positive integer less than the length of the shortest sequence in the benchmarking dataset.

The protein P can be represented as a $20 + \lambda R$ dimension vector as

$$\mathbf{Q}(P) = \left[q_1(P), q_2(P), \ldots, q_{20+\lambda R}(P)\right]^T. \quad (10)$$

When the PSSM is not available, $A_{i\to R_i} = 1$ would be assumed. The whole sequence representation would automatically degrade to AmPseAAC.

2.3. Prediction Algorithm. We use SVM (support vector machine) as the prediction algorithm in this study. It searches for an optimal separating hyperplane, which maximizes the margin in feature space [38]. We used an RBF (radial basis function) kernel in this study, as the RBF kernel is the most flexible and the most widely used kernel function. The RBF kernel function can be formulated as follows:

$$K\left(\mathbf{Q}(P_x), \mathbf{Q}(P_y)\right) = \exp\left(-\gamma\left|\mathbf{Q}(P_x) - \mathbf{Q}(P_y)\right|^2\right), \quad (11)$$

where γ is a parameter, $\mathbf{Q}(P_x)$ and $\mathbf{Q}(P_y)$ are $20 + \lambda R$ dimension vectors representing proteins P_x and P_y, and "$|\cdot|$" is the operator that computes the Euclidean length of a vector.

2.4. Performance Evaluations. The jackknife test, which is deemed to be the most objective and rigorous protocol for evaluating predictive bioinformatics methods, was applied in evaluating the performance of our method [39]. The following summary statistics were used to measure the prediction performance:

$$\mathrm{Acc}_s = \frac{\mathrm{TP}_s}{\mathrm{TP}_s + \mathrm{FN}_s}, \quad s = 1, 2, 3,$$

$$\mathrm{MCC}_s = \frac{\mathrm{TP}_s\mathrm{TN}_s - \mathrm{FP}_s\mathrm{FN}_s}{\sqrt{(\mathrm{TP}_s + \mathrm{FP}_s)(\mathrm{TP}_s + \mathrm{FN}_s)(\mathrm{TN}_s + \mathrm{FP}_s)(\mathrm{TN}_s + \mathrm{FN}_s)}}, \quad s = 1, 2, 3,$$

$$\mathrm{ACC} = \frac{\sum_{s=1}^{3} \mathrm{TP}_s}{\sum_{s=1}^{3} \mathrm{TP}_s + \mathrm{FN}_s}, \tag{12}$$

where Acc_s is the prediction accuracy for the sth location, MCC_s is the Mathew's correlation coefficient [40] for the sth location, ACC is the overall prediction accuracy, and TP_s, TN_s, FP_s, and FN_s are the numbers of true positives, true negatives, false positives, and false negatives of the sth location in the jackknife test, respectively.

2.5. Parameter Calibrations. There are several parameters in our method. The value of these parameters will affect the prediction performance of our method. These parameters were calibrated to optimize the jackknife test overall accuracy. Nine different types of physicochemical properties, which are the same as the SubMito method, were applied in this method. These physicochemical properties can be found in Table 2. The parameters w and λ were selected by enumerations. The parameter w was enumerated in the range 0.05 to 1.0 with step 0.05. The parameter λ was enumerated in the range 2 to 20 with step 1. Altogether, 380 combinations of w and λ were tested. For every combination, a grid search was carried out using LIBSVM software package [41] to optimize the jackknife test performance by finding the best values of the parameters γ and C, which are the cost parameters in training SVM models.

3. Results and Discussions

3.1. Prediction Performance. The jackknife test on SML3-983 dataset was shown in Table 3. The optimal performance was achieved when $w = 0.15$, $\lambda = 11$, $\gamma = 0.125$, and $C = 8$. The optimal jackknife test performance on SML3-983 was 89.01%.

Since all existing methods reported their jackknife test performance on SML3-317 dataset, we also optimized our method on that dataset for a performance comparison. On SML3-317 dataset, we achieved the best performance when $w = 0.15$, $\lambda = 9$, $\gamma = 0.125$, and $C = 2$. The optimal performance of our method on SML3-317 was listed in Table 4 with the comparison to the other existing methods.

Table 2: Physicochemical properties used in this method.

AAIndex ID	Property description
BULH740101	Transfer free energy to surface
EISD840101	Consensus normalized hydrophobicity
HOPT810101	Hydrophilicity value
RADA880108	Mean polarity
ZIMJ680104	Isoelectric point
MCMT640101	Refractivity
BHAR880101	Average flexibility indices
CHOC750101	Average volume of buried residue
COSI940101	Electron-ion interaction potential values

Table 3: Prediction performance on SML3-983 dataset.

Submitochondrial location	ACC	MCC
Inner membrane	95.46%	0.77
Outer membrane	77.93%	0.83
Matrix	74.01%	0.73
Overall	89.01%	

On SML3-317 dataset, the overall prediction accuracy of our method achieved over 93%, which is comparable to all state-of-the-art methods. Obviously, some other methods have achieved about 1% higher overall accuracy than our method. Nevertheless, no existing method achieved better prediction accuracy on all three submitochondrial locations. It should also be noticed that our method achieved 98% accuracy on the inner membrane class, which is higher than SubIdent, MitoLoc, and Fan and Li's method. The only method that has higher prediction accuracy on the inner membrane class is the TetraMito. However, TetraMito has a lower MCC value on the inner membrane class, which indicates that the 100% accuracy on the inner membrane class may be on the cost of decreasing accuracy of the other locations. As anticipated, TetraMito has only 66% prediction accuracy on the outer membrane class with a similar MCC value to our method. The only drawback of our method is the performance on matrix. The prediction accuracy is slightly lower than existing methods. However, the MCC on matrix location is still higher than most of the existing methods. Therefore, it is fair to say that our method is comparable to all state-of-the-art methods in predicting protein submitochondrial locations.

To further validate the performance of our method, we carried out an independent dataset test. For both SML3-983 and SML3-317 datasets, 80% sequences were randomly selected as the training dataset. The predictor was trained with these 80% sequences. The prediction performance was estimated using the remaining 20% sequences. These procedures were repeated 20 times for every dataset. The average prediction performance and the standard deviation of the accuracy were shown in Table 5. The independent dataset test performance is similar to the jackknife test performance. These results proved that the performance of our method was not overestimated.

Table 4: Performance comparison on SML3-317 dataset.

Methods	Inner membrane		Matrix		Outer membrane		Overall
	ACC	MCC	ACC	MCC	ACC	MCC	
SubMito [21]	85.50%	0.79	94.50%	0.77	51.20%	0.64	85.20%
GPLoc [22]	83.20%	0.80	97.20%	0.85	78.10%	0.77	89.00%
SubIdent [23]	91.60%	0.86	97.30%	0.79	82.90%	0.88	93.10%
Predict_SubMito [26]	91.80%	0.79	96.40%	0.79	66.10%	0.63	89.70%
MitoLoc [25]	97.70%	0.94	99.00%	0.93	68.30%	0.81	94.70%
Fan and Li [24]	94.70%	0.91	99.30%	0.96	80.50%	0.84	94.90%
TetraMito [17]	100.00%	0.90	96.60%	0.95	65.90%	0.79	94.00%
This work	98.60%	0.92	93.90%	0.89	70.70%	0.79	93.10%

Table 5: Independent dataset test of the current method.

Dataset	Average ACC	Standard deviation of ACC
SML3-317	90.24%	3.27%
SML3-983	87.17%	1.81%

The values in this table are obtained by 20 times 20% independent dataset test.

3.2. Advantages of PSPCP. In the method section, we have already described how to generate the PSPCP features. We will now discuss why we use (4) to define a replacement of physicochemical properties in the PseAAC.

The protein functions, including its subcellular locations, are largely determined by the physicochemical properties of the residues on the sequence. However, not all residues contribute to the protein functions equally. Some of the residues are important, while others are not. In the evolutionary process, the important residues tend to be conserved, or at least can only vary to limited types that possess similar physicochemical properties. But the unimportant residues would not be conserved. Thus, we can assume that all unimportant residues would have similar replacement propensity patterns in the evolutionary history. Although it is difficult to figure out which residue is important and which is not, based on our assumption, the average physicochemical properties in the evolution history would be similar for all unimportant residues. Thus, if we compute the average physicochemical properties in the evolution history, the important residues would possess physicochemical properties that are much more different to those unimportant ones. This is why we use PSPCP, which is the average physicochemical properties of all residues in the evolution history, to replace the conventional physicochemical properties in the PseAAC.

Developing novel methods for predicting protein submitochondrial locations is not only a race of prediction performance. There are many different quality terms other than prediction accuracy that can be used to describe how good a prediction method is. There are two major advantages of our method, the simpleness and the potential to improve all existing PseAAC-based methods.

The feature vectors in all state-of-the-art methods usually have several hundreds to over a thousand dimensions, which is a number much larger than the number of the samples in the benchmarking dataset. In the general concept of machine learning, a feature vector with lower dimensions is usually preferred when a similar performance can be achieved when other conditions are the same. Our method uses only about 100 dimensions feature vectors, which is lower in dimension than all existing methods except SubIdent.

Our method also has the potential to improve all existing PseAAC-based methods. Actually, the current method only replaces the physicochemical properties in the SubMito method with the PSPCP, which is derived from the same physicochemical properties in SubMito and the PSSM information. This simple replacement resulted in 8% performance improvement, which proved that the PSSM information is very useful in classifying protein sequences. Our method also gives a simple and effective way on how to integrate the PSSM information into all existing PseAAC-based methods. PsePSSM, which only extracts the information from PSSM, has achieved great success. Therefore, it can be anticipated that our method, which integrates PSSM within the PseAAC, could start a new way to utilize PSSM information more efficiently.

As pointed out by TetraMito, the GO-based methods usually achieve better performances, like Fan and Li's work. There is no doubt that GO-based methods are very useful in computationally determining protein subcellular locations. In the view of a user, today's GO-based methods require the same input as the sequence-based *ab initio* methods and provide a better result, which is very promising in practical studies. However, this cannot conceal the following fact. When a protein sequence was given to predict its locations, the performance of GO-based methods relies on whether similar sequences of the given sequence can be found in the UniProtKB database. Therefore, almost every existing GO-based method tried to incorporate some sequence-based information as its complement. Our method provides a perfect complement to the GO-based methods, as all GO-based methods, which used to incorporate PseAAC as the complement, can now be upgraded to use PSPCP within PseAAC. Actually, these methods can work side by side to help each other in a practical study.

3.3. Software Availability. We have developed an online service called SubMito-PSPCP. This service can be accessed using the following URL: http://www.pufengdu.org/srv/bioinfo/submito-pspcp/. The datasets SML3-983 and SML3-317 can both be downloaded from the "download" page of this service.

4. Conclusions

We developed a computational method that can predict the protein submitochondrial locations. We proposed the positional specific physicochemical properties concept and used this concept along with the pseudoamino acid compositions to generate protein descriptors. With only about 100 dimensions of the descriptors, we achieved comparable prediction performance to those methods using over a thousand descriptors. We hope this method can be an alternative choice in predicting protein submitochondrial locations.

Acknowledgments

This work was supported by the National Science Foundation of China (NSFC 61005041), Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP 20100032120039), Tianjin Natural Science Foundation (no. 12JCQNJC02300), China Postdoctoral Science Foundation (2012T50240 and 2013M530114), and the Seed Foundation of Tianjin University (nos. 60302006 and 60302024).

References

[1] K. Henze and W. Martin, "Evolutionary biology: essence of mitochondri," *Nature*, vol. 426, no. 6963, pp. 127–128, 2003.

[2] H. M. McBride, M. Neuspiel, and S. Wasiak, "Mitochondria: more than just a powerhouse," *Current Biology*, vol. 16, no. 14, pp. R551–R560, 2006.

[3] S. W. Taylor, E. Fahy, B. Zhang et al., "Characterization of the human heart mitochondrial proteome," *Nature Biotechnology*, vol. 21, no. 3, pp. 281–286, 2003.

[4] J. Zhang, X. Li, M. Mueller et al., "Systematic characterization of the murine mitochondrial proteome using functionally validated cardiac mitochondria," *Proteomics*, vol. 8, no. 8, pp. 1564–1575, 2008.

[5] E. J. Lesnefsky, S. Moghaddas, B. Tandler, J. Kerner, and C. L. Hoppel, "Mitochondrial dysfunction in cardiac disease: ischemia—reperfusion, aging, and heart failure," *Journal of Molecular and Cellular Cardiology*, vol. 33, no. 6, pp. 1065–1089, 2001.

[6] Z.-C. Wu, X. Xiao, and K.-C. Chou, "ILoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites," *Molecular Biosystems*, vol. 7, no. 12, pp. 3287–3297, 2011.

[7] P. Du and C. Xu, "Predicting multisite protein subcellular locations: progress and challenges," *Expert Review of Proteomics*, vol. 10, no. 3, pp. 227–237, 2013.

[8] K.-C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Molecular BioSystems*, vol. 9, no. 6, pp. 1092–1100, 2013.

[9] L. Li, Y. Zhang, L. Zou et al., "An ensemble classifier for eukaryotic protein subcellular location prediction using gene ontology categories and amino acid hydrophobicity," *PLoS ONE*, vol. 7, no. 1, Article ID e31057, 2012.

[10] P. Du, T. Li, and X. Wang, "Recent progress in predicting protein sub-subcellular locations," *Expert Review of Proteomics*, vol. 8, no. 3, pp. 391–404, 2011.

[11] A. Pierleoni, P. L. Martelli, and R. Casadio, "MemLoci: predicting subcellular localization of membrane proteins in eukaryotes," *Bioinformatics*, vol. 27, no. 9, pp. 1224–1230, 2011.

[12] P. Du, Y. Tian, and Y. Yan, "Subcellular localization prediction for human internal and organelle membrane proteins with projected gene ontology scores," *Journal of Theoretical Biology*, vol. 313, pp. 61–67, 2012.

[13] K. Lee, K. Byun, W. Hong et al., "Proteome-wide discovery of mislocated proteins in cancer," *Genome Research*, 2013.

[14] G.-Z. Li, X. Wang, X. Hu, J.-M. Liu, and R.-W. Zhao, "Multilabel learning for protein subcellular location prediction. Ieee Trans," *Nanobioscience*, vol. 11, no. 3, pp. 237–243, 2012.

[15] X. Wang, G.-Z. Li, and W.-C. Lu, "Virus-ECC-mPLoc: a multi-label predictor for predicting the subcellular localization of virus proteins with both single and multiple sites based on a general form of Chou's pseudo amino acid composition," *Protein & Peptide Letters*, vol. 20, no. 3, pp. 309–317, 2013.

[16] X. Wang and G.-Z. Li, "A multi-label predictor for identifying the subcellular locations of singleplex and multiplex eukaryotic proteins," *PLos ONE*, vol. 7, no. 5, Article ID e36317, 2012.

[17] H. Lin, W. Chen, L.-F. Yuan, Z.-Q. Li, and H. Ding, "Using over-represented tetrapeptides to predict protein submitochondria locations," *Acta Biotheoretica*, vol. 61, no. 2, pp. 259–268, 2013.

[18] H. Lin, C. Ding, L.-F. Yuan et al., "Predicting subchloroplast locations of proteins based on the general form of Chou's pseudo amino acid composition: approached from optimal tripeptide composition," *International Journal of Biomathematics*, vol. 6, no. 2, Article ID 1350003, 2013.

[19] C. Ding, L.-F. Yuan, S.-H. Guo, H. Lin, and W. Chen, "Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions," *Journal of Proteomics*, vol. 77, pp. 321–328, 2012.

[20] H. Ding, S.-H. Guo, E.-Z. Deng et al., "Prediction of Golgi-resident protein types by using feature selection technique," *Chemometrics and Intelligent Laboratory Systems*, vol. 124, pp. 9–13, 2013.

[21] P. Du and Y. Li, "Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence," *BMC Bioinformatics*, vol. 7, article 518, 2006.

[22] L. Nanni and A. Lumini, "Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization," *Amino Acids*, vol. 34, no. 4, pp. 653–660, 2008.

[23] S.-P. Shi, J.-D. Qiu, X.-Y. Sun et al., "Identify submitochondria and subchloroplast locations with pseudo amino acid composition: approach from the strategy of discrete wavelet transform feature extraction," *Biochimica et Biophysica Acta*, vol. 1813, no. 3, pp. 424–430, 2011.

[24] G.-L. Fan and Q.-Z. Li, "Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition," *Amino Acids*, vol. 43, no. 2, pp. 545–555, 2012.

[25] P. Zakeri, B. Moshiri, and M. Sadeghi, "Prediction of protein submitochondria locations based on data fusion of various features of sequences," *Journal of Theoretical Biology*, vol. 269, no. 1, pp. 208–216, 2011.

[26] Y. Zeng, Y. Guo, R. Xiao, L. Yang, L. Yu, and M. Li, "Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach," *Journal of Theoretical Biology*, vol. 259, no. 2, pp. 366–372, 2009.

[27] S. Mei, "Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization," *Journal of Theoretical Biology*, vol. 293, pp. 121–130, 2012.

[28] K.-C. Chou, "Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology," *Current Proteomics*, vol. 6, no. 4, pp. 262–274, 2009.

[29] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[30] W. Chen, P.-M. Feng, H. Lin, and K.-C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Research*, vol. 41, no. 6, article e68, 2013.

[31] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[32] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins*, vol. 43, no. 3, pp. 246–255, 2001.

[33] W.-Z. Lin, J.-A. Fang, X. Xiao, and K.-C. Chou, "iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins," *Molecular BioSystems*, vol. 9, no. 4, pp. 634–644, 2013.

[34] K.-C. Chou, Z.-C. Wu, and X. Xiao, "iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins," *PLoS ONE*, vol. 6, no. 3, Article ID e18258, 2011.

[35] K.-C. Chou, Z.-C. Wu, and X. Xiao, "ILoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Molecular BioSystems*, vol. 8, no. 2, pp. 629–641, 2012.

[36] H.-B. Shen and K.-C. Chou, "Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM," *Protein Engineering, Design and Selection*, vol. 20, no. 11, pp. 561–567, 2007.

[37] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[38] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA.

[39] H.-B. Shen, J. Yang, and K.-C. Chou, "Methodology development for predicting subcellular localization and other attributes of proteins," *Expert Review of Proteomics*, vol. 4, no. 4, pp. 453–463, 2007.

[40] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[41] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 227, no. 3, pp. 1–27, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 686090, 11 pages
http://dx.doi.org/10.1155/2013/686090

Research Article

An Approach for Identifying Cytokines Based on a Novel Ensemble Classifier

Quan Zou,[1,2,3] Zhen Wang,[1,2] Xinjun Guan,[1] Bin Liu,[3,4] Yunfeng Wu,[1] and Ziyu Lin[1,2]

[1] *School of Information Science and Technology, Xiamen University, Xiamen, Fujian, China*
[2] *Center for Cloud Computing and Big Data, Xiamen University, Xiamen, Fujian, China*
[3] *Shanghai Key Laboratory of Intelligent Information Processing, Shanghai, China*
[4] *School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China*

Correspondence should be addressed to Ziyu Lin; ziyulin@xmu.edu.cn

Received 12 May 2013; Revised 2 July 2013; Accepted 15 July 2013

Academic Editor: Lei Chen

Biology is meaningful and important to identify cytokines and investigate their various functions and biochemical mechanisms. However, several issues remain, including the large scale of benchmark datasets, serious imbalance of data, and discovery of new gene families. In this paper, we employ the machine learning approach based on a novel ensemble classifier to predict cytokines. We directly selected amino acids sequences as research objects. First, we pretreated the benchmark data accurately. Next, we analyzed the physicochemical properties and distribution of whole amino acids and then extracted a group of 120-dimensional (120D) valid features to represent sequences. Third, in the view of the serious imbalance in benchmark datasets, we utilized a sampling approach based on the synthetic minority oversampling technique algorithm and K-means clustering undersampling algorithm to rebuild the training set. Finally, we built a library for dynamic selection and circulating combination based on clustering (LibD3C) and employed the new training set to realize cytokine classification. Experiments showed that the geometric mean of sensitivity and specificity obtained through our approach is as high as 93.3%, which proves that our approach is effective for identifying cytokines.

1. Introduction

Cytokines are proteins or micromolecular polypeptides mainly secreted by immune cells. They play an important regulatory role in many cellular activities, such as growth, differentiation, and interactions between cells. Research on cytokine identification and classification has important theoretical and practical significance that may assist in the elucidation of immune regulatory mechanisms at the molecular level and contribute to disease prevention, diagnosis, and treatment. The classification and identification of proteins are of great importance in the postgenomic era. Since the 1990s, with the evolution of the human genome project, studies on biological information excavation have developed rapidly, and large numbers of protein sequences have been obtained. The scale of original bioinformatics data has grown rapidly and continues to double every ten months [1]. At present, protein classification is based mostly on their structures and functions in molecular biology [2]; thus, more information on protein classification and prediction is necessary. Cytokines are a type of proteins produced by immunocytes or related cells that regulate the functions of certain cells. They play important roles in many physiological activities. Only through accurate classification and recognition to the original sequences of cytokines can the structure and functions of unknown types of cytokines be understood. Such information will contribute to future endeavors to detect the nature of diseases at the molecular level and prevent, diagnose, and treat human diseases.

The major biological laboratories in the world have predicted the classification of all kinds of genes, protein structures, and their functions by artificial experiments. The basic method used to identify cytokines involves obtaining their sequence structures and functions by manual prediction [1], which can yield small-scale data. However, this approach is inappropriate when the data is large. Several methods

for cytokines identification have emerged over the last two decades. These methods include (1) hidden Markov model (HMM) [3, 4] and artificial neutral network (ANN) [5–7], which is based on statistical learning theory but presents significant limitations for finite sample processing; (2) Basic Local Alignment Search Tool (BLAST) [8] and FASTA [9, 10], which are approaches that utilize sequence alignments based on similarity but can only effectively identify and classify the sequences of homologous structures; (3) CTKPred, a method proposed by Huang in 2005 [11] based on support vector machine (SVM); this method extracts the dipeptide composition properties of cytokines and shows improved prediction accuracy; and (4) CytoPred, a method proposed by Lata [12] at the beginning of 2008 based on the PSI-BLAST; while this method yields favorable results, it is also unstable because it relies heavily on samples, and different samples may yield different performance.

In our approach, we selected amino acids composed of cytokines as research objects. We obtained benchmark datasets from the PFAM [13] database and deleted similar and redundant sequences. We then extracted a group of valid 120-dimensional (120D) features to represent the protein sequences of cytokines. These 120D features are the distribution features of amino acid (AA) with certain physicochemical properties [14], including hydrophobicity, normalized Van der Waals volume, polarity, polarizability, change, surface tension, secondary structure (SS), and solvent accessibility. Because the sequence numbers of positive (cytokines) and negative instances are extremely imbalanced (the number of negative instances is 84 times the number of positive instances), we utilized a sampling approach based on K-means clustering the undersampling algorithm [15] and the synthetic minority oversampling technique (SMOTE) oversampling algorithm [16]. We built a library for dynamic selection and circulating combination based on clustering (LibD3C) on the rebuilt training sets to realize cytokine classification. We achieved a success rate of 93.3%, which is higher than the result obtained using Cai's approach [17]. Cai et al. utilized 188D features of the AA composition, such as content, distribution, and bivalent frequency. The experiments prove that our approach effectively achieves cytokine identification.

Our work shows improved prediction accuracy for large-scale data and extends the prediction range of cytokine families. Compared with prior studies, we not only focused on features extraction but also extended our work to four aspects: accurate pretreatment of the benchmark data, extraction of multidimensional feature vectors [18], rebuilding training sets through the oversampling and undersampling approaches, and adoption of a novel ensemble classifier.

2. Methods

We developed several procedures to achieve cytokine identification and classification.

2.1. Data. Cytokine identification refers to the process of determining whether a protein is a cytokine or not. This classification process divides proteins into two categories, cytokines and non cytokines, which are positive and negative instances, respectively.

Due to the low number of cytokines currently available, building a representative and nonredundant negative set is very important. We chose the protein family database (PFAM [13]) based on structural information as the data source and built a negative dataset according to two principles: (1) every negative instance comes from different protein families and is the longest one in its family, and (2) negative instances from positive families cannot be selected.

We downloaded 16245 cytokines from the UniProt (Universal Protein, release 2012_09) [19–21] database website (http://www.uniprot.org/uniprot/) and obtained the family numbers of these cytokines. We removed duplicate numbers and extracted the longest cytokine sequences of their families corresponding to the non-duplicate numbers from PFAM. We obtained 126 representative cytokines as the positive set.

We then excluded positive protein families (126) from the PFAM database (10714) and obtained 10588 negative protein families. We extracted the longest sequences from the negative protein families and obtained 10588 negative instances as the negative set. Positive and negative instances constitute the original imbalanced dataset.

2.2. Features Extraction. The developmental direction of protein classification is the extraction of the characteristic properties of protein sequences and determination of the relationships between positions and structural functions in original sequence mode using appropriate mathematical tools. We extracted a group of 120 valid features to represent the protein sequence based on the distribution of AAs with certain physicochemical properties [22]. We adopted $S = R_1R_2R_3 \ldots R_L$ to represent a protein sequence, where R_i represents the amino acid in position i and L represents the sequence length, in other words, the number of amino acids. Twenty amino acids are expressed as

$$AA = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}. \quad (1)$$

2.2.1. Algorithm Based on AA Composition. The algorithm based on AA composition [23] has been previously formulated. By calculating the frequencies of 20 amino acids in the protein sequence and using these frequencies to represent a specific protein sequence, each sequence becomes a 20D vector after features conversion:

$$\left(v_1, v_2, v_3, \ldots, v_{20}\right)^T = \left(\frac{n_1}{L}, \frac{n_2}{L}, \frac{n_3}{L}, \ldots, \frac{n_{20}}{L}\right), \quad (2)$$

where n_i ($i = 1, 2, 3, \ldots, 20$) represents the quantity of an AA in the protein sequence. Obviously, $\sum_{i=1}^{20} v_i = 1$.

2.2.2. Algorithm Based on the Distribution of AAs with Certain Physicochemical Properties. The nature of AAs is determined by their side chains, and these side chains vary in shape, charge, and hydrophobicity. AAs sequences thus have different structural features and physiological functions. Based on this perspective, we employed eight physicochemical

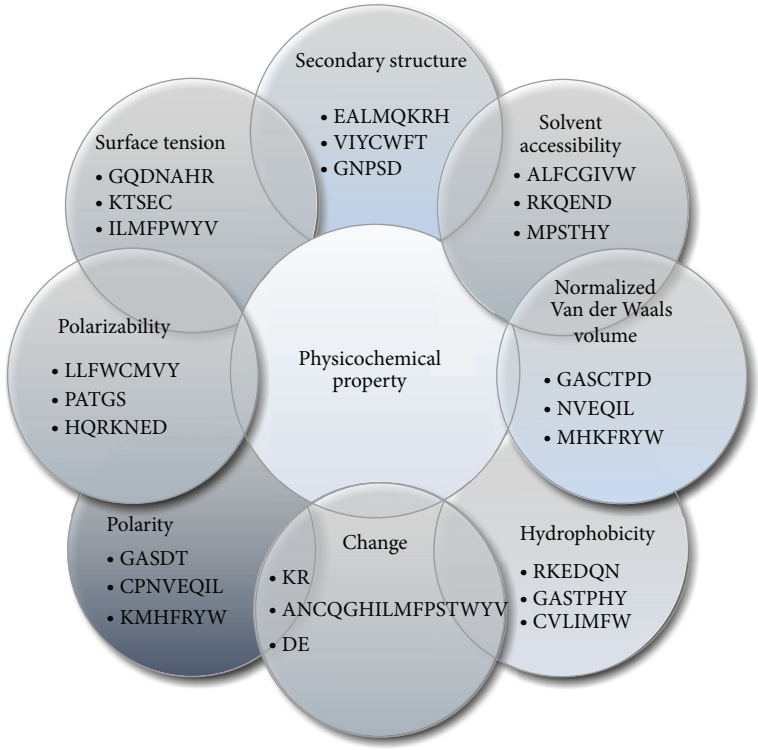


Figure 1: Division of amino acids into 3 different groups by different physicochemical properties.

properties [24–29] of AAs such as SS, solvent accessibility, normalized Van der Waals volume, hydrophobicity, change, polarizability, polarity, and surface tension. The eight physicochemical properties and the basis for their division are shown in Figure 1.

We calculated the characteristic value of the distribution of AAs with certain physicochemical properties [29] (D). Using SS [26] as an example.

To the AAs of EALMQKRH group, making the position of the first, 25%, 50%, 75%, and 100% of AAs chain represented by $p_{11}, p_{12}, \ldots, p_{15}$, respectively, and the lengths from $p_{11}, p_{12}, \ldots, p_{15}$ to the head of this protein sequence are $DSS_{11}, DSS_{12}, \ldots, DSS_{15}$, respectively. We can calculate similar parameters of two other AA SS as $DSS_{21}, DSS_{22}, \ldots, DSS_{25}, DSS_{31}, DSS_{32}, \ldots, DSS_{35}$. $V_1, V_2, \ldots, V_{15}$ can then be represented as

$$\begin{bmatrix} V_1 & V_6 & V_{11} \\ V_2 & V_7 & V_{12} \\ V_3 & V_8 & V_{13} \\ V_4 & V_9 & V_{14} \\ V_5 & V_{10} & V_{15} \end{bmatrix} = \begin{bmatrix} \frac{DSS_{11}}{L} & \frac{DSS_{21}}{L} & \frac{DSS_{31}}{L} \\ \frac{DSS_{12}}{L} & \frac{DSS_{22}}{L} & \frac{DSS_{32}}{L} \\ \frac{DSS_{13}}{L} & \frac{DSS_{23}}{L} & \frac{DSS_{33}}{L} \\ \frac{DSS_{14}}{L} & \frac{DSS_{24}}{L} & \frac{DSS_{34}}{L} \\ \frac{DSS_{15}}{L} & \frac{DSS_{25}}{L} & \frac{DSS_{35}}{L} \end{bmatrix}. \quad (3)$$

Thus, 15d feature vectors may be extracted from the SS property. We can extract 120D feature vectors after the eight physicochemical properties are analyzed. This process is presented in Figure 2.

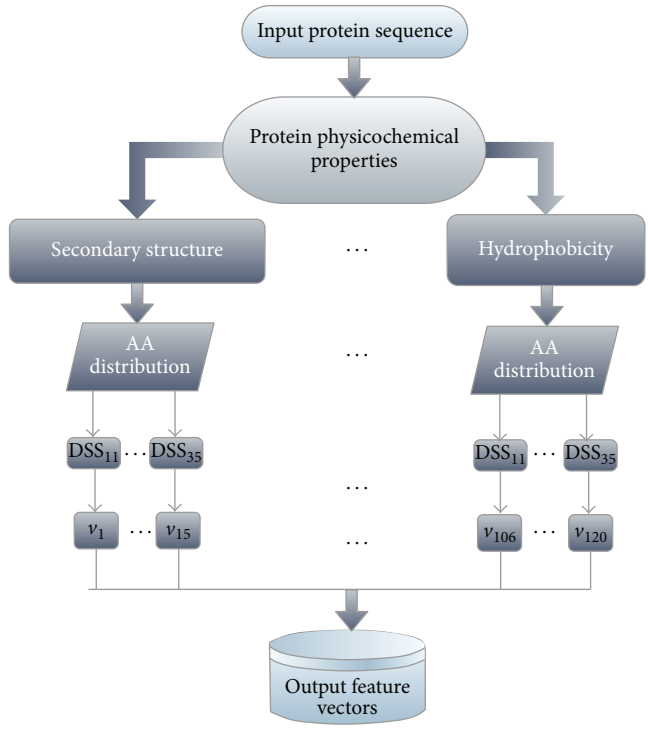


Figure 2: Extraction process of the 120-dimensional (120D) feature vectors (v).

In 2003, Cai et al. [17] established a method of features extraction based on the composition and distribution of

amino acids combined with their physicochemical properties. A total of 188D features were extracted, including the 120D features we used in this paper (3), 20D features of AA compositions (2), 24d features based on the contents of AAs with certain physicochemical properties (4), and 24d features of bivalent frequency (5) based on the eight physicochemical properties described above. We will demonstrate that the effectiveness of our 120D features is superior to that of the 188D combined features through multiple sets of experiments

$$\begin{bmatrix} \varphi_{11} & \varphi_{12} & \varphi_{13} \\ \varphi_{21} & \varphi_{22} & \varphi_{23} \\ \varphi_{31} & \varphi_{32} & \varphi_{33} \\ \varphi_{41} & \varphi_{42} & \varphi_{43} \\ \varphi_{51} & \varphi_{52} & \varphi_{53} \\ \varphi_{61} & \varphi_{62} & \varphi_{63} \\ \varphi_{71} & \varphi_{72} & \varphi_{73} \\ \varphi_{81} & \varphi_{82} & \varphi_{83} \end{bmatrix} = \begin{bmatrix} \frac{CSS_{11}}{L} & \frac{CSS_{12}}{L} & \frac{CSS_{13}}{L} \\ \frac{CSS_{21}}{L} & \frac{CSS_{22}}{L} & \frac{CSS_{23}}{L} \\ \frac{CSS_{31}}{L} & \frac{CSS_{32}}{L} & \frac{CSS_{33}}{L} \\ \frac{CSS_{41}}{L} & \frac{CSS_{42}}{L} & \frac{CSS_{43}}{L} \\ \frac{CSS_{51}}{L} & \frac{CSS_{52}}{L} & \frac{CSS_{53}}{L} \\ \frac{CSS_{61}}{L} & \frac{CSS_{62}}{L} & \frac{CSS_{63}}{L} \\ \frac{CSS_{71}}{L} & \frac{CSS_{72}}{L} & \frac{CSS_{73}}{L} \\ \frac{CSS_{81}}{L} & \frac{CSS_{82}}{L} & \frac{CSS_{83}}{L} \end{bmatrix}, \quad (4)$$

$$\begin{bmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \\ \phi_{41} & \phi_{42} & \phi_{43} \\ \phi_{51} & \phi_{52} & \phi_{53} \\ \phi_{61} & \phi_{62} & \phi_{63} \\ \phi_{71} & \phi_{72} & \phi_{73} \\ \phi_{81} & \phi_{82} & \phi_{83} \end{bmatrix} = \begin{bmatrix} \frac{BSS_{11}}{L} & \frac{BSS_{12}}{L} & \frac{BSS_{13}}{L} \\ \frac{BSS_{21}}{L} & \frac{BSS_{22}}{L} & \frac{BSS_{23}}{L} \\ \frac{BSS_{31}}{L} & \frac{BSS_{32}}{L} & \frac{BSS_{33}}{L} \\ \frac{BSS_{41}}{L} & \frac{BSS_{42}}{L} & \frac{BSS_{43}}{L} \\ \frac{BSS_{51}}{L} & \frac{BSS_{52}}{L} & \frac{BSS_{53}}{L} \\ \frac{BSS_{61}}{L} & \frac{BSS_{62}}{L} & \frac{BSS_{63}}{L} \\ \frac{BSS_{71}}{L} & \frac{BSS_{72}}{L} & \frac{BSS_{73}}{L} \\ \frac{BSS_{81}}{L} & \frac{BSS_{82}}{L} & \frac{BSS_{83}}{L} \end{bmatrix}. \quad (5)$$

2.3. Sampling. Random sampling may miss samples with strong feature prediction capability. To compensate for this shortcoming, we applied the undersampling approach using K-means clustering [15]. To avoid extremely sparse numbers of samples in the datasets by undersampling, we generated samples artificially using the SMOTE algorithm [16] to increase the size of the minimum class. The ensemble algorithm of undersampling combined with oversampling not only avoids producing excessive noise but also solves the problem of sample shortage.

The SMOTE oversampling algorithm and K-means undersampling algorithm are illustrated in Algorithms 1 and 2, respectively.

The distances between samples and clustering centroids were measured using the square of the Euclidean distance

$$d = \left\| x_p - \widehat{\mu}_i \right\|, \quad p = 1, 2, \ldots, n; \; i = 1, 2, \ldots, \lambda_n, \quad (6)$$

where x_p represents clustering samples and $\widehat{\mu}_i$ represents clustering centroids.

The process of undersampling by K-means clustering is illustrated in Figure 3.

K-means clustering is simple and rapid. Its time complexity is $O(nkt)$, and n, k, and t represent the negative sample size, initial negative cluster size, and iteration, respectively. The initial parameters directly influence the time performance of clustering, and the effective parameters significantly reduce the iterations.

To solve the problems of missing samples and introducing noise through the ensemble algorithm, we considered oversampling and undersampling to achieve balance. The ensemble algorithm is illustrated in Algorithm 3.

2.4. Ensemble Classifier. Ensemble classification is a method used to combine various basic classifiers that each has independent decision-making ability. Generally speaking, the prediction ability of an ensemble classifier is superior to that of a single classifier because the former can address the diversities produced by the latter more efficiently when faced with different problems [30]. According to the principle that the effect of the ensemble classifier is closer to the globally optimal solution than that of the single classifier, we further improved the prediction accuracy of our proposed technique by increasing the diversity of basic classifiers.

We adopted the K-means algorithm [31] to cluster all classification results of basic classifiers, and the diversity of basic classifiers selected from each category was further improved. Classifiers were selected through a circulating combined dynamic selective strategy (Circulatory Ensemble Forward Selection, CEFS), and voted for the last result. The classifier architecture is illustrated in Figure 4.

We utilized 18 basic classifiers to create the training set. The basic classifiers utilized in this study are sequential minimal optimization (SMO), support vector machine (SVM), logistic regression, Instance-based 1 (IB1), Instance-based 5 (IB5), instance-based 10 (IB10), decision table, conjunctive rule, one rule (one R), simple cart, JRip, Zero R, random tree, naïve Bayes, random forest (RF), decision stump, J48, and functional trees (FT), which are labeled as $C_1, C_2, \ldots, C_{18}$, respectively. These basic classifiers were applied to the training set independently, and the training results are represented as

$$R_{ij} = \{0, 1\}; \quad i = 1, 2, \ldots, 18; \; j = 1, 2, \ldots, m, \quad (7)$$

where m is the number of training samples.

If $R_{ij} = 0$, the sample j is classified wrongly by classifier i; otherwise, it is correct. Figure 4 shows the results matrix obtained using the K-means clustering algorithm.

We used $K = 9$ as the initial number of clustering centroids in the K-means algorithm. These centroids were divided into nine groups based on the training results of basic

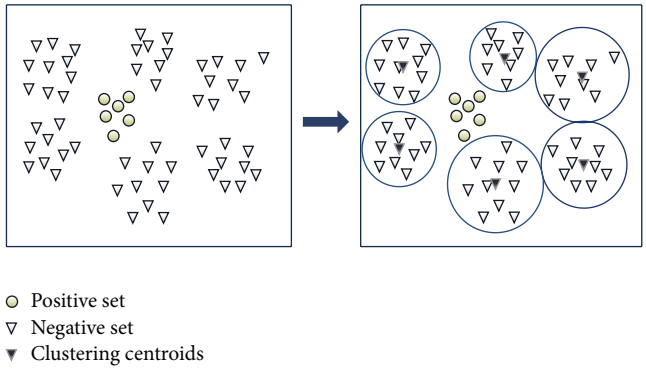


Figure 3: The process of undersampling applies K-means clustering.

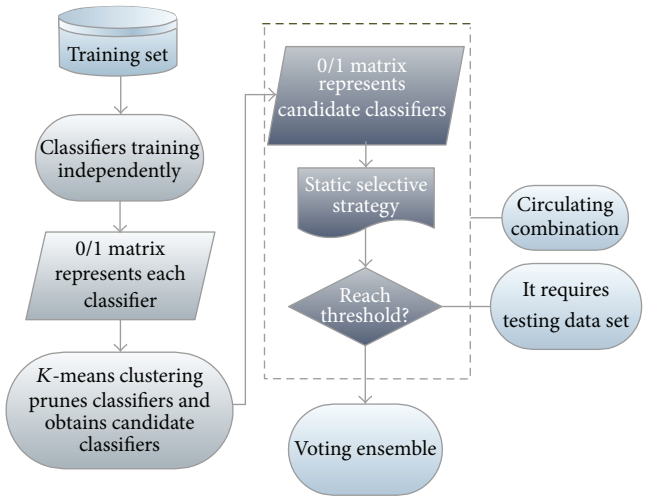


Figure 4: Classifier architecture.

classifiers. The basic classifiers with the best performance in each cluster were sorted in descending order according to their classification accuracy to form a set of selected classifiers.

The classifier combination was processed continuously with the circulating combination methodology to further optimize its effects. We set up a new variable CC (chosen classifier) to store the selected basic classifiers. In each cycle, the CEFS algorithm was employed to basic classifiers continuously to choose the best performing classifiers and create classifiers combination with these classifiers abiding by the vote rule. If the process results in a decline in diversity and an increase in accuracy at the same time, the classifier is added to the CC. This process is considered completed once the accuracy obtained is superior to the initial goal. The detailed algorithm description is illustrated in Figure 5.

The target accuracy, optimal accuracy, and step were initialized to 1, 0, and 0.05, respectively. The diversity was set to infinity, and the accuracy of classification and number of selected basic classifiers were set to 0.

The ensemble classifier described in this section is highly focused on the selection of basic classifiers. Through comprehensive application of various methods, we integrated

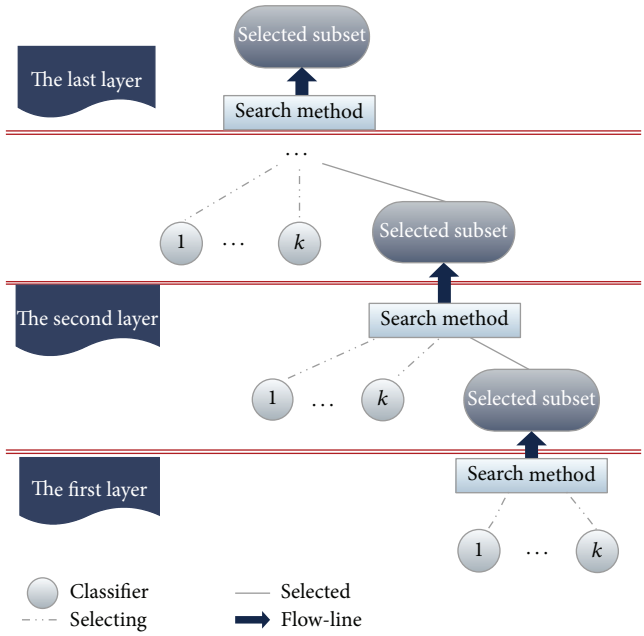


Figure 5: Circulating combination of CEFS.

the most effective basic classifiers so as to optimize the classification results.

3. Experiments

We performed a series of experiments to confirm the effectiveness of our method. First, we analyzed the effectiveness of the extracted 120D feature vectors. Second, we showed the performance of other sampling strategies and compared findings with the performance of the ensemble classifier we developed. Finally, we tested all known proteins and determined 4151 cytokines. These experiments are discussed in detail in this section.

3.1. Performance of Evaluation Standards. Sensitivity (SN) (8), specificity (SP) (9), GM (10), and overall accuracy (ACC) (11) are often used to evaluate the results of prediction or classification in bioinformatics

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{8}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{9}$$

$$\mathrm{GM} = \sqrt{\mathrm{SN} \times \mathrm{SP}}, \tag{10}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{11}$$

These four parameters are recognized as reliable measures for evaluating the performance of machine learning methods. TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

Due to the extreme imbalance of positive and negative instances in this paper, the ACC value roughly equaled the SP

(1) Input: small sample set S, over-sampling magnification N;
(2) Output: the new small sample set S', sample number k;
(3) For each small class sample x, find y nearest neighbors of the same kind with x;
(4) Choose N samples according to the magnification of over-sampling randomly, then do random linear interpolation between x and each neighbor selected;

Algorithm 1: SMOTE over-sampling.

(1) Input: positive sample set S', negative sample set B ($|S'| \ll |B|$);
(2) Output: the new negative set B' ($|B'| = |S'|$);
(3) Calculate the number of samples in two sets, k to S', n to B;
(4) Select λn (λ is defined as under-sampling ratio, $0 < \lambda < 1$) samples randomly from set B as initial clustering centroids, $\lambda = 0.2$ in our paper;
(5) Repeat;
(6) Calculate distances (Euclidean Distance) of each sample to all the clustering centroids;
(7) Choose the nearest clustering centroids and add them to certain clusters;
(8) Find the new centroids of all the new clusters;
(9) Until each cluster stability;
(10) Define the final λn centroids as B';
(11) Output B'.

Algorithm 2: Under-sampling applies K-means clustering.

(1) Input: small sample set S, big sample set B, the sample number of output sets k;
(2) Output: balanced sample set U ($U = S' + B'$, $|S'| = |B'| = k$);
(3) Extending S to a new set S' that has k samples according to SMOTE algorithm;
(4) Down sample set B to a new set B' that has k samples according to K-mean clustering;
(5) Output U.

Algorithm 3: Ensemble algorithm of under-sampling combined with over-sampling.

value (12). Hence, only SN, SP, and GM were adopted as evaluation standards in our study

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \approx \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = \mathrm{SP}. \quad (12)$$

3.2. Performance of Sampling Strategies. The test dataset consisted of 126 positive feature samples and 10588 negative feature samples; thus, it may be considered extremely imbalanced. They are extracted by 120D feature extraction algorithm in agreement with the one mentioned in Section 3.3. After directly performing 10-fold cross-validation on the test dataset without sampling by LibD3C classifier, we achieved an SP value as high as 99.9% but an SN value as low as 0.80% and a GM value of only 8.90%. The effect of that is even worse than random sampling effect. We conducted SMOTE oversampling on the positive set and K-means clustering undersampling on the negative set. The rebuilt testing set was balanced and contained 2019 positive feature samples and 1996 negative feature samples. The detailed algorithms refer to Section 2.

SN, SP, and GM values of classification results obtained from 10-fold cross-validation on the unsampled and sampled datasets are illustrated in Figure 6.

Figure 6 shows that the effect of 10-fold cross-validation on the sampled dataset is quite good. The values of SN, SP, and GM reached 96.8%, 97.7%, and 97.2%, respectively, far better than the training results of the unsampled dataset. These results provide strong evidence that oversampling and undersampling processes on the testing set are necessary.

3.3. Performance of 120D Feature Vectors. We extracted 120D feature vectors of positive and negative instances based on the distribution of AAs with certain physicochemical properties. The validity was verified by Experiments 1 and 2.

Experiment 1. The sampled dataset with 120D feature was trained, and the results of 10-fold cross-validation were analyzed. The training model was saved as model_1 by Weka (version 3.7.9). We calculated the SN, SP, and GM values of model_1 and illustrated the results in Figure 7.

Experiment 2. The imbalanced test set was tested by model_1 achieved in Experiment 1, and the SN, SP, and GM values of

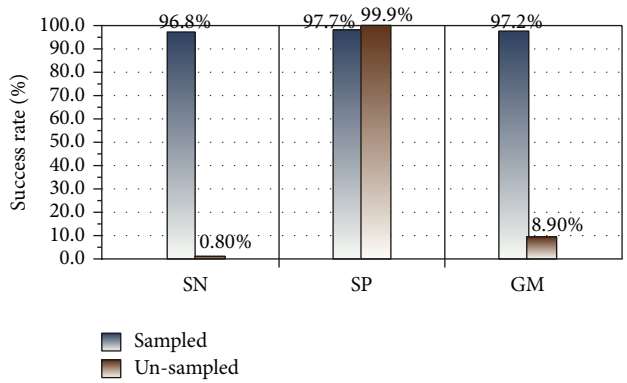


Figure 6: Comparison of validation on sampled dataset and unsampled dataset.

the test results were calculated. The findings are shown in Figure 8.

The SN, SP and GM are 96.8%, 97.7%, and 97.2%, respectively, as shown in Figure 7. In addition, testing on the imbalanced testing dataset by $model_1$ yielded favorable results, with SN, SP, and GM values of 93.7%, 92.9%, and 93.3%, respectively. These findings demonstrate that the classification works well.

To demonstrate that the performance of the 120D features we used is better than that of Cai's 188D features [17] for classifying cytokines, we conducted Experiments 3 and 4 and compared their effects. A comprehensive comparison of results illustrated the superiority of our method for cytokine identification.

Experiment 3. We used five training sets with different properties by LibD3C. These sets included 120D, 20D, 24d (content), 24d (bivalent frequency), and 188D feature vectors. The method of obtaining 20D, 24d (content), and 24d (bivalent frequency) feature vectors is used to eliminate redundant attributes from the 188D feature vectors and preserve the required attributes utilizing Weka. The results were analyzed, and the five training models were saved as $model_1$, $model_2$, $model_3$, $model_4$, and $model_5$. $Model_1$ to $model_5$ are shown in Table 1. Five groups of SN, SP, and GM values corresponding to the five training sets are shown in Figure 9. Five groups of feature vectors are detailed Section 2.

Experiment 4. We tested the imbalanced testing dataset with $model_1$, $model_2$, $model_3$, $model_4$, and $model_5$ in this order. The SN, SP, and GM values of the five testing results are shown in Figure 10.

The results show that the extraction method used in Experiments 3 and 4 is effective. The performance of the 120D feature vectors is better than that of the 188D feature vectors for the classification of cytokines. Thus, the 120D feature vectors are highly suitable for cytokine identification.

3.4. Performance of the Ensemble Classifier. To validate the classification effect of LibD3C, we conducted eight experiments (Experiments 5 to 12) using Weka (version 3.7.9).

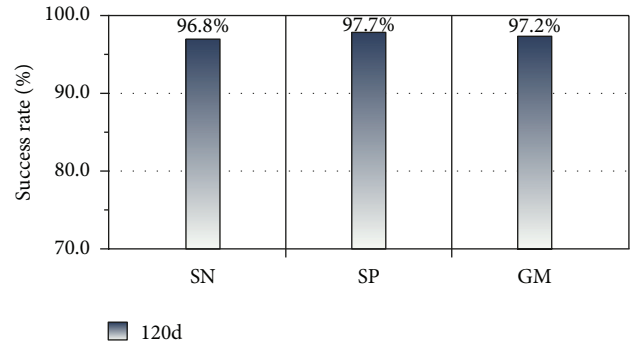


Figure 7: 10-fold cross-validation result of training set (120D features).

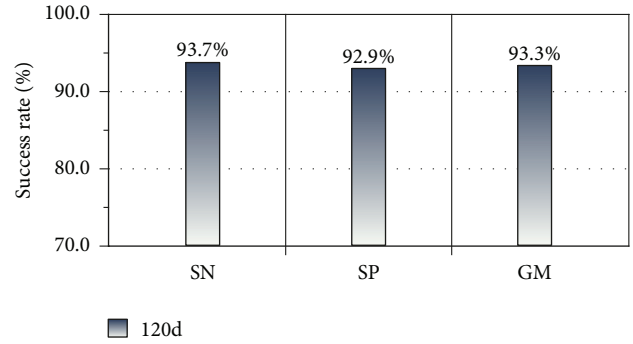


Figure 8: Testing results of original imbalanced testing set (120D features).

Experiment 5 includes the training and testing processes used in Experiments 1 and 2. The results of training and testing are shown in Figures 7 and 8, respectively.

We chose 7 simple classifiers from 18 basic classifiers from LibD3C for Experiments 6 to 12, which are similar to Experiment 5. These simple classifiers included RF, Libsvm, decision stump, SMO, naive Bayes, IB1, and J48, corresponding to Experiments 6, 7, 8, 9, 10, 11, and 12, respectively.

The training model used for Experiment 5 was $model_1$. Training models of Experiments 6 to 12 were $model_6$ to $model_{12}$, respectively, as shown in Table 2.

The 10-fold cross-validation results of ensemble classifier LibD3C and simple classifiers are shown in Figure 11.

SN, SP, and GM values of the testing results are shown in Figure 12.

Figures 11 and 12 show the optimal performance of LibD3C based on dynamic selection clustering and circulating combination. The training results of LibD3C were 96.8%, 97.7%, and 97.2%, respectively, and SN, SP, and GM values of testing results reached 93.7%, 92.9%, and 93.3%, respectively. Compared with other simple classifiers, LibD3C has very high and stable SN, SP, and GM values.

3.5. Comparison with Other Softwares. There are just few software tools or web server available on line, which can predict cytokines from protein primary sequences. We develop a web server named CytoPre (Cytokine Prediction System)

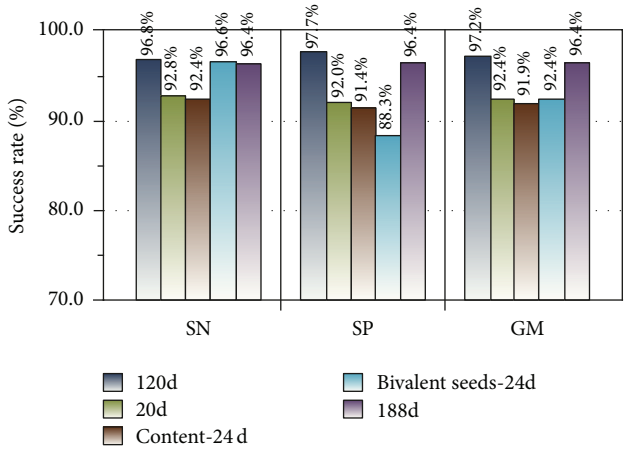


Figure 9: The comparison of 10-fold cross-validation results of five training sets.

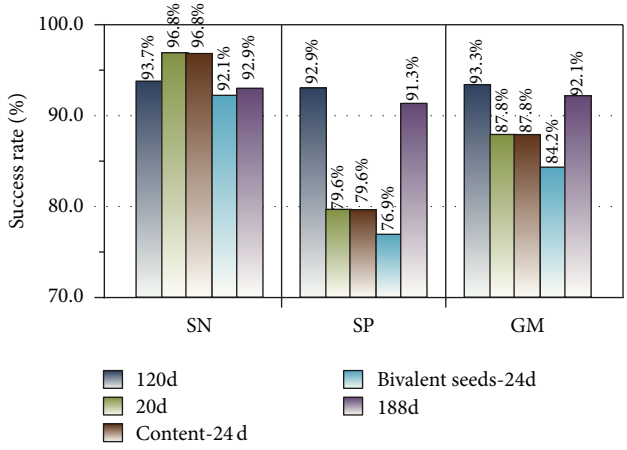


Figure 10: The comparison of testing results.

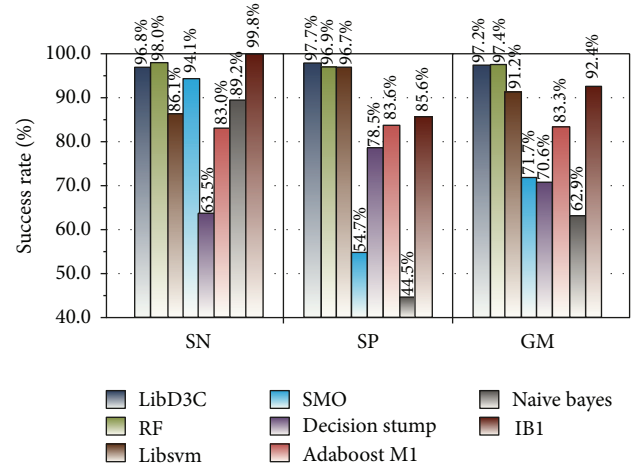


Figure 11: Performance comparison of 8 classifiers training on training set.

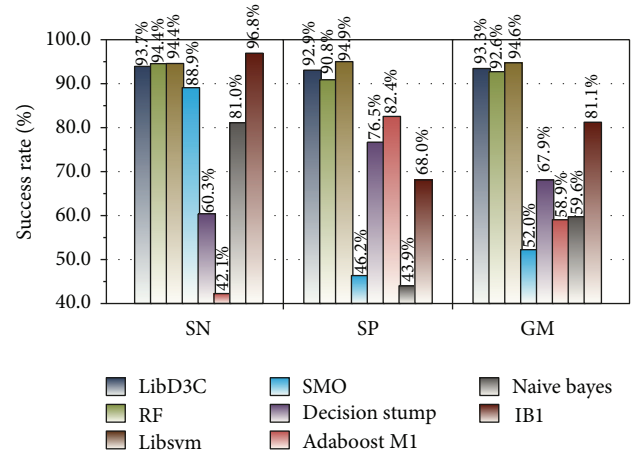


Figure 12: Performance comparison of 8 models testing on imbalanced testing set.

and compare it with CTKPred [11] and CytoKey (http://www.biomed.ecnu.edu.cn/CN/GPCR/Tools/BioAnalysistools/CytoRAP/CytoKey.aspx).

CTKPred was proposed for identifying cytokine using SVM. It extracted features from dipeptide composition and compared with Pfam searching. It was proved that CTKPred can outperform homologous searching, including HMM alignment. The sensitivity, specificity, and accuracy can get 92.5%, 97.2%, and 95.3%. CytoKey added amino acid composition and length features and gets 93.4%, 97.5%, 95.9% as sensitivity, specificity, and accuracy each.

We compared our CytoPre with CytoKey and CTKPred. Experiments showed that our system can outperforms the other two software, as shown in Figure 13, which suggested that the 188D protein composition and physical chemical properties features are more suitable for cytokine identification. Furthermore, the ensemble classifier can work better than single SVM.

3.6. Undiscovered Cytokines. We downloaded a total of 539616 protein sequences from the UniProt [19–21] database. Our goal was to predict all cytokines from whole protein sequences utilizing our training model. We detected 4151 candidate cytokine sequences (about 0.77%) from 539616 proteins. Of the 4151 candidate sequences, 39 were annotated as cytokines in UniProt. The other ones were done BLASTP to the known 16245 cytokines. Out of 4151 sequences, 444 showed regions with over 90% similarity to known cytokines, and another 697 sequences showed regions with over 50% similarity. The BLAST results and related data are supplied in the Supplementary Material (see Supplementary Material available online at http://dx.doi.org/10.1155/2013/686090) included in this paper.

Several conclusions may be made from the above experiments. First, not all of the cytokines have similar primary sequences. As well, BLAST is incapable of detecting all of the cytokines. Machine learning methods are necessary for detection. Finally, the experiments suggest that many cytokines have yet to be discovered.

3.7. Discussion about the Experiments. Our preparatory work aimed to identify positive and negative families from

Table 1: Training models of five training sets.

Features	Name of model	Save model
120d	$Model_1$	LibD3C.model
20d	$Model_2$	RF.model
24d (content)	$Model_3$	Libsvm.model
24d (bivalent frequency)	$Model_4$	SMO.model
188d	$Model_5$	J48.model

Table 2: Training models of 8 classifiers.

classifier	Name of model	Save model
LibD3C	$Model_1$	LibD3C.model
RF	$Model_6$	RF.model
Libsvm	$Model_7$	Libsvm.model
SMO	$Model_8$	SMO.model
Decision stump	$Model_9$	Decision stump.model
Naive Bayes	$Model_{10}$	Naive Bayes.model
IB1	$Model_{11}$	IB1.model
J48	$Model_{12}$	J48.model

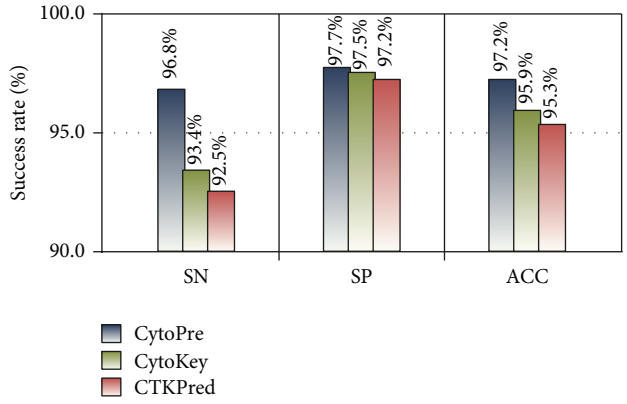


Figure 13: Performance comparison of 3 cytokine prediction systems.

the PFAM database. We then extracted the longest protein sequence in each family. To establish an effective classification model without deviation, we removed redundant sequences based on a sequence consistency standard. We extracted 120D feature vectors of positive and negative sequences based on the distribution of AAs with certain physicochemical properties and further sampled these to set up a training set. We then developed an ensemble classifier LibD3C to improve the stability and accuracy of cytokine classification. Cytokine identification was improved significantly in this paper in terms of accuracy and precision.

A series of experiments demonstrated the effectiveness of our method. We designed two group experiments to compare our methods (120D features) with Cai's (188D various features). The training results of our methods by LibD3C yielded SN = 96.8%, SP = 97.7%, and GM = 97.2%. In addition, the testing results of our methods were SN = 93.7%, SP = 92.9%, and GM = 93.3%. Two experimental sets of data generated by Cai's are SN = 96.4%, SP = 96.4%, and GM = 96.4%; SN = 92.9%, SP = 92.9%, and GM = 92.1%. The experimental results demonstrate that our method is superior to Cai's method in terms of classification validity because the hybrid approach may increase the weight of some information content and it is not conducive to all kinds of feature information extraction.

To prove that sampling has a significant influence on classification accuracy, we trained two groups of datasets by LibD3C. The first group used the test dataset (126 positive instances and 10588 negative instances) without sampling, while the second group used the rebuilding test dataset with SMOTE oversampling and K-means clustering undersampling, and they are extracted by 120D feature extraction algorithm. Experimentally, sampling is necessary to obtain good results. The SMOTE and K-means clustering algorithms were applied to small class and big class datasets, respectively. It avoids introducing excessive noise to the sampling set by SMOTE and effectively solves the problem of sample sparsity in the training set. In our approach, we employed a new type of ensemble classifier called LibD3C [32], which is a library for dynamic selection and circulating combination based on clustering. The ensemble classifier contained 18 basic classifiers and integrated some of these classifiers dynamically according to different objects of classification. Our goal is to achieve a classification result with the highest stability and accuracy. We developed eight groups of experiments to test the performance of LibD3C and conducted 10-fold cross validation using the rebuilding training set with LibD3C and seven basic classifiers. The results showed that the performance of the RT and Libsvm classifiers approached that of the ensemble classifier LibD3C. However, considering the sensitivity and specificity of the classifiers overall, LibD3C has obvious advantages.

Finally, we tested all protein sequences (539616) obtained from the UniProt database with the model trained through the method described prev and obtained 4151 cytokines. These cytokines are shown in the Supplementary Materials in FASTA format.

4. Conclusions

As a new interdisciplinary technology in the bioinformatics field, cytokine identification plays a very important role in the study of human disease. Studies that aim to improve the accuracy of cytokine prediction are of particular importance. To systematically present our experimental results and improve ease of use, we developed an online web server for cytokine prediction. Users input protein sequences that need to be predicted, and the server indicates which sequences are cytokines and displays geometric mean (GM) values of prediction. The results response to the HTML interface display whether it is cytokine and the prediction probability. The cytokine online prediction system can be accessed through http://datamining.xmu.edu.cn/software/CytoPre. The web site also provides related datasets and software for download.

Acknowledgments

The work was supported by the Natural Science Foundation of China (no. 61001013, 81101115, 61102136), the Natural Science Foundation of Fujian Province of China (no. 2011J05158, 2011J01371), the Natural Scientific Research Innovation Foundation in Harbin Institute of Technology (HIT.NSRIF.2013103), the Natural Science Foundation of Guangdong Province (no. S2012040007390), and Shanghai Key Laboratory of Intelligent Information Processing, China (no. IIPL-2012-002, no. IIPL-2011-004). Yunfeng Wu was also supported by the 2013 Program for New Century Excellent Talents in Fujian Province University.

References

[1] Q. Zou, W. C. Chen, Y. Huang, X. R. Liu, and Y. Jiang, "Identifying multi-functional enzyme with hierarchical multi-label classifier," *Journal of Computational and Theoretical Nanoscience*, vol. 10, no. 4, pp. 1038–1043, 2013.

[2] C. Lin, Y. Zou, J. Qin et al., "Hierarchical classification of protein folds using a novel ensemble classifier," *PLoS ONE*, vol. 8, no. 2, Article ID e56499, 2013.

[3] Y. Yabuki, T. Muramatsu, T. Hirokawa, H. Mukai, and M. Suwa, "GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model," *Nucleic Acids Research*, vol. 33, no. 2, pp. W148–W153, 2005.

[4] P. K. Papasaikas, P. G. Bagos, Z. I. Litou, and S. J. Hamodrakas, "A novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden markov models," *SAR and QSAR in Environmental Research*, vol. 14, no. 5-6, pp. 413–420, 2003.

[5] C.-S. Yu, Y.-C. Chen, C.-H. Lu, and J.-K. Hwang, "Prediction of protein subcellular localization," *Proteins: Structure, Function and Genetics*, vol. 64, no. 3, pp. 643–651, 2006.

[6] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne, "A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites," *International Journal of Neural Systems*, vol. 8, no. 5-6, pp. 581–599, 1997.

[7] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[8] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[9] W. R. Pearson, "Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms," *Genomics*, vol. 11, no. 3, pp. 635–650, 1991.

[10] G. S. Ladics, G. A. Bannon, A. Silvanovich, and R. F. Cressman, "Comparison of conventional FASTA identity searches with the 80 amino acid sliding window FASTA search for the elucidation of potential identities to known allergens," *Molecular Nutrition and Food Research*, vol. 51, no. 8, pp. 985–998, 2007.

[11] N. Huang, H. Chen, and Z. Sun, "CTKPred: an SVM-based method for the prediction and classification of the cytokine superfamily," *Protein Engineering, Design and Selection*, vol. 18, no. 8, pp. 365–368, 2005.

[12] S. Lata and G. P. S. Raghava, "CytoPred: a server for prediction and classification of cytokines," *Protein Engineering, Design and Selection*, vol. 21, no. 4, pp. 279–282, 2008.

[13] A. Bateman, L. Coin, R. Durbin et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 32, pp. D138–D141, 2004.

[14] T. Huang, L. Chen, Y.-D. Cai, and K.-C. Chou, "Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property," *PLoS ONE*, vol. 6, no. 9, Article ID e25297, 2011.

[15] H. Altınçay and C. Ergün, "Clustering based under-sampling for improving speaker verification decisions using AdaBoost," in *Structural, Syntactic, and Statistical Pattern Recognition*, pp. 698–706, Springer, New York, NY, USA, 2004.

[16] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *Advances in intelligent computing*, vol. 3644 of *Lecture Notes in Computer Science*, pp. 878–887, Springer, August 2005.

[17] C. Z. Cai, L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen, "SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3692–3697, 2003.

[18] K.-C. Chou and Y.-D. Cai, "Predicting protein structural class by functional domain composition," *Biochemical and Biophysical Research Communications*, vol. 321, no. 4, pp. 1007–1009, 2004.

[19] A. Bairoch, R. Apweiler, C. H. Wu et al., "The Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 33, pp. D154–D159, 2005.

[20] R. Apweiler, A. Bairoch, C. H. Wu et al., "UniProt: the universal protein knowledgebase," *Nucleic Acids Research*, vol. 32, supplement 1, pp. D115–D119, 2004.

[21] C. H. Wu, R. Apweiler, A. Bairoch et al., "The Universal Protein Resource (UniProt): an expanding universe of protein information," *Nucleic Acids Research*, vol. 34, supplement 1, pp. D187–D191, 2006.

[22] Y.-D. Cai, X.-J. Liu, X.-B. Xu, and K.-C. Chou, "Prediction of protein structural classes by support vector machines," *Computers & Chemistry*, vol. 26, no. 3, pp. 293–296, 2002.

[23] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *Journal of Molecular Biology*, vol. 238, no. 1, pp. 54–61, 1994.

[24] J. R. Bock and D. A. Gough, "Predicting protein-protein interactions from primary structure," *Bioinformatics*, vol. 17, no. 5, pp. 455–460, 2001.

[25] R. Karchin, K. Karplus, and D. Haussler, "Classifying G-protein coupled receptors with support vector machines," *Bioinformatics*, vol. 18, no. 1, pp. 147–159, 2002.

[26] S. Hua and Z. Sun, "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach," *Journal of Molecular Biology*, vol. 308, no. 2, pp. 397–407, 2001.

[27] Z. Yuan, K. Burrage, and J. S. Mattick, "Prediction of protein solvent accessibility using Support Vector Machines," *Proteins: Structure, Function and Genetics*, vol. 48, no. 3, pp. 566–570, 2002.

[28] C. H. Q. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, no. 4, pp. 349–358, 2001.

[29] L. Nanni and A. Lumini, "MppS: an ensemble of support vector machine based on multiple physicochemical properties of amino acids," *Neurocomputing*, vol. 69, no. 13–15, pp. 1688–1690, 2006.

[30] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: many could be better than all," *Artificial Intelligence*, vol. 137, no. 1-2, pp. 239–263, 2002.

[31] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: a k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C*, vol. 28, no. 1, pp. 100–108, 1979.

[32] Q. Zou, X. B. Li, Y. Jiang, Y. M. Zhao, and G. H. Wang, "BinMemPredict: a web server and software for predicting membrane protein types," *Current Proteomics*, vol. 10, no. 1, pp. 2–9, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 246761, 11 pages
http://dx.doi.org/10.1155/2013/246761

Research Article

Optimal Control of Gene Regulatory Networks with Effectiveness of Multiple Drugs: A Boolean Network Approach

Koichi Kobayashi and Kunihiko Hiraishi

School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa 923-1292, Japan

Correspondence should be addressed to Koichi Kobayashi; k-kobaya@jaist.ac.jp

Received 30 April 2013; Revised 25 June 2013; Accepted 12 July 2013

Academic Editor: Lei Chen

Developing control theory of gene regulatory networks is one of the significant topics in the field of systems biology, and it is expected to apply the obtained results to gene therapy technologies in the future. In this paper, a control method using a Boolean network (BN) is studied. A BN is widely used as a model of gene regulatory networks, and gene expression is expressed by a binary value (0 or 1). In the control problem, we assume that the concentration level of a part of genes is arbitrarily determined as the control input. However, there are cases that no gene satisfying this assumption exists, and it is important to consider structural control via external stimuli. Furthermore, these controls are realized by multiple drugs, and it is also important to consider multiple effects such as duration of effect and side effects. In this paper, we propose a BN model with two types of the control inputs and an optimal control method with duration of drug effectiveness. First, a BN model and duration of drug effectiveness are discussed. Next, the optimal control problem is formulated and is reduced to an integer linear programming problem. Finally, numerical simulations are shown.

1. Introduction

In the field of systems biology, there have been a lot of studies on modeling, analysis, and control of gene regulatory networks. Especially, control of gene regulatory networks corresponds to therapeutic interventions, which are realized by radiation, chemotherapy, and so on. In order to develop gene therapy technologies (see, e.g., [1]) in the future, developing control theory of gene regulatory networks is important. Furthermore, in recent years, the important result on control of gene regulatory networks has been obtained in [2]. That is, feedback control of synthetic biological circuits has been implemented, and the experimental result in which cellular behavior is regulated by control has been obtained. This result suggests that control methods of gene regulatory networks can be realized. Motivated by the above background, we study control methods of gene regulatory networks.

Gene regulatory networks are in general expressed by ordinary/partial differential equations with high nonlinearity and high dimensionality. In order to deal with such a system, it is important to consider a simple model, and various models such as Bayesian networks, Boolean networks (BNs) [3], hybrid systems (piecewise affine models), and Petri nets have been developed so far (see, e.g., [4] for further details). In control problems, BNs and hybrid systems are frequently used [5–8]. In the hybrid systems-based approach, the class of gene regulatory networks are limited to low-dimensional systems, because the computation time to solve the control problem is too long. In a BN, dynamics such as interactions between genes are expressed by the Boolean functions [3]; that is, gene expression is expressed by a binary value (0 or 1). There is a criticism that a Boolean network is too simple as a model of gene regulatory networks (see, e.g., [9]), but this model can be relatively applied to large-scale systems. In addition, since the behavior of gene regulatory networks is stochastic by the effects of noise, it is appropriate that a Boolean function is randomly decided at each time among the candidates of the Boolean functions. From this viewpoint, a probabilistic Boolean network (PBN) has been proposed in [10]. Furthermore, a context-sensitive PBN (CS-PBN) in which the deciding time is randomly selected has been proposed as a general form of PBNs [11, 12].

Furthermore, in the control theory of gene regulatory networks, the control input is given by the concentration level

of a part of genes; that is, we assume that the concentration level of a part of genes can be arbitrarily determined. However, in the case where this assumption is not satisfied, it is important to consider structural control via external stimuli [13, 14]. These controls are realized by multiple drugs, and it is also important to consider multiple effects such as duration of effect and side effects [15]. To our knowledge, a unified method considering these properties has not been proposed so far.

Thus, in this paper, we propose a BN model with two types of the control inputs and an optimal control method with duration of drug effectiveness. The first control input is the control input satisfying the assumption that the binary value is arbitrarily determined. The second control input is called a structural control input herein, and the dynamics, that is, the Boolean functions, are selected among the candidates of the dynamics. However, it is difficult to uniquely select one Boolean function. Hence, we suppose that one Boolean function is selected probabilistically, and the probability distribution is switched by using the structural control input. A structural control method has been discussed in [13, 14], but the notion of the structural control input defined in this paper is different from that in those existing methods. Since the proposed BN model has a switch of the probability distribution, it may be regarded as a generalized version of PBNs.

In optimal control of PBNs and CS-PBNs, many results have been obtained so far (see, e.g., [11, 12, 16–21]). In many existing results, state transition diagrams with 2^n nodes (i.e., $2^n \times 2^n$ transition probability matrices) must be computed for a PBN with n states. As a result, in order to compute state transition diagrams, several issues such as memory consumption must be considered in implementation, and it is desirable to directly use a given Boolean function. The authors have proposed in [22] a control method in which state transition diagrams are not computed. In many existing results, we consider finding a control input such that the expected value of the cost function is minimized. In [22], we consider finding a control input such that the lower bound of the cost function is minimized under a certain constraint condition. Owing to this difference, state transition diagrams are not computed in the method in [22], and the optimal control problem is reduced to an integer linear programming (ILP) problem. Also in [16], ILP-based methods were proposed for other optimal control problems, and in those methods, solving multiple ILP problems is required.

In this paper, based on our previously proposed method [22], the optimal control problem with duration of drug effectiveness is reduced to an ILP problem. Since a given Boolean function is directly used, duration of drug effectiveness can be easily described as a linear inequality constraint. The proposed method provides us with a basic in control theory of gene regulatory networks. The conference paper [23] is a preliminary version of this paper. In this paper, we provide improved formulations and explanations, a discussion on duration of drug effectiveness, and a numerical simulation using the large-scale BN.

This paper is organized as follows. In Section 2.1, the Boolean networks with two kinds of the control inputs are proposed. In Section 2.2, duration of drug effectiveness is introduced. In Section 2.3, the optimal control problem is formulated. In Section 2.4, its solution method is proposed. In Section 3, two numerical examples are presented. In Section 4, we conclude this paper.

Notation. Let $\mathscr{R}$ denote the set of real numbers. Let $\{0,1\}^n$ denote the set of n-dimensional vectors, which consists of elements 0 and 1. Let I_n and $0_{m\times n}$ denote the $n \times n$ identity matrix and the $m \times n$ zero matrix, respectively. For simplicity, we sometimes use the symbol 0 instead of $0_{m\times n}$ and and the symbol I instead of I_n. For a matrix M, $\ln M$ denotes the matrix such that the (i, j)th element is given as the natural logarithm of the (i, j)th element in M. For a matrix M, M^T denotes the transpose of M.

2. Materials and Methods

2.1. The Boolean Networks with Control Inputs. A Boolean network (BN) with n states is given by

$$x(k+1) = f_a(x(k)), \tag{1}$$

where $x \in \{0,1\}^n$ is the state (e.g., the concentration of genes) and $k = 0, 1, 2, \ldots$ is the discrete time. The function $f_a : \{0,1\}^n \to \{0,1\}^n$ is a given Boolean function with logical operators such as AND ($\wedge$), OR ($\vee$), and NOT ($\neg$). If the BN (1) is deterministic, then the next state $x(k+1)$ is uniquely determined for a given $x(k)$. See also Example 1 for an example.

Next, the control inputs are added to a BN (1). For the BN (1) with n state, consider two types of the control inputs. First, in a similar way to that of the conventional control method, the control input is added to the BN (1) as follows:

$$x(k+1) = f(x(k), u(k)), \tag{2}$$

where $u \in \{0,1\}^m$ is the control input; that is, the value of u (e.g., the concentration of genes) can be arbitrarily given, and $f : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}^n$ is a given Boolean function. The ith element of the state x and the ith element of the control input u are denoted by x_i and u_i, respectively. In the BN (2), $x(k+1)$ is uniquely determined for the given $x(k)$ and $u(k)$.

Then, consider the structural control input. Suppose that the candidates of f are given by f_i, $i = 1, 2, \ldots, l$. It will be difficult to select one Boolean function uniquely. In this paper, we assume that one discrete probability distribution is selected among m_s discrete probability distributions. Probabilistic distributions are derived from experimental results, but details are one of the future works. Then, a method for inferring a probabilistic Boolean network will be useful (see, e.g., [24]). Let $r_{i,j}$ denote the probability that the Boolean function f_j is selected in the ith discrete probability distribution. Then,

$$\sum_{j=1}^{l} r_{i,j} = 1, \quad i = 1, 2, \ldots, m_s, \tag{3}$$

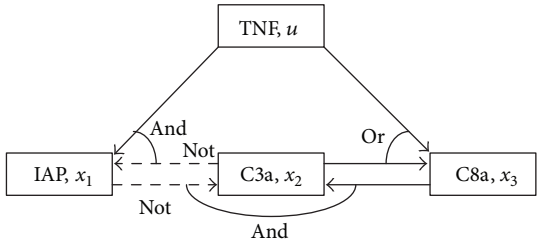


Figure 1: A simplified model of an apoptosis network: activation (solid) and inhibition (broken).

hold. In addition, m_s-dimensional binary variables $u^s \in \{0,1\}^{m_s}$ are assigned to m_s discrete probability distributions, and let u_i^s denote the ith element of u^s. The structural control input u^s corresponds to m_s kinds of external stimuli. Then, the equality constraint

$$\sum_{i=1}^{m_s} u_i^s(k) = 1 \tag{4}$$

is imposed. Here, we show a simple example.

Example 1. As a simple example, consider the simplified model of an apoptosis network in Figure 1 [25]. Then, the Boolean network model expressing this apoptosis network is given by

$$\begin{aligned} x_1(k+1) &= \neg x_2(k) \wedge u(k), \\ x_2(k+1) &= \neg x_1(k) \wedge x_3(k), \\ x_3(k+1) &= x_2(k) \vee u(k), \end{aligned} \tag{5}$$

where the concentration level (high or low) of the inhibitor of apoptosis proteins (IAPs) is denoted by x_1, the concentration level of the active caspase 3 (C3a) is denoted by x_2, and the concentration level of the active caspase 8 (C8a) is denoted by x_3. The concentration level of the tumor necrosis factor (TNF, a stimulus) is denoted by u and is regarded as the control input. Since the caspase C3a is responsible for cleaving or breaking many other proteins, a high level of the C3a concentration, that is, $x_2 = 1$, implies cell near-death, otherwise, cell survival. As seen in (5), if the concentration of IAP is high ($x_1 = 1$) or the concentration of the caspase C8a is low ($x_3 = 0$), then the concentration of C3a becomes low; that is, $x_2 = 0$. On the other hand, x_1 and x_3 at the next time depend on the value of x_2 as well as u. In this way, some dynamical interactions exist. See [25, 26] for further details.

Suppose that $l = 2$ and $m_s = 2$. Then, as an example of the candidates of the Boolean functions, we consider the following:

$$\begin{aligned} f_1 &= \begin{bmatrix} \neg x_2(k) \wedge u(k) \\ \neg x_1(k) \wedge x_3(k) \\ x_2(k) \vee u(k) \end{bmatrix}, \quad r_{1,1} = 0.8, \quad r_{2,1} = 0.1, \\ f_2 &= \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \end{bmatrix}, \quad r_{1,2} = 0.2, \quad r_{2,2} = 0.9. \end{aligned} \tag{6}$$

We suppose that the Boolean function f_1 expresses the situation that the dynamics of an apoptosis network are selected with high probability and that the Boolean function f_2 expresses the situation that the state is not changed with high probability. By using u_1^s and u_2^s, one of the two discrete probability distributions $\{r_{1,1}, r_{1,2}\}$ and $\{r_{2,1}, r_{2,2}\}$ is selected at each time.

A BN with two types of the control inputs includes the probabilistic behavior, and we assume that the probability distribution can be controlled. From these facts, a BN studied in this paper can be regarded as a generalized form of a probabilistic Boolean network (PBN). To explain the relation between the proposed BN model and a PBN, we show a simple example.

Example 2. As a simple example, consider the PBN with three states and one control input. Suppose that the Boolean functions are given as follows:

$$\begin{aligned} x_1(k+1) &= \begin{cases} x_3(k) \vee u(k), & \text{with the probability } 0.8, \\ \neg x_3(k), & \text{with the probability } 0.2, \end{cases} \\ x_2(k+1) &= x_1(k) \wedge \neg x_3(k), \text{ with the probability } 1.0, \\ x_3(k+1) &= \begin{cases} x_1(k) \wedge \neg x_2(k), & \text{with the probability } 0.7, \\ x_2(k) \vee u(k), & \text{with the probability } 0.3. \end{cases} \end{aligned} \tag{7}$$

This PBN corresponds to the cases of $l = 4$ and $m_s = 1$. The candidates of the Boolean functions f_i, $i = 1,2,3,4$, and the probabilities $r_{1,j}$, $j = 1,2,3,4$, are obtained as follows:

$$\begin{aligned} f_1 &= \begin{bmatrix} x_3(k) \vee u(k) \\ x_1(k) \wedge \neg x_3(k) \\ x_1(k) \wedge \neg x_2(k) \end{bmatrix}, \quad r_{1,1} = 0.56, \\ f_2 &= \begin{bmatrix} x_3(k) \vee u(k) \\ x_1(k) \wedge \neg x_3(k) \\ x_2(k) \vee u(k) \end{bmatrix}, \quad r_{1,2} = 0.24, \\ f_3 &= \begin{bmatrix} \neg x_3(k) \\ x_1(k) \wedge \neg x_3(k) \\ x_1(k) \wedge \neg x_2(k) \end{bmatrix}, \quad r_{1,3} = 0.14, \\ f_4 &= \begin{bmatrix} \neg x_3(k) \\ x_1(k) \wedge \neg x_3(k) \\ x_2(k) \vee u(k) \end{bmatrix}, \quad r_{1,4} = 0.06. \end{aligned} \tag{8}$$

Next, consider the state orbit of this PBN. In PBNs, one Boolean function is probabilistically selected at each time. Then, for $x(0) = [0\ 0\ 0]^T$ and $u(0) = 0$, we obtain the following:

$$\begin{aligned} &\text{Prob}\left(x(1) = [0\ 0\ 0]^T \mid x(0) = [0\ 0\ 0]^T\right) = 0.8, \\ &\text{Prob}\left(x(1) = [1\ 0\ 0]^T \mid x(0) = [0\ 0\ 0]^T\right) = 0.2. \end{aligned} \tag{9}$$

In this example, the cardinality of the finite state set $\{0,1\}^3$ is given by $2^3 = 8$, and we obtain the state transition diagram of

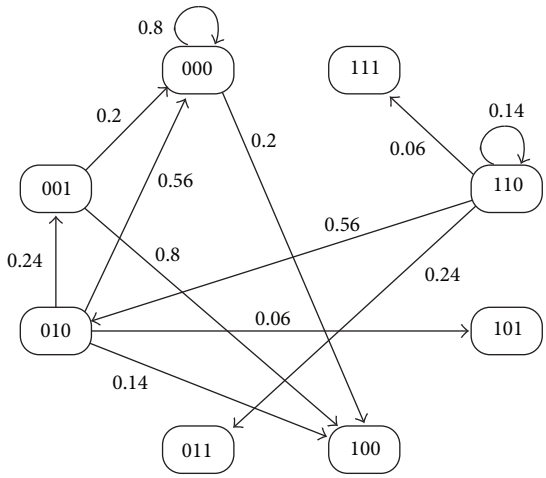


FIGURE 2: The state transition diagram with $u(k) = 0$.

Figure 2 by computing the transition from each value of the state. In Figure 2, the number assigned to each node denotes x_1, x_2, and x_3 (each element of the state), and the number assigned to each arc denotes the transition probability from some state to another state. For simplicity of illustration, the state transitions from $x(k) = [0\ 0\ 0]^T, [0\ 0\ 1]^T, [0\ 1\ 0]^T, [1\ 1\ 0]^T$ are illustrated in Figure 2. In the existing solution methods for optimal control of PBNs, the optimal control input is computed using dynamic programming with state transition diagrams.

As shown in this example, computing state transition diagrams with 2^n nodes (n is the number of the state) is required in the existing solution methods for optimal control of PBNs with n states, and this computation is hard for large-scale systems (see also Section 3.2). Thus, it is important to consider a new solution method. In this paper, for BNs with two types of the control inputs, a solution method using integer programming is proposed based on our previously proposed work in [22]. In the proposed method, computation of state transition diagrams such as that in Figure 2 is not needed.

Remark 3. By adding the candidates of the Boolean functions, BNs with two types of the control inputs can be transformed into BNs with only the structural control input. That is, the control input u can be eliminated from (2) by fixing the value of u in (2). Then, the number of the candidates of Boolean functions is $2^m l$, and 2^m combinations for u must be computed in advance. To avoid this computation, we consider two types of the control inputs.

2.2. Duration of Drug Effectiveness. The control input u and the structural control input u^s are realized by using multiple drugs. Then, we must consider the multiple effects such as duration of effect and the side effects. In this paper, we focus on the duration of drug effectiveness. In, for example, chemotherapy, therapeutic intervention is generally applied to the target cell in a cyclic manner [15]. Each therapeutic window is started by delivering the drug. The drug delivered is effective on the target cell for some period of time. This is followed by a recovery phase. However, when the drug is not delivered, the drug may be delivered in the timing that is faster than the next time in a cyclic [15]. Therefore, it is necessary to model several situations on duration of drug effectiveness. To model the duration of effect, three parameters $L_{u_i}, W^1_{u_i}$, and $W^0_{u_i}$ are defined for each input u_i (or u^s_i). The parameters L_{u_i} and $W^1_{u_i}$ have been already defined in [15].

The parameter L_{u_i} is the length of the drug effectiveness period. That is, if $u_i(k) = 1$, then $u_i(k+1) = u_i(k+2) = \cdots = u_i(k+L_{u_i}) = 1$ holds. Next, $W^1_{u_i}(> L_{u_i})$ is explained. If $u_i(k) = 1$, then $u_i(k+1), u_i(k+2), \ldots, u_i(k+W^1_{u_i}-1)$ is uniquely determined depending on L_{u_i}, and $u_i(k+W^1_{u_i})$ is a decision variable. Then, $W^1_{u_i} - L_{u_i} - 1$ corresponds to the length of a recovery phase. Finally, $W^0_{u_i}$ is explained. If $u_i(k) = 0$, then $u_i(k+1) = u_i(k+2) = \cdots = u_i(k+W^0_{u_i}-1) = 0$ holds, and $u_i(k+W^0_{u_i})$ is a decision variable. By using $L_{u_i}, W^1_{u_i}, W^0_{u_i}$, we can consider several situations, and we show two typical examples.

Example 4. First, suppose that, for the control input $u \in \{0,1\}^1$, L_u, W^1_u, and W^0_u are given as $L_u = 1$, $W^1_u = 3$, and $W^0_u = 2$, respectively. Consider the case of $u(k) = 1$. Then, $u(k+1)$ and $u(k+2)$ are uniquely determined as $u(k+1) = 1$ and $u(k+2) = 0$, respectively, and $u(k+3)$ is a decision variable. Then, $u(k+2) = 0$ is the recovery phase, and $W^1_u - L_u - 1 = 1$ is its length. In the case of $u(k) = 0$, the relation $u(k+1) = 0$ holds, and $u(k+2)$ is a decision variable.

Another example is shown. Suppose that, for the control input $u \in \{0,1\}^1$, L_u, W^1_u, and W^0_u are given as $L_u = 0$, $W^1_u = 3$, and $W^0_u = 3$, respectively. In both, the case of $u(k) = 1$ and the case of $u(k) = 0$, $u(k+1) = u(k+2) = 0$ holds, and $u(k+3)$ is a decision variable. In this case, $u(k+1) = u(k+2) = 0$ is the recovery phase, and $W^1_u - L_u - 1 = 2$ is its length.

By using the three parameters $L_{u_i}, W^1_{u_i}$, and $W^0_{u_i}$, several situations on the duration of effect can be modeled (see also [15]). In addition, since these parameters can be given for each u_i (or u^s_i), effectiveness of multiple drugs can be evaluated. Thus, in this paper, we consider not only two types of the control inputs but also duration of drug effectiveness.

2.3. Optimal Control Problem. First, the following two notations are defined. Let $\pi_i(k)$ denote the probability that some Boolean function f_i is selected at time k. In addition, the probability that some time sequence of theBoolean functions $f_{i(k_1)}, f_{i(k_1+1)}, \ldots, f_{i(k_2)}$ is selected at time interval $[k_1, k_2]$ is denoted by

$$\pi(k_1, k_2) := \prod_{k=k_1}^{k_2} \pi_{i(k)}(k). \tag{10}$$

For simplicity of notation, $i(k_1), i(k_1+1), \ldots, i(k_2)$ are omitted in $\pi(k_1, k_2)$.

Next, for the Boolean networks with n states and two types of the control inputs, consider the following optimal control problem.

Problem 1. Suppose that, for the Boolean network with n states and two types of the control inputs, the initial state $x(0) = x_0$, ρ satisfying $0 \leq \rho \leq 1$, the control time N, the parameters on duration of drug effectiveness $L_{u_i(u_i^s)}$, $W^1_{u_i(u_i^s)}$, and $W^0_{u_i(u_i^s)}$ are given. Then, for all combinations of the Boolean functions satisfying the constraint

$$\pi(0, N-1) \geq \rho, \tag{11}$$

find two control input sequences $u(0), u(1), \ldots, u(N-1)$ and $u^s(0), u^s(1), \ldots, u^s(N-1)$ minimizing the lower bound of the cost function

$$J = \sum_{k=0}^{N-1} \left\{ Qx(k) + Ru(k) + R_s u^s(k) \right\} + Q_f x(N) \tag{12}$$

subject to the constraint on duration of drug effectiveness, where $Q, Q_f \in \mathcal{R}^{1\times n}$, $R \in \mathcal{R}^{1\times m}$, and $R \in \mathcal{R}^{1\times m_s}$ are weighting vectors whose elements are nonnegative real numbers.

For simplicity of discussion, a linear function with respect to x, u, and u^s is considered as a cost function. We consider that a linear cost function is appropriate from the following two reasons.

(i) For a binary variable $\delta \in \{0, 1\}$, the relation $\delta^2 = \delta$ holds. That is, in the cost function, the quadratic term such as $x_i^2(k)$ is not necessary.

(ii) In control of gene regulatory networks, the expression of a certain gene is frequently focused (see, e.g., [18]). For example, in the gene regulatory network related to melanoma, it important to inhibit the concentration level of the gene WNT5A [27]. In this case, it is enough to consider the cost function (12).

Furthermore, in many existing methods on optimal control of PBNs, the expected value of a nonnegative function is frequently used as a cost function (see, e.g., [11, 12, 17–21]). However, the expected value of the state must be computed from all combinations of the Boolean functions, and this computation is hard for large-scale systems. To avoid this computation, in this paper, we evaluate the control performance by using the lower bound. If the constraint (11) is not included in Problem 1, then the behaviors are regarded as uncertain (nondeterministic) behaviors, and the best performance is derived in Problem 1. Since the combinations of the Boolean functions selected with low probability are included, performance evaluation is not appropriate. In order to exclude such combinations, we impose the constraint (11). Similar problem formulations have been considered in optimal control of stochastic hybrid systems (see, e.g., [28–30]). Thus, since the performance index in this paper is different from that in existing methods, it is difficult to directly compare the performance of the proposed method with those of existing methods. On the other hand, in [22], we discussed this topic from the qualitative viewpoint. In [22], the upper bound is also computed by using the control input such that the lower bound is minimized. If the lower bound and the upper bound are not improved by control, then the expected value will not be improved. Then, it is important to suitably set ρ in the constraint (11). See [22] for further details.

We show an example for setting weighting vectors from the biological viewpoint.

Example 5. Consider the Boolean network expressing an apoptosis network in Example 1 again. For this system, we consider finding a control strategy such that a stimulus u is not applied as much as possible, and cell survival is achieved; $u = 0$ implies that a stimulus is not applied to the system, and $x_1 = 1$ and $x_2 = 0$ express cell survival [25]. Then, as one of the appropriate cost functions, we can consider the following cost function:

$$\begin{aligned} J = \sum_{i=0}^{N-1} & \left\{ 10\left|x_1(i) - 1\right| + 10\left|x_2(i) - 0\right| + u(i) \right\} \\ & + 100\left|x_1(N) - 1\right| + 100\left|x_2(N) - 0\right|. \end{aligned} \tag{13}$$

By the coordinate transformation of x_1 into $1 - x_1$, this cost function can be rewritten as the form of (12).

2.4. Solution Method. We propose a solution method for Problem 1. First, two lemmas are introduced as preparations. Next, Problem 1 is reduced to an integer linear programming (ILP) problem.

As preparations, two lemmas are introduced. To reduce Problem 1 to an ILP problem, it is necessary to transform a Boolean function into a polynomial on the real number field. First, the following lemma [31] is used.

Lemma 6. *Consider the two binary variables δ_1, and δ_2. Then, the following relations hold:*

(i) $\neg\delta_1$ *is equivalent to* $1 - \delta_1$,

(ii) $\delta_1 \vee \delta_2$ *is equivalent to* $\delta_1 + \delta_2 - \delta_1\delta_2$,

(iii) $\delta_1 \wedge \delta_2$ *is equivalent to* $\delta_1\delta_2$.

For example, $\delta_1 \vee \neg\delta_2$ is equivalently transformed into $\delta_1 + (1-\delta_2) - \delta_1(1-\delta_2) = 1 - \delta_2 + \delta_1\delta_2$. Furthermore, the product of binary variables such as $\delta_1\delta_2$ can be linearized by using the following lemma [32].

Lemma 7. *Suppose that the binary variables $\delta_j \in \{0, 1\}$ and $j \in \mathcal{J}$ are given, where $\mathcal{J}$ is some index set. Then, $z = \prod_{j\in\mathcal{J}}\delta_j$ is equivalent to the following linear inequalities:*

$$\sum_{j\in\mathcal{J}} \delta_j - z \leq |\mathcal{J}| - 1, \qquad -\sum_{j\in\mathcal{J}} \delta_j + |\mathcal{J}|\, z \leq 0, \tag{14}$$

where $|\mathcal{J}|$ is the cardinality of $\mathcal{J}$.

From Lemmas 6 and 7, we see that any Boolean function can be equivalently transformed into a pair of some linear function and some linear inequality. See [31, 32] for further details. For example, $\delta_1 \vee \neg\delta_2$ is equivalent to a pair of $1 - \delta_2 + z$ and $z = \delta_1\delta_2$. By using Lemma 7, $z = \delta_1\delta_2$ can be expressed as a set of linear inequalities.

Now, we consider reducing Problem 1 to an ILP problem.

By using Lemma 6, the candidates of the Boolean functions $f_i(x(k), u(k))$, $i = 1, 2, \ldots, l$, are transformed into a polynomial on the real number field. Let $\widehat{f}_i(x(k), u(k))$ denote the polynomial obtained. Then, consider the following system using $\widehat{f}_i(x(k), u(k))$:

$$x(k+1) = \sum_{i=1}^{l} \left\{ \delta_i(k)\, \widehat{f}_i(x(k), u(k)) \right\}, \tag{15}$$

where $\delta_1(k), \delta_2(k), \ldots, \delta_l(k)$ are binary variables satisfying

$$\sum_{i=1}^{l} \delta_i(k) = 1. \tag{16}$$

The binary vector $\delta(k) := \left[\delta_1(k) \;\; \delta_2(k) \;\; \cdots \;\; \delta_l(k)\right]^T$ is used to select the polynomial $\widehat{f}_i$ and to express (11) as a linear form. Here, we define the following vector:

$$S_i := \left[r_{i,1} \;\; r_{i,2} \;\; \cdots \;\; r_{i,l}\right]. \tag{17}$$

Then, by using the natural logarithm, $\pi(0, N-1)$ in (11) is expressed as

$$\ln \pi(0, N-1) = \sum_{k=0}^{N-1} \left(\sum_{i=1}^{m_s} \ln S_i u_i^s(k) \right) \delta(k). \tag{18}$$

In this expression, one probability distribution is selected by using $u_i^s(k)$, and the probability that a certain Boolean function is selected is determined by $\delta(k)$. Then, Problem 1 is equivalent to the following problem.

Problem A.

$$\begin{aligned}
&\text{find} && u(k), u^s(k), \delta(k), k = 0, 1, \ldots, N-1,\\
&\min && \text{Cost function (12)},\\
&\text{subject to} && \text{System (15)}, x(0) = x_0,\\
&&& \text{Inequality constraint:}\\
&&& \sum_{k=0}^{N-1} \left(\sum_{i=1}^{m_s} \ln S_i \delta_i^s(k) \right) \delta(k) \geq \ln \rho,\\
&&& \text{Equality constraint (4), (16)},\\
&&& \text{Constraint on duration of drug effectiveness.}
\end{aligned} \tag{19}$$

By using Lemma 7, the system (15) and $(\sum_{i=1}^{m_s} \ln S_i \delta_i^s(k))\delta(k)$ can be equivalently expressed in the following linear form:

$$x(k+1) = Ax(k) + B_u u(k) + B_s u^s(k) + B_b z_b(k), \tag{20}$$

$$\left(\sum_{i=1}^{m_s} \ln S_i \delta_i^s(k) \right) \delta(k) = \sum_{i=1}^{m_s} \ln S_i z_i^s(k), \tag{21}$$

$$Ex(k) + F_u u(k) + F_s u^s(k) + F_z z(k) \leq G, \tag{22}$$

where $z_i^s(k) := \delta_i^s(k)\delta(k)$, and (22) is the linear inequality obtained by applying Lemma 7 to (15). The vector $z_b(k) \in \{0,1\}^b$ is an auxiliary binary variable obtained by using Lemma 7, and the dimension of $z_b(k)$, that is, b, is determined depending on the form of the given Boolean functions. In addition, $z(k)$ is defined as

$$\begin{aligned}
z(k) &:= \left[\left(z_b(k)\right)^T \;\; \left(z_1^s(k)\right)^T \;\; \left(z_2^s(k)\right)^T \;\; \cdots \;\; \left(z_{m_s}^s(k)\right)^T \right]^T\\
&\in \{0,1\}^{b+m_s l}.
\end{aligned} \tag{23}$$

Here in after, for simplicity of notation, $B_b z_b(k)$ is rewritten as $B_z z(k)$, $B_z := \left[B_b \;\; 0\right]$, and $\sum_{i=1}^{m_s} \ln S_i z_i^s(k)$ is rewritten as $Cz(k)$, $C := \left[0 \;\; \ln S_1 \;\; \ln S_2 \;\; \cdots \;\; \ln S_{m_s}\right]$.

Now, we consider transforming Problem A by using (20), (21), and (22). By using

$$x(k) = A^k x_0 + \sum_{i=1}^{k} A^{i-1} \left(B_u u(k-i) + B_s u^s(k-i) + B_z z(k) \right) \tag{24}$$

obtained from the state equation in (20), we can obtain

$$\overline{x} = \overline{A} x_0 + \overline{B}_u \overline{u} + \overline{B}_s \overline{u}_s + \overline{B}_z \overline{z}, \tag{25}$$

where

$$\begin{aligned}
\overline{x} &:= \left[\left(x(0)\right)^T \;\; \left(x(1)\right)^T \;\; \cdots \;\; \left(x(N)\right)^T \right]^T,\\
\overline{u} &:= \left[\left(u(0)\right)^T \;\; \left(u(1)\right)^T \;\; \cdots \;\; \left(u(N-1)\right)^T \right]^T,\\
\overline{u}_s &:= \left[\left(u^s(0)\right)^T \;\; \left(u^s(1)\right)^T \;\; \cdots \;\; \left(u^s(N-1)\right)^T \right]^T,\\
\overline{z} &:= \left[\left(z(0)\right)^T \;\; \left(z(1)\right)^T \;\; \cdots \;\; \left(z(N-1)\right)^T \right]^T,\\
\overline{A} &= \begin{bmatrix} I \\ A \\ A^2 \\ \vdots \\ A^N \end{bmatrix}, \qquad \overline{B} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ I & 0 & \cdots & 0 \\ A & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ A^{N-1} & \cdots & A & I \end{bmatrix},\\
\overline{B}_u &= \overline{B} \begin{bmatrix} B_u & & 0 \\ & \ddots & \\ 0 & & B_u \end{bmatrix}, \qquad \overline{B}_s = \overline{B} \begin{bmatrix} B_s & & 0 \\ & \ddots & \\ 0 & & B_s \end{bmatrix},\\
\overline{B}_z &= \overline{B} \begin{bmatrix} B_z & & 0 \\ & \ddots & \\ 0 & & B_z \end{bmatrix}.
\end{aligned} \tag{26}$$

Next, the inequality constraint $\sum_{k=0}^{N-1} (\sum_{i=1}^{m_s} \ln S_i \delta_i^s(k))\delta(k) \geq \ln \rho$ in Problem A is equivalent to

$$-\overline{C}\overline{z} \leq -\ln \rho, \tag{27}$$

where $\overline{C} = \left[C \;\; C \;\; \cdots \;\; C\right]$. Furthermore, from (22), we can obtain

$$\overline{E}\overline{x} + \overline{F}_u \overline{u} + \overline{F}_s \overline{u}_s + \overline{F}_z \overline{z} \leq \overline{G}, \tag{28}$$

where

$$\overline{E} = \begin{bmatrix} E & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & E & 0 \end{bmatrix},$$

$$\overline{F}_u = \begin{bmatrix} F_u & \ddots & 0 \\ 0 & & F_u \end{bmatrix}, \qquad \overline{F}_s = \begin{bmatrix} F_s & \ddots & 0 \\ 0 & & F_s \end{bmatrix}, \tag{29}$$

$$\overline{F}_z = \begin{bmatrix} F_z & \ddots & 0 \\ 0 & & F_z \end{bmatrix}, \qquad \overline{G} = \begin{bmatrix} G \\ G \\ \vdots \\ G \end{bmatrix}.$$

Next, consider the constraint on duration of drug effectiveness. This constraint can be expressed as a Boolean function. Then, by using Lemmas 6 and 7, it can be transformed into the following form:

$$\overline{u} = V^1 \overline{v} + V^2, \qquad W^1 \overline{v} \le W^2, \tag{30}$$

$$\overline{u}_s = V_s^1 \overline{v}_s + V_s^2, \qquad W_s^1 \overline{v}_s \le W_s^2, \tag{31}$$

where $\overline{v}$ and $\overline{v}_s$ are binary decision variables with certain dimensions. Deriving a general form of coefficient matrices will be difficult, but for the given L_{u_i}, $W_{u_i}^1$, and $W_{u_i}^0$, deriving coefficient matrices is easy.

We show two examples.

Example 8. Consider Example 4 again. First, consider the case of $L_u = 1$, $W_u^1 = 3$, and $W_u^0 = 2$. Then, noting explanations in Example 4, we can obtain

$$\begin{aligned} u(0) &= v_0, \\ u(1) &= v_0, \\ u(2) &= (1 - v_0) v_2, \\ u(3) &= (1 - v_0) v_2 + v_0 v_3, \\ u(4) &= v_0 v_3 + (1 - v_2) v_4, \\ u(5) &= (1 - v_2) v_4 + v_2 v_5, \\ &\vdots \end{aligned} \tag{32}$$

and in this case, these are equivalent to

$$\begin{aligned} u(0) &= v_0, \\ u(1) &= v_0, \\ u(2) &= v_2, \quad v_2 \le 1 - v_0, \\ u(3) &= v_2 + v_3, \quad v_2 \le 1 - v_0,\ v_3 \le v_0, \\ u(4) &= v_3 + v_4, \quad v_3 \le v_0,\ v_4 \le 1 - v_2,\ 0 \le v_3 + v_4 \le 1, \\ u(5) &= v_4 + v_5, \quad v_4 \le 1 - v_2,\ v_2 \le v_5,\ 0 \le v_4 + v_5 \le 1, \\ &\vdots \end{aligned} \tag{33}$$

We explain $u(2) = v_2$, and $v_2 \le 1 - v_0$ as an example. If $v_0 = 1$, then $v_2 \le 0$ holds. Since v_2 is binary, we can obtain $v_2 = 0$; that is, $u(2) = 0$. If $v_0 = 0$, then $v_2 \le 1$ holds, and we can obtain $u(2) = v_2$. That is, $u(2)$ can take on either 0 or 1. From the previous discussion, we see that a pair of $u(2) = v_2$ and $v_2 \le 1 - v_0$ is equivalent to $u(2) = (1 - v_0)v_2$. Thus, we can obtain the forms of (30) and (31). In the case of $N = 5$ (N is the control time in Problem 1), (30) can be obtained as

$$\begin{aligned} &\underbrace{\begin{bmatrix} u(0) \\ u(1) \\ u(2) \\ u(3) \\ u(4) \end{bmatrix}}_{\overline{u}} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}}_{V^1} \underbrace{\begin{bmatrix} v_0 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}}_{\overline{v}}, \\ &\underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 1 & 1 \end{bmatrix}}_{W^1} \underbrace{\begin{bmatrix} v_0 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}}_{\overline{v}} \le \underbrace{\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}}_{W^2}, \end{aligned} \tag{34}$$

where $V^2 = 0$. We remark that, in a general case, the product such as $z = v_0 v_2$ must be transformed into linear inequalities by using Lemma 7.

Next, consider the case of $L_u = 0$, $W_u^1 = 3$, and $W_u^0 = 3$. Then, we can obtain the following:

$$\begin{aligned} u(0) &= v_0, \\ u(1) &= 0, \\ u(2) &= 0, \\ u(3) &= v_3, \\ u(4) &= 0, \\ u(5) &= 0, \\ &\vdots \end{aligned} \tag{35}$$

and this case is one of the simplest cases. In the case of $N = 5$ (N is the control time in Problem 1), (30) can be obtained as

$$\underbrace{\begin{bmatrix} u(0) \\ u(1) \\ u(2) \\ u(3) \\ u(4) \end{bmatrix}}_{\overline{u}} = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}}_{V^1} \underbrace{\begin{bmatrix} v_0 \\ v_3 \end{bmatrix}}_{\overline{v}}, \tag{36}$$

where $V^2 = 0$, $W^1 = 0$, and $W^2 = 0$.

Finally, the cost function (12) is rewritten as

$$J = \overline{Q}\overline{x} + \overline{R}\overline{u} + \overline{R}_s\overline{u}_s, \tag{37}$$

where $\overline{Q} = [Q \cdots Q\ Q_f]$, $\overline{R} = [R \cdots R]$, and $\overline{R}_s = [R_s \cdots R_s]$. By substituting (25), (30), and (31) into (28) and (37), Problem A is equivalent to the following ILP problem.

Problem B.

$$\begin{aligned}
&\text{find} && \overline{v}, \overline{v}_s, \overline{z}, \\
&\text{min} && \left(\overline{R} + \overline{QB}_u\right) V^1 \overline{v} + \left(\overline{R}_s + \overline{QB}_s\right) V_s^1 \overline{v}_s + \overline{QB}_z \overline{z} \\
&&& + \overline{QA} x_0, \\
&\text{subject to} && -\overline{C}\overline{z} \le -\ln \rho \\
&&& \left(\overline{F}_u + \overline{EB}_u\right) V^1 \overline{v} + \left(\overline{F}_s + \overline{EB}_s\right) V_s^1 \overline{v}_s + \left(\overline{F}_z + \overline{EB}_z\right) \overline{z} \\
&&& \le \overline{G} - \overline{F}_u V^1 - \overline{F}_s V_s^1 - \overline{E}\left(\overline{A} x_0 + \overline{B}_u V^2 + \overline{B}_s V_s^2\right), \\
&&& W^1 \overline{v} \le W^2, \\
&&& W_s^1 \overline{v}_s \le W_s^2.
\end{aligned} \tag{38}$$

Problem B can be solved by a suitable solver such as IBM ILOG CPLEX [33].

3. Results and Discussion

In this section, we show numerical simulations. First, we consider the WNT5A network [14]. Next, in order to evaluate the proposed method from the viewpoint of the computation time, we consider an artificial example.

3.1. WNT5A Network. The gene regulatory network with the gene WNT5A is related to melanoma, and it has been extensively studied (see, e.g., [27]). The BN model $x(k+1) = f_a(x(k))$ of the WNT5A network is given by the following:

$$\begin{aligned}
x_1(k+1) &= \neg x_6(k), \\
x_2(k+1) &= \left(\neg x_2(k) \wedge x_4(k) \wedge x_6(k)\right) \\
&\quad \vee \left\{\neg x_2(k) \wedge \left(x_4(k) \vee x_6(k)\right)\right\}, \\
x_3(k+1) &= \neg x_7(k), \\
x_4(k+1) &= x_4(k), \\
x_5(k+1) &= x_2(k) \vee \neg x_7(k), \\
x_6(k+1) &= x_3(k) \vee x_4(k), \\
x_7(k+1) &= \neg x_2(k) \vee x_7(k),
\end{aligned} \tag{39}$$

where the concentration level (high or low) of the gene WNT5A is denoted by x_1, the concentration level of the gene pirin by x_2, the concentration level of the gene S100P is denoted by x_3, the concentration level of the gene RET1 is denoted by x_4, the concentration level of the gene MART1 is denoted by x_5, the concentration level of the gene HADHB is denoted by x_6, and the concentration level of the gene STC2 is denoted by x_7. See [14] for further details. In a WNT5A network, it is important to inhibit the concentration level of the gene WNT5A [27].

The optimal control problem is formulated. For simplicity, we consider only the structural control input. Then, suppose that the number of the structural control inputs is two. If $u_1^s(k) = 1$, then the system is given as:

$$x(k+1) = \begin{cases} f_a(x(k)) & \text{with the probability 0.8,} \\ x(k) & \text{with the probability 0.2.} \end{cases} \tag{40}$$

If $u_2^s(k) = 1$, then the system is given as:

$$x(k+1) = \begin{cases} f_a(x(k)) & \text{with the probability 0.1,} \\ x(k) & \text{with the probability 0.9.} \end{cases} \tag{41}$$

The case of $u_1^s(k) = 1$ corresponds to the situation such that the dynamics of the WNT5A network, that is, $x(k+1) = f_a(x(k))$, are selected with high probability. The case of $u_2^s(k) = 1$ corresponds to the situation such that the state is not changed; that is, $x(k+1) = x(k)$ is selected with high probability. From the previous setting, $r_{1,1} = 0.8$, $r_{1,2} = 0.2$, $r_{2,1} = 0.1$, and $r_{2,2} = 0.9$. For this WNT5A network with structural control inputs, consider solving Problem 1. Q, Q_f, and R_s in Problem 1 are given as $Q = [1\ 0\ 0\ 0\ 0\ 0\ 0]$, $Q_f = [10\ 0\ 0\ 0\ 0\ 0\ 0]$, and $R_s = [0\ 0]$, respectively. The initial state is given as $x_0 = [1\ 0\ 1\ 0\ 1\ 0\ 0]^T$. In addition, the control time N in Problem 1 is given by $N = 5$. Finally, the constraint on duration of drug effectiveness is imposed for only $u_2^s(k)$. The parameters $L_{u_2^s}$, $W_{u_2^s}^1$, and $W_{u_2^s}^0$ are given as $L_{u_2^s} = 1$, $W_{u_2^s}^1 = 3$, and $W_{u_2^s}^0 = 2$, respectively. Thus, we can obtain the ILP problem (Problem B), where the dimension of binary variables is 130 and the number of inequalities is 264.

We show the computational result. Let $\underline{J}^*$ denote the optimal value of the lower bound of a given cost function in Problem 1. Let $\overline{J}^*$ denote the upper bound of the cost function derived by using the optimal control input. First, consider the case of $\rho = 10^{-5}$. Then, we can obtain $\underline{J}^* = 2$ and $\overline{J}^* = 15$, and $u^s(k)$ is obtained as

$$u^s(0) = \cdots = u^s(4) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{42}$$

Noting that $r_{2,1} = 0.1$ and $\rho = 10^{-5} (= 0.1^5)$, all combinations of the Boolean functions are considered, and the value of ρ is not appropriate. In particular, $\overline{J}^* = 15$ implies that $x_1(k) = 1$, $k = 0, 1, \ldots, 5$, and is the trivial upper bound.

Next, consider the case of $\rho = 0.2$. Then, we can obtain $\underline{J}^* = \overline{J}^* = 4$, and $u^s(k)$ is obtained as

$$u^s(0) = u^s(1) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad u^s(2) = u^s(3) = u^s(4) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{43}$$

From the obtained inputs, we see that the system is controlled by switching two discrete probability distributions, and the obtained inputs satisfy the constraint on duration of drug effectiveness. Noting that the trivial value of $\overline{J}^*$ is 15, we see that in this case the effectiveness of control synthesis is clear.

Finally, we discuss the computation time for solving Problem 1. The computation time of the ILP problem was less than 20 [msec], where we used IBM ILOG CPLEX 11.0 as an ILP solver on the computer with Windows Vista 32-bit, the Intel Core 2 Duo CPU 3.0 GHz, and the 4 GB memory. Since the WNT5A network considered here is small size, Problem 1 can be solved fast.

3.2. Artificial Example. In order to evaluate the computation time for solving Problem 1, we consider one artificial example of a BN with 15 states and 3 control inputs. We stress that the existing method [11, 12, 17–19, 21] cannot be applied to such a BN. This is because it is necessary to compute the state transition diagram such as that in Figure 2, that is, the transition probability matrix with $2^n \times 2^n$. In naive implementation using MATLAB [34], matrices with $2^{15} \times 2^{15}$ cannot be created due to memory consumption, where we used the computer described previously.

The optimal control problem is formulated. In this example, we consider 3 control inputs and 2 structural control inputs. If $u_1^s(k) = 1$, then the system is given as follows:

$$x(k+1) = \begin{cases} f_1(x(k), u(k)) & \text{with the probability } 0.8, \\ f_2(x(k), u(k)) & \text{with the probability } 0.2. \end{cases} \tag{44}$$

If $u_2^s(k) = 1$, then the system is given as follows:

$$x(k+1) = \begin{cases} f_1(x(k), u(k)) & \text{with the probability } 0.2, \\ f_2(x(k), u(k)) & \text{with the probability } 0.8. \end{cases} \tag{45}$$

The Boolean function f_1 is given by the following:

$$\begin{aligned}
x_1(k+1) &= x_1(k) \wedge \neg x_6(k) \vee u_3(k), \\
x_2(k+1) &= \neg x_4(k) \wedge u_1(k) \vee u_3(k), \\
x_3(k+1) &= x_5(k) \wedge u_1(k) \vee \neg x_{10}(k) \wedge x_{12}(k) \wedge u_3(k), \\
x_4(k+1) &= x_2(k) \wedge x_5(k) \wedge \neg u_1(k) \vee \neg x_{14}(k), \\
x_5(k+1) &= \neg u_1(k) \wedge x_6(k) \wedge x_7(k) \vee x_{12}(k) \wedge x_{14}(k), \\
x_6(k+1) &= x_1(k) \wedge x_6(k) \wedge x_{10}(k) \vee \neg x_{15}(k), \\
x_7(k+1) &= x_6(k) \wedge x_7(k) \wedge x_8(k) \vee u_2(k) \wedge \neg u_3(k), \\
x_8(k+1) &= x_5(k) \wedge \neg u_1(k) \vee x_{10}(k) \wedge u_2(k) \wedge x_{13}(k), \\
x_9(k+1) &= x_3(k) \wedge u_1(k) \vee \neg x_8(k) \wedge x_{11}(k), \\
x_{10}(k+1) &= x_6(k), \\
x_{11}(k+1) &= x_6(k) \wedge x_{10}(k) \vee \neg u_2(k) \wedge u_3(k), \\
x_{12}(k+1) &= x_{12}(k) \wedge \neg x_{15}(k), \\
x_{13}(k+1) &= \neg u_3(k), \\
x_{14}(k+1) &= \neg x_{14}(k) \wedge u_3(k), \\
x_{15}(k+1) &= x_{14}(k) \wedge x_{15}(k).
\end{aligned} \tag{46}$$

The Boolean function f_2 is given by the following:

$$\begin{aligned}
x_1(k+1) &= x_2(k) \wedge x_4(k) \wedge \neg x_8(k), \\
x_2(k+1) &= \neg x_2(k) \wedge x_3(k) \vee u_3(k), \\
x_3(k+1) &= x_1(k) \vee \neg x_2(k) \wedge x_3(k) \wedge x_4(k), \\
x_4(k+1) &= \neg x_1(k) \wedge x_2(k) \wedge u_1(k) \vee x_{14}(k), \\
x_5(k+1) &= \neg u_2(k) \wedge x_{13}(k) \wedge x_{14}(k) \wedge x_{15}(k), \\
x_6(k+1) &= \neg x_2(k) \vee x_5(k) \wedge u_1(k) \wedge \neg u_3(k), \\
x_7(k+1) &= u_1(k) \wedge x_{13}(k), \\
x_8(k+1) &= x_5(k) \wedge x_{13}(k), \\
x_9(k+1) &= \neg x_6(k), \\
x_{10}(k+1) &= x_2(k) \wedge u_2(k) \wedge \neg x_{12}(k) \wedge u_3(k), \\
x_{11}(k+1) &= \neg x_5(k) \wedge u_1(k) \vee \neg x_{15}(k) \wedge u_3(k), \\
x_{12}(k+1) &= x_7(k) \wedge x_3(k), \\
x_{13}(k+1) &= x_7(k) \wedge u_2(k) \wedge u_3(k), \\
x_{14}(k+1) &= x_{12}(k) \vee x_{14}(k) \wedge u_3(k), \\
x_{15}(k+1) &= x_8(k).
\end{aligned} \tag{47}$$

From the previous setting, $r_{1,1} = 0.8$, $r_{1,2} = 0.2$, $r_{2,1} = 0.2$, $r_{2,2} = 0.8$ hold. In Problem 1, Q, Q_f, R, and R_s are given as $Q = [1 \ \cdots \ 1]$, $Q_f = [10 \ \cdots \ 10]$, $R = [1 \ 0 \ 1]$, and $R_s = [0 \ 0]$,

respectively. The initial state and the parameter ρ are given as $x_0 = [1 \ \cdots \ 1]^T$ and $\rho = 10^{-4}$, respectively. The constraint on duration of drug effectiveness is imposed for $u_1(k)$ and $u_2^s(k)$. For u_1, the parameters L_{u_1}, $W_{u_1}^1$, and $W_{u_1}^0$ are given as $L_{u_1} = 0$, $W_{u_1}^1 = 3$, and $W_{u_1}^0 = 3$, respectively. For u_2^s, the parameters $L_{u_2^s}$, $W_{u_2^s}^1$, and $W_{u_2^s}^0$ are given as $L_{u_2^s} = 1$, $W_{u_2^s}^1 = 3$, and $W_{u_2^s}^0 = 2$, respectively.

Next, we discuss the computation time. Consider the two cases of $N = 10$ and $N = 20$. Then, in the ILP problem (Problem B) obtained, the dimension of binary variables is 1420 for $N = 10$ and 2840 for $N = 20$, and the number of inequalities is 3381 for $N = 10$ and 6731 for $N = 20$. In the case of $N = 10$, the computation time of the ILP problem was 96 [sec], where we used the computer described previously. In the case of $N = 20$, the computation time of the ILP problem was 238 [sec]. We remark that BNs with such a size are large scale in control problems of gene regulatory networks. Thus, we conclude that Problem 1 can be solved within the practical computation time.

4. Conclusions

In this paper, we have proposed a Boolean network (BN) model with two types of the control inputs and an optimal control method. By using this model, several situations in control of gene regulatory networks can be modeled. To model more realistic situations, duration of drug effectiveness has also been introduced. Since duration is given for each control input, effectiveness of multiple drugs can be evaluated. Furthermore, for this BN model, the optimal control problem has been formulated, and this problem is reduced to an integer linear programming problem. Finally, numerical simulations have been shown. The proposed method provides us with a basic in the control theory of gene regulatory networks.

Recently, to simplify state transition diagrams such as that in Figure 2, a stochastic Boolean network has been proposed in [35]. The authors proposed in [36] a similar method using polynomial optimization. In addition, to simplify a given Boolean network, the Karnaugh map realization of a Boolean network has been proposed in [37]. These methods are useful for reducing the computational burden. It is one of the future works to consider the control problem with duration of drug effectiveness based on these methods.

Acknowledgment

This work was partially supported by the Grant-in-Aid for Young Scientists (B) no. 23760387.

References

[1] H. Santos-Rosa and C. Caldas, "Chromatin modifier enzymes, the histone code and cancer," *European Journal of Cancer*, vol. 41, no. 16, pp. 2381–2402, 2005.

[2] A. Milias-Argeitis, S. Summers, J. Stewart-Ornstein et al., "*In silico* feedback for *in vivo* regulation of a gene expression circuit," *Nature Biotechnology*, vol. 29, no. 12, pp. 1114–1116, 2011.

[3] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, 1969.

[4] H. D. Jong, "Modeling and simulation of genetic regulatory systems: a literature review," *Journal of Computational Biology*, vol. 9, no. 1, pp. 67–103, 2002.

[5] T. Akutsu, M. Hayashida, W. Ching, and M. K. Ng, "Control of Boolean networks: hardness results and algorithms for tree structured networks," *Journal of Theoretical Biology*, vol. 244, no. 4, pp. 670–679, 2007.

[6] S. Azuma, E. Yanagisawa, and J. Imura, "Controllability analysis of biosystems based on piecewise-affine systems approach," *IEEE Transactions on Automatic Control*, vol. 53, pp. 139–152, 2008.

[7] K. Kobayashi, J. Imura, and K. Hiraishi, "Polynomial-time algorithm for controllability test of a class of Boolean biological networks," *Eurasip Journal on Bioinformatics and Systems Biology*, vol. 2010, Article ID 210685, 12 pages, 2010.

[8] C. J. Langmead and S. K. Jha, "Symbolic approaches for finding control strategies in boolean networks," *Journal of Bioinformatics and Computational Biology*, vol. 7, no. 2, pp. 323–338, 2009.

[9] A. Mochizuki, "An analytical study of the number of steady states in gene regulatory networks," *Journal of Theoretical Biology*, vol. 236, no. 3, pp. 291–310, 2005.

[10] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2002.

[11] B. Faryabi, G. Vahedi, J. F. Chamberland, A. Datta, and E. R. Dougherty, "Intervention in context-sensitive probabilistic boolean networks revisited," *Eurasip Journal on Bioinformatics and Systems Biology*, vol. 2009, Article ID 360864, 13 pages, 2009.

[12] R. Pal, A. Datta, M. L. Bittner, and E. R. Dougherty, "Intervention in context-sensitive probabilistic Boolean networks," *Bioinformatics*, vol. 21, no. 7, pp. 1211–1218, 2005.

[13] I. Shmulevich, E. R. Dougherty, and W. Zhang, "Control of stationary behavior in probabilistic Boolean networks by means of structural intervention," *Journal of Biological Systems*, vol. 10, no. 4, pp. 431–445, 2002.

[14] Y. Xiao and E. R. Dougherty, "The impact of function perturbations in Boolean networks," *Bioinformatics*, vol. 23, no. 10, pp. 1265–1273, 2007.

[15] M. R. Yousefi, A. Datta, and E. R. Dougherty, "Optimal intervention strategies for therapeutic methods with fixed-length duration of drug effectiveness," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4930–4944, 2012.

[16] X. Chen, T. Akutsu, T. Tamura, and W. K. Ching, "Finding optimal control policy in probabilistic Boolean networks with hard constraints by using integer programming and dynamic programming," *International Journal of Data Mining and Bioinformatics*, vol. 7, no. 3, pp. 321–343, 2013.

[17] W. K. Ching, S. Q. Zhang, Y. Jiao, T. Akutsu, N. K. Tsing, and A. S. Wong, "Optimal control policy for probabilistic Boolean networks with hard constraints," *IET Systems Biology*, vol. 3, no. 2, pp. 90–99, 2009.

[18] A. Datta, A. Choudhary, M. L. Bittner, and E. R. Dougherty, "External control in Markovian genetic regulatory networks," *Machine Learning*, vol. 52, no. 1-2, pp. 169–191, 2003.

[19] A. Datta, A. Choudhary, M. L. Bittner, and E. R. Dougherty, "External control in Markovian genetic regulatory networks:

the imperfect information case," *Bioinformatics*, vol. 20, no. 6, pp. 924–930, 2004.

[20] S. Z. Denic, B. Vasic, C. D. Charalambous, and R. Palanivelu, "Robust control of uncertain context-sensitive probabilistic Boolean networks," *IET Systems Biology*, vol. 3, no. 4, pp. 279–295, 2009.

[21] R. Pal, A. Datta, and E. R. Dougherty, "Optimal infinite-horizon control for probabilistic Boolean networks," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2375–2387, 2006.

[22] K. Kobayashi and K. Hiraishi, "An integer programming approach to optimal control problems in context-sensitive probabilistic Boolean networks," *Automatica*, vol. 47, no. 6, pp. 1260–1264, 2011.

[23] K. Kobayashi and K. Hiraishi, "Probabilistic control of Boolean networks with multiple dynamics: Towards control of gene regulatory," in *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, pp. 4371–4376, Orlando, Fla, USA, 2011.

[24] S. Marshall, L. Yu, Y. Xiao, and E. R. Dougherty, "Inference of a probabilistic boolean network from a single observed temporal sequence," *Eurasip Journal on Bioinformatics and Systems Biology*, vol. 2007, Article ID 32454, 15 pages, 2007.

[25] M. Chaves, "Methods for qualitative analysis of genetic networks," in *Proceedings of the European Control Conference*, pp. 671–676, 2009.

[26] L. Tournier and M. Chaves, "Uncovering operational interactions in genetic networks using asynchronous Boolean dynamics," *Journal of Theoretical Biology*, vol. 260, no. 2, pp. 196–209, 2009.

[27] A. T. Weeraratna, Y. Jiang, G. Hostetter et al., "Wnt5a signaling directly affects cell motility and invasion of metastatic melanoma," *Cancer Cell*, vol. 1, no. 3, pp. 279–288, 2002.

[28] F. Adamek, M. Sobotka, and O. Stursberg, "Stochastic optimal control for hybrid systems with uncertain discrete dynamics," in *Proceedings of the 4th IEEE Conference on Automation Science and Engineering (CASE '08)*, pp. 23–28, Arlington, Va, USA, August 2008.

[29] A. Bemporad and S. Di Cairano, "Optimal control of discrete hybrid stochastic automata," in *Hybrid Systems: Computation and Control*, vol. 3414 of *Lecture Notes in Computer Science*, pp. 417–432, 2005.

[30] K. Kobayashi, K. Matou, and K. Hiraishi, "Probabilistic-constrained optimal control of a class of stochastic hybrid systems," *International Journal of Control, Automation and Systems*, vol. 10, no. 5, pp. 897–904, 2012.

[31] H. P. Williams, *Model Building in Mathematical Programming*, John Wiley & Sons, New York, NY, USA, 4th edition, 1999.

[32] T. M. Cavalier, P. M. Pardalos, and A. L. Soyster, "Modeling and integer programming techniques applied to propositional calculus," *Computers and Operations Research*, vol. 17, no. 6, pp. 561–570, 1990.

[33] IBM ILOG CPLEX Optimizer, http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/index.html.

[34] MathWorks MATLAB, http://www.mathworks.com/.

[35] J. Liang and J. Han, "Stochastic Boolean networks: an efficient approach to modeling gene regulatory networks," *BMC Systems Biology*, vol. 6, article 113, 20 pages, 2012.

[36] K. Kobayashi and K. Hiraishi, "Optimal control of probabilistic Boolean networks using polynomial optimization," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 95, no. 9, pp. 1512–1517, 2012.

[37] R. K. Layek, A. Datta, and E. R. Dougherty, "From biological pathways to regulatory networks," *Molecular BioSystems*, vol. 7, no. 3, pp. 843–851, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 414069, 7 pages
http://dx.doi.org/10.1155/2013/414069

Research Article

Method for Rapid Protein Identification in a Large Database

Wenli Zhang[1,2,3] and Xiaofang Zhao[1]

[1] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China*
[2] *State Key Laboratory of Computer Architecture, ICT, CAS, Beijing 100190, China*
[3] *Graduate University of Chinese Academy of Sciences, Beijing 100049, China*

Correspondence should be addressed to Wenli Zhang; zhangict@gmail.com

Received 17 May 2013; Revised 10 July 2013; Accepted 14 July 2013

Academic Editor: Lei Chen

Protein identification is an integral part of proteomics research. The available tools to identify proteins in tandem mass spectrometry experiments are not optimized to face current challenges in terms of identification scale and speed owing to the exponential growth of the protein database and the accelerated generation of mass spectrometry data, as well as the demand for nonspecific digestion and post-modifications in complex-sample identification. As a result, a rapid method is required to mitigate such complexity and computation challenges. This paper thus aims to present an open method to prevent enzyme and modification specificity on a large database. This paper designed and developed a distributed program to facilitate application to computer resources. With this optimization, nearly linear speedup and real-time support are achieved on a large database with nonspecific digestion, thus enabling testing with two classical large protein databases in a 20-blade cluster. This work aids in the discovery of more significant biological results, such as modification sites, and enables the identification of more complex samples, such as metaproteomics samples.

1. Introduction

Proteomics is an emerging discipline based on the human genome project. Proteomics primarily aims to determine the presence and quantity of proteins. Similar to gene sequencing, the identification of protein sequences is important to facilitate a systematic understanding of key biological knowledge including protein structure, function, and evolutionary relationships.

The basic principle of protein identification [1–3] based on mass spectrometry is to deduce protein amino acid sequences based on the mass/charge ratio of digested peptides and peptide fragments. The main strategies employed for protein identification by mass spectrometry include database searching, de novo sequencing, and peptide sequence tag. Among these strategies, database searching is the most popular. In this approach, experimental spectra are compared with theoretical spectra from database peptides to identify the best fit.

Commercial search engines such as Mascot [4] and SEQUEST [5] as well as open-source search engines such as X!Tandem [6], OMSSA [7, 8] and pFind (one of the first available protein identification software designed and developed in China) [9, 10] are very popular. With the rapid development of high-throughput biological mass spectrometry along with the exponential expansion of the protein database [11], computing technologies are presented with both an opportunity and a challenge to serve a notable function in solving biological problems. Commonly used protein identification software, such as Mascot, SEQUEST, and X!Tandem, currently possesses implemented cluster or cloud versions [12–17].

The main idea of parallel versions is to split the input spectra and to process each subset independently, which limits the identification speed to a stand-alone computer. Thus, problems arise once the index space becomes larger than the main memory size. Moreover, the types of modifications and the occurrence of digestion are unknown in most cases. If all modifications in the Unimod [18] database are specified, all search engines would not work properly. To complete the task within an acceptable duration under specific machine resources, including CPU frequency and

memory size, protein identification has to employ restricted search, in which a small sequence database is used to specify a limited number of specific or semispecific enzymatic digestions as well as a limited number of commonly used modifications to limit the use of time and space. This condition makes most of the spectra produced by a mass spectrometer in proteomics experiments difficult to interpret. Only 5% to 30% of spectra can usually be identified [19]. An important reason is that several digestion methods and unknown or unexpected modifications likely exist. Thus, wrong candidate peptides will affect the subsequent protein identification process. Otherwise, the interpretation rate may double.

Thus, the development and analysis of a rapid identification method that can support a large protein database with any type of digestion and modification is urgently needed to facilitate protein deep analysis, like the expansion of metaproteomics [20].

Therefore, this paper designs and develops a distributed protein identification tool based on pFind to support nonspecific enzymatic digestion on a large database with unrestricted modification. The goal is to design and implement a practical, scalable, and efficient system to identify proteins rapidly and to identify more modification sites on a large protein sequence database. Actually, pFind has been proven more accurate than similar tools. In this work, we will focus on the acceleration of pFind to achieve greater identification speed.

The remainder of the paper is organized as follows. Section 2 describes the materials and methods used in this work. Section 3 shows the evaluation of the proposed method and discusses the results. Finally, the conclusion and future work are given in Section 4.

2. Materials and Methods

In this section, we provide the background of open protein identification and discuss our optimization work.

2.1. Basic Principle and Trends. Database searching is frequently employed to identify unknown amino acid sequences of peptides/proteins with high throughput. The main idea of this approach is shown in Figure 1. In this approach, proteins in the sample are digested into a peptide mixture. A mass spectrometer is then used to produce tandem mass (briefly as MS/MS) spectra, which are to be identified to query in a known protein database. On the other hand, the theoretical MS/MS spectra are predicted according to enzymatic digestion rules, that is, simulated digestion, based on peptide sequences from a known protein database. The most common method is to use a search algorithm to identify peptides by correlating experimental and theoretical MS/MS data generated from possible peptides in the protein sequence database through simulated digestion.

Simulated digestion theoretically refers to digestion based on the known protein sequences and enzyme specificity. Simulated digestion generally includes three types: specific, semispecific, and nonspecific. In specific digestion, protein sequence hydrolysis only occurs at a specific amino acid. For example, trypsin will cut polypeptide chain after lysine (K) or arginine (R) under the premise that proline (P) is not the next residue. In semispecific digestion, hydrolysis only occurs at some particular amino acids in one terminus, whereas the other will be disconnected at any amino acid. Nonspecific digestion occurs when disconnection occurs at any amino acid in both termini, that is, equivalent to any substrings of the amino acid sequence. Nonspecific digestion is usually avoided, especially when identifying on large databases, because of its running time and high memory demand.

Protein posttranslational modification on eukaryotic cells is of great significance in presenting the protein structure and function and explaining the mechanisms of major diseases. Over 1000 kinds of modifications are currently available in the database. Searching for an excessive number of modification types is thus unrealistic. Therefore, not more than 10 types of variable modification could be assigned for current mature search engines, such as SEQUEST and Mascot, which obviously cannot meet actual needs.

The types of digestion and modification are generally restricted. As the mainstream approach to database searching, the most significant advantage of the restricted method is its reduction of the scale of candidate peptides because this method assigns some factors depending on experience. However, individual experience is not always accurate. Despite appearing to be a perfect solution for the open method to support a large database, any type of digestion, and any type of modification, the search speed has restricted the development of this approach because of the large search space.

Meanwhile, the exponential growth of the protein database, the rapid generation of mass spectrometry data, and the requirement for nonspecific digestion and postmodifications in complex-sample identification also pose a significant challenge on the identification scale and speed. The size of the genomic and protein sequence database grows exponentially, exceeding even Moore's Law in terms of the requirements for computing hardware. As shown in Figure 2, the increasing trend of the protein database UniProtKB/TrEMBL is a representative case.

The opinion of Patterson in [21] is curt and to the point: "...our ability to generate data now outstrips our ability to analyze it."

2.2. Optimization Approaches. Figure 3 shows the typical identification workflow, usually including (A.) spectra preprocess, (B.) build index, and (C.) search index to identify.

Based on an analysis of the protein identification process, three main methods are available to accelerate search engines at present. First, preprocess protein database can be secured, such that a more efficient index structure can be constructed. This design is a high-performance solution in a small-scale protein database. However, for protein databases with a scale of tens of MB, the index created by this method has to use several GB of storage space. Moreover, building an index for a large database is time consuming. Second, efficient search algorithms or technologies can be presented for search engine acceleration, such as an inverted index. Third, a parallel

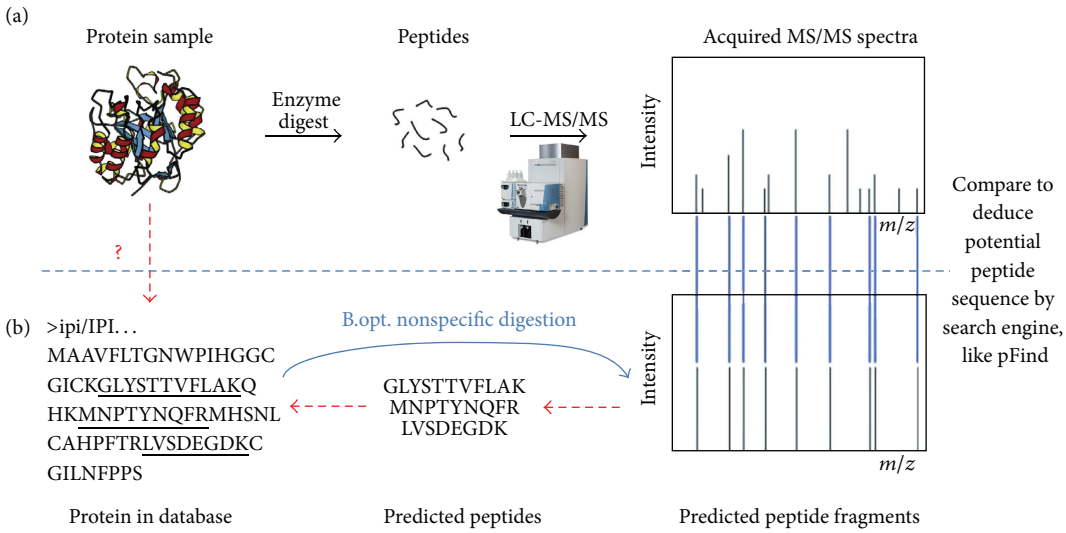


Figure 1: Principle of protein identification using MS/MS data.

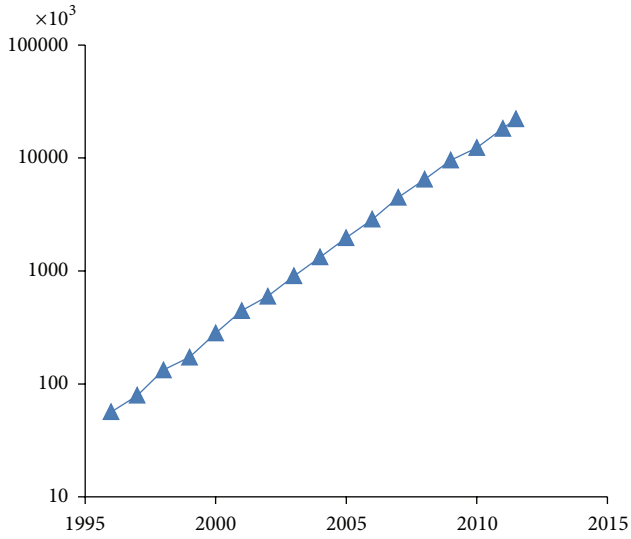


Figure 2: Increasing trend of protein database UniProtKB/TrEMBL (in entries).

search can be conducted to improve query efficiency in clusters.

With the popularity of cluster applications, successive parallel versions of some mainstream protein identification tools have been introduced. Most of these versions are based on the simple task of partitioning technology among spectra. As opposed to a stand-alone version, the identification speed can be increased several times. However, online digestion and fragmentation cannot be avoided for each retrieval.

To prepare for large-scale protein identification, identification on a large-scale protein database, with any type of restriction and modification, must be supported. Based on pFind, we designed a scalable and efficient system to meet the rapid identification needs.

pFind is one of the first protein identification software designed and developed in China. In terms of accuracy and speed, pFind has reached the level of international mainstream commercial software, such as SEQUEST and Mascot. As early as 2008, pFind has participated in the international evaluation on protein identification organized by the Association of Biomolecular Resource Facilities and has demonstrated strong performance in terms of identification accuracy and false positive rate control capability. pFind is currently the only protein search engine devoted to first-line research that was developed in China and is used by hundreds of groups around the world, including Duke University and MIT.

Search engines usually need to digest protein sequences online as well as filter peptides according to the mass error, which may add unnecessary overhead. When spectrum data are large, the overhead for online digestion will unnecessarily increase because the process will have to be performed repeatedly for each batch. If the index space can be guaranteed, nonspecific digestion on the protein database would significantly improve efficiency.

In nonspecific digestion, the protein sequence may cleave at any amino acid to form peptide fragments, which indicates that the hydrolysis of peptide can be any substring of the protein sequence. For each protein sequence, all subsequences within a specified length and mass range are generated. This optimization is step B.opt., as shown in Figure 1. Step B.2, as shown in Figure 3, can be handled in one way by nonspecific digestion for all enzymes and even offline. This condition not only lays the foundation for acceleration but also reduces the dependence on expertise.

In this work, we built a reverted index of peptide fragments generated by nonspecific digestion in mass prior to spectrum queries. The index generation process helps eliminate the overhead of simulated digestion during a search while naturally supporting the retrieval of nonspecific digestion. All subsequences generated from the protein

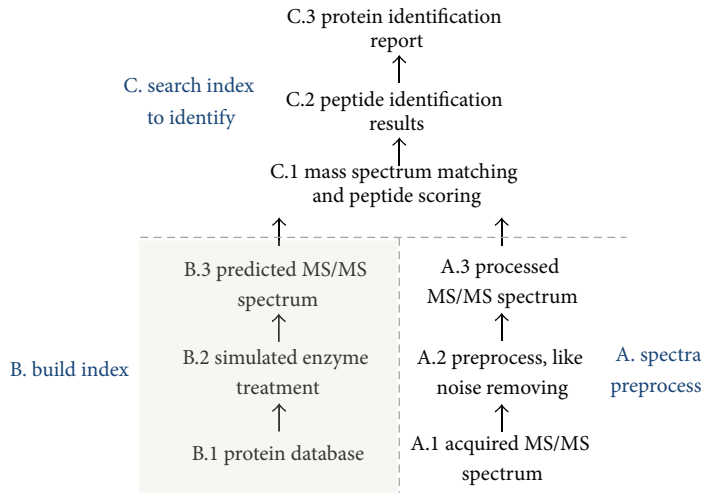


Figure 3: The typical workflow of protein identification by database searching.

database are sorted by their masses in ascending order, and an index table is constructed in which the key is the mass of each peptide represented by three integers: protein ID, start position, and amino acid length. Therefore, all index terms are recorded with equal lengths. Given an explicit mass or mass range to be queried, the time complexity of finding the first valid position in the datasheet is $O(1)$. Some range searches obtained from spectrum peaks with mass tolerance can then quickly retrieve the unique index. Undoubtedly, this approach will simplify identification process and will save both the index build and search time.

Modification identification is another time-consuming process, and unrestricted posttranslational modification identification remains inadequate. InsPecT describes an unrestrictive PTM search algorithm that searches for any possible type of modification at once in a "blind search" mode, which does not depend on any given modification list. Such ideas can be used to identify more types of modifications, but its operation speed will be affected to a certain extent. By contrast, the number of modification types in Unimod is almost complete; at least in the vast majority of mass spectrometry experiments, Unimod is sufficient. As of Mascot version 2.3, support for Error Tolerant, an earlier proposed open identification method, is provided to iterate search over conventional identification. Moreover, this approach only supports semidigestion or all modifications in the Unimod database.

Using pFind with the DeltAMT algorithm [22], the Beijing Proteome Research Center identified core fucosylation (CF) modification. Over 100 CF glycoproteins and CF modification sites were identified from plasma samples of human liver cancer, the greatest number among all reports. The scale of identification results indicates significant progress in finding potential biomarkers. The discovery of a large number of modifications is of great significance for follow-up research and would aid in the early discovery of cancer markers [23]. Therefore, we have reasons to believe that accelerating pFind is an efficient method to accelerate postmodification prediction.

The two-stage method can be used to determine one terminus of the peptide as well as to obtain a smaller number of candidates. The other terminus of the peptide can then be determined, taking the mass difference as the modification to seek in the inverted index of modification that was initially built according to Unimod, based on the smaller number of candidates. The complexity in determining one modification based on mass is $O(1)$. To focus on acceleration, we will concentrate on dealing with nonspecific digestion.

2.3. Data Access Characteristics. In this section, we monitor the data usage of our open protein identification method and verify the robustness of this approach by implementing a measurable identification system based on Hadoop software.

We checked the index and query distributions and obtained the results shown in Figure 4. Basic analysis revealed that the query is mainly concentrated in less than 2000 Dalton; that is, 97% of total queries is within this range. However, only 40% of the data in the database is within 2000 Dalton. Meanwhile, the other 60% of the index will only account for only 3% of the queries, which is largely idle for these parts. The natural imbalance will inevitably cause some hot spots and low efficiency.

At first impression, the process of protein identification over a large-scale protein sequence database is similar to large-scale text information retrieval. Mature technologies in the large-scale Internet, such as Google GFS, map/reduce, and bigtable, can serve as references for protein identification. Such technologies also provide an important template for the construction of a large-scale protein identification cloud system with mechanisms such as distributed storage, load balance, and fault tolerance with the thousands-of-nodes cluster of Google.

However, the above results imply that success in the large-scale Internet with cloud system architecture is unsuitable for large-scale protein identification. This experiment is not similar in nature to experiments conducted using Amazon to distribute spectra [16, 17].

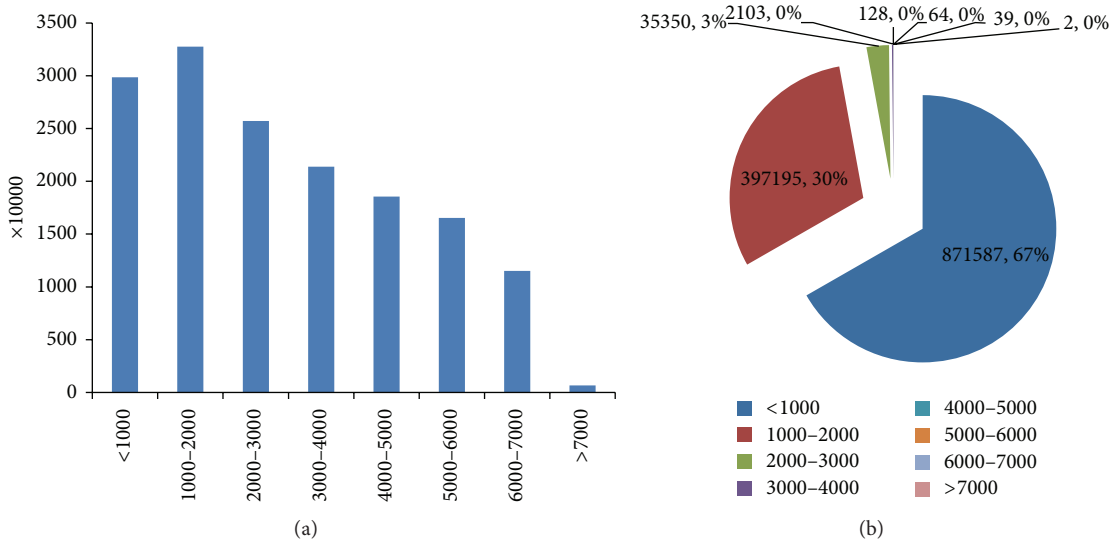


FIGURE 4: (a) Example of index distribution (x-axis is the mass range in Dalton, whereas y-axis is the statistic number). (b) Example of query distribution (legends are mass ranges in Dalton, and numbers shown in the pie include the statistic number and ratio).

Nevertheless, we implement a cloud version of our approach with Apache Hadoop, a popular open-source software framework derived from Google techniques. This approach uses Hadoop map/reduce streaming to build the index for a protein database in advance. The database is then stored into the Hadoop HBase system for searching. The performance was just as expected.

2.4. Distributed Speedup. In this section, we design a distributed acceleration program for digestion-open protein identification based on the above analysis. To cope with the massive computational challenges brought by large-scale databases and nonspecific protein digestion, a parallel protein identification process must be efficiently implemented. As shown in Table 1, the number of peptide fragments generated by nonspecific digestion will increase by approximately 10,000 times compared with the protein number in conservative estimates. This number will potentially require a considerably large index space. Thus, a viable idea to self-design a system based on the map/reduce ideology to achieve unlimited expansion and to control data locality in the index arrangement.

To scale up support for identification on large protein databases, we introduce a method to partition and distribute the protein sequence database to build the index separately among cluster nodes that are parallel among proteins. With a large search space from nonspecific digestion, the index cannot be handled using the common stand-alone computer memory, like for ipi.HUMAN and UNIPROT-SPROT shown in Table 1, not saying a larger database and more complex analysis required. To address this issue, a simple and direct strategy is to partition the protein database into as many subsets as there are processors and then build the index in the memory of each subset independently.

To scale up the system further, the database is partitioned according to the number of processors, and the subdatabase is again partitioned inside the processor according to the computer memory capacity. Within a node, the subdatabase can be partitioned efficiently. Partial identification is then conducted, and all results are collected as spectra. This process will help mitigate computer memory limitations. If speed is not an issue, the large database can be processed on any number of nodes. Moreover, the database is partitioned by accounting for the number of amino acids to balance the load. More detailed analysis can be referred to [24].

Latency hiding is another approach to develop scalable parallel machines. The main technology is data prefetching and multithreading. Data prefetching facilitates the prompt access or transfer of delayed data. Using multi-threading, computation and communication can be overlapped when a thread is computing to store data into the memory in preparation for the subsequent identification. Using this strategy, index-building time can be saved if the subblock index can be generated in advance to be prefetched when needed.

We distribute spectra among CPU cores in multithread to speed up the system. To improve the performance of the spectrum identification later, the spectra are sorted in mass before division into subtask blocks. Thereafter, mass spectra are assigned with close mass to the same sub-task. We achieved linear speedup in 320 processors with the spectrum distribution strategy on a small protein database [25]. Linear results were likewise obtained in 1024 processors in subsequent experiments.

In parallel computing, task scheduling is one of the most critical issues. The scheduling algorithm aims to achieve load balancing among compute nodes. In other words, parallel computing ensures that the tasks among compute nodes can be completed within close time points to minimize the overall running time. For simplicity, we only use the static scheduling according to the current application characterizations. In our

Table 1: Scale change in nonspecific digestion.

DB	Version	Fasta size (MB)	No. of proteins	No. of peptide fragments	Index size (GB)
ipi.HUMAN	3.87	48.71	91,491	1,939,985,477	44.44
UNIPROT-SPROT	201108	240.51	532,234	9,964,797,926	235.51

Note: peptide mass is 300–8000 Dalton. Peptide length is 3–60 amino acids.

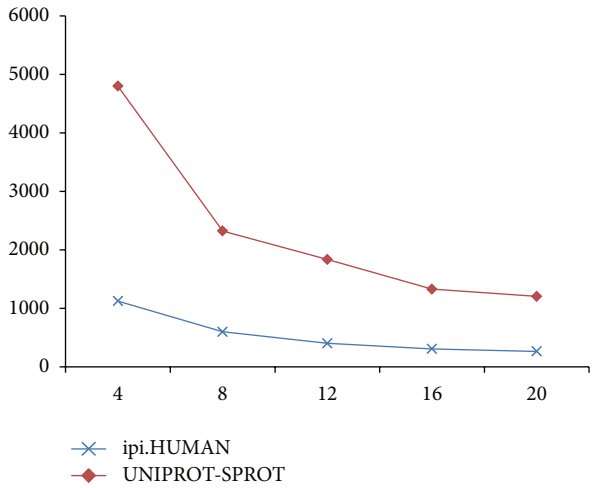


Figure 5: Performance test on the distributed identification system (no. of nodes—seconds).

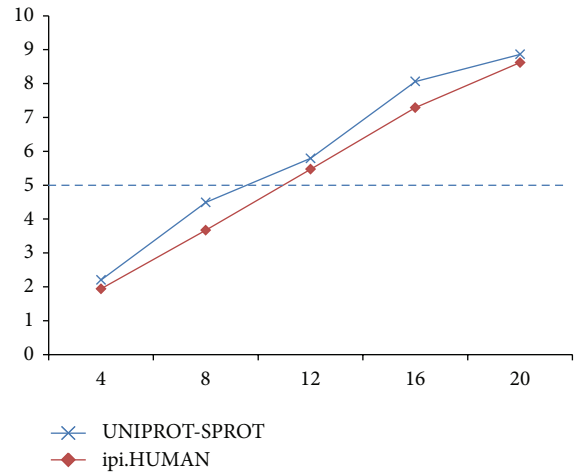


Figure 6: Identification throughput in spectra/s/100M B (no. of nodes—spectra/s/100 MB).

experiments, static scheduling presents good performance in a homogeneous computer system.

3. Results and Discussion

In this section, we show some evaluation results in terms of identification speed and throughput of the distributed acceleration program on a 20-blade cluster.

3.1. Linear Speedup. To evaluate the proposed approach, we tested a 20-blade cluster with an Intel two-quad-core CPU and 8 GB memory. We took the two databases mentioned in Table 1 and a raw file randomly sampled from TTE experiments as input. The raw file was obtained from LTQ-Orbitrap Velos with HCD collision. To save test time, we regularly sampled 1096 spectra as a test case.

Figure 5 shows good scalability in the near linear area from 4 to 20 nodes on both ipi.HUMAN and UNIPROT-SPROT. In our analysis, the system can easily be scaled up to support a larger database. Users can add new nodes in the cluster to meet the growing demand.

3.2. Support for Real-Time Identification over Large Protein Databases. In analyzing the system overhead, we found that over 70% of time is used to build the index. We thus consider building the index offline with nonspecific digestion support and then perform partial prefetching when needed. Simulated digestion is only required once for each protein sequence database if the index is stored. When an index block is used for identification, the next block can be fetched by another process in background to prepare for the next round. Given that prefetching time can overlap, that is, the part B overhead in Figure 3 can be saved, nearly four times of improvement can be achieved. Considering LTQ as an example, the 16-node system is sufficient to support real-time identification on it, which generates approximately 5 MS/MS spectra per second. These results in Figure 6 show that real-time identification is feasible.

4. Conclusion

Motivated by the significance of protein identification, we identify the considerable computation demand primarily based on the development of high-throughput spectrometry and the expansion of known protein databases. We then propose an open identification method to support nonspecific digestion, which lays the foundation for acceleration and reduces the dependence on digestion expertise. We likewise accelerated the identification speed to real-time through appropriate distribution.

In the future, we will still focus on improving the speed and throughput of protein identification using algorithms and workflows. Moreover, we plan to conduct more work on open identification. We will continue contributing to practical protein identification systems similar to Google for large-scale Internet.

Acknowledgments

This research is supported by the Institute of Computing Technology, Chinese Academy of Sciences, and supported by the National Basic Research Program (973) of China under Grant no. 2010CB912701. The authors would gratefully acknowledge the generous financial support of the Institute of Computing Technology and the warm-hearted support and encouragement of many colleagues in the Institute of Computing Technology. In particular, Zhuojian Li was patient with system support. The authors are also grateful to pFind group, especially Professor Simin He, Hao Chi, Leheng Wang, Zuofei Yuan, Long Wu, and Kun Zhang, for open and helpful technical discussions. Furthermore, Wenli Zhang would like to thank her advisor Professor Jianping Fan for giving her the chance to do open research. Finally, the authors would also appreciate reviewers and editors for valuable and patient comments.

References

[1] J. S. Cottrell, "Protein identification using MS/MS data," *Journal of Proteomics*, vol. 74, no. 10, pp. 1842–1851, 2011.

[2] R. S. Johnson, M. T. Davis, J. A. Taylor, and S. D. Patterson, "Informatics for protein identification by mass spectrometry," *Methods*, vol. 35, no. 3, pp. 223–236, 2005.

[3] N. J. Edwards, "Protein identification from tandem mass spectra by database searching," *Methods in Molecular Biology*, vol. 694, pp. 119–138, 2011.

[4] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, no. 18, pp. 3551–3567, 1999.

[5] J. K. Eng, A. L. McCormack, and J. R. Yates III, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *Journal of the American Society for Mass Spectrometry*, vol. 5, no. 11, pp. 976–989, 1994.

[6] R. Craig and R. C. Beavis, "TANDEM: matching proteins with tandem mass spectra," *Bioinformatics*, vol. 20, no. 9, pp. 1466–1467, 2004.

[7] L. Y. Geer, S. P. Markey, J. A. Kowalak et al., "Open mass spectrometry search algorithm," *Journal of Proteome Research*, vol. 3, no. 5, pp. 958–964, 2004.

[8] L. Y. Geer et al., "Reducing false positive rates in MS/MS sequence searching and incorporating intensity into match based statistics," ftp://ftp.ncbi.nih.gov/pub/lewisg/presentations/asms06poster.pdf.

[9] D. Li, Y. Fu, R. Sun et al., "pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry," *Bioinformatics*, vol. 21, no. 13, pp. 3049–3050, 2005.

[10] L. Wang, D. Li, Y. Fu et al., "pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 21, no. 18, pp. 2985–2991, 2007.

[11] EBI, http://www.ebi.ac.uk/uniprot/TrEMBLstats/.

[12] R. D. Bjornson, N. J. Carriero, C. Colangelo et al., "X!!Tandem, an improved method for running X!Tandem in parallel on collections of commodity computers," *Journal of Proteome Research*, vol. 7, no. 1, pp. 293–299, 2008.

[13] D. T. Duncan, R. Craig, and A. J. Link, "Parallel tandem: a program for parallel processing of tandem mass spectra using PVM or MPI and X!Tandem," *Journal of Proteome Research*, vol. 4, no. 5, pp. 1842–1847, 2005.

[14] A. Quandt, P. Hernandez, P. Kunzst et al., "Grid-based analysis of tandem mass spectrometry data in clinical proteomics," *Studies in health technology and informatics*, vol. 126, pp. 13–22, 2007.

[15] D. Zosso, M. Podvinec, M. Müller, R. Aebersold, M. C. Peitsch, and T. Schwede, "Tandem mass spectrometry protein identification on a PC grid," *Studies in Health Technology and Informatics*, vol. 126, pp. 3–12, 2007.

[16] B. D. Halligan, J. F. Geiger, A. K. Vallejos, A. S. Greene, and S. N. Twigger, "Low cost, scalable proteomics data analysis using Amazon's cloud computing services and open source search algorithms," *Journal of Proteome Research*, vol. 8, no. 6, pp. 3148–3153, 2009.

[17] B. Pratt, J. J. Howbert, N. I. Tasman, and E. J. Nilsson, "Mr-tandem: parallel x!Tandem using hadoop MapReduce on Amazon web services," *Bioinformatics*, vol. 28, no. 1, Article ID btr615, pp. 136–137, 2012.

[18] Unimod, http://www.unimod.org/.

[19] E. Ahrné, M. Müller, and F. Lisacek, "Unrestricted identification of modified proteins using MS/MS," *Proteomics*, vol. 10, no. 4, pp. 671–686, 2010.

[20] P. Wilmes and P. L. Bond, "Metaproteomics: Studying functional gene expression in microbial ecosystems," *Trends in Microbiology*, vol. 14, no. 2, pp. 92–97, 2006.

[21] S. D. Patterson, "Data analysis—the Achilles heel of proteomics," *Nature Biotechnology*, vol. 21, no. 3, pp. 221–222, 2003.

[22] Y. Fu, L. Xiu, W. Jia et al., "DeltAMT: a statistical algorithm for fast detection of protein modifications from LC-MS/MS data," *Molecular and Cellular Proteomics*, vol. 10, no. 5, 2011.

[23] W. Jia, Z. Lu, Y. Fu et al., "A strategy for precise and large scale identification of core fucosylated glycoproteins," *Molecular and Cellular Proteomics*, vol. 8, no. 5, pp. 913–923, 2009.

[24] W. Zhang, H. Chi, Y. Lu, Y. Huang, X. Zhao, and S. He, "Preliminary search engine for open protein identification," in *Proceeding of the 13th International Conference on Parallel and Distributed Computing, Applications and Technologies*, pp. 410–415, 2012.

[25] L. Wang, S. He, R. Sun et al., "An efficient parallelization of phosphorylated peptide and protein identification," *Rapid Communications in Mass Spectrometry*, vol. 24, no. 12, pp. 1791–1798, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 210253, 20 pages
http://dx.doi.org/10.1155/2013/210253

Research Article

Identification of Interconnected Markers for T-Cell Acute Lymphoblastic Leukemia

Emine Guven Maiorov,[1] Ozlem Keskin,[1] Ozden Hatirnaz Ng,[2] Ugur Ozbek,[2] and Attila Gursoy[1]

[1] *Center for Computational Biology and Bioinformatics and College of Engineering, Koç University, Rumelifeneri Yolu, Sariyer, 34450 Istanbul, Turkey*
[2] *Department of Genetics, Institute for Experimental Medicine (DETAE), Istanbul University, 34393 Istanbul, Turkey*

Correspondence should be addressed to Ozlem Keskin; okeskin@ku.edu.tr and Attila Gursoy; agursoy@ku.edu.tr

Received 29 April 2013; Accepted 4 June 2013

Academic Editor: Tao Huang

T-cell acute lymphoblastic leukemia (T-ALL) is a complex disease, resulting from proliferation of differentially arrested immature T cells. The molecular mechanisms and the genes involved in the generation of T-ALL remain largely undefined. In this study, we propose a set of genes to differentiate individuals with T-ALL from the nonleukemia/healthy ones and genes that are not differential themselves but interconnected with highly differentially expressed ones. We provide new suggestions for pathways involved in the cause of T-ALL and show that network-based classification techniques produce fewer genes with more meaningful and successful results than expression-based approaches. We have identified 19 significant subnetworks, containing 102 genes. The classification/prediction accuracies of subnetworks are considerably high, as high as 98%. Subnetworks contain 6 nondifferentially expressed genes, which could potentially participate in pathogenesis of T-ALL. Although these genes are not differential, they may serve as biomarkers if their loss/gain of function contributes to generation of T-ALL via SNPs. We conclude that transcription factors, zinc-ion-binding proteins, and tyrosine kinases are the important protein families to trigger T-ALL. These potential disease-causing genes in our subnetworks may serve as biomarkers, alternative to the traditional ones used for the diagnosis of T-ALL, and help understand the pathogenesis of the disease.

1. Introduction

T-lineage acute lymphoblastic leukemia (T-ALL) is known to result from malignant transformation of hematopoietic precursor cells at different maturation stages of T cells, the so-called thymocytes [1]. Proliferation of developmentally arrested T cells gives rise to T-ALL. Differentiation arrest may take place at almost all stages of thymocyte development [2]. T-ALLs are a heterogeneous set of diseases, in terms of cytogenetics, molecular aberrations, and clinical characteristics [1]. Its pathogenesis and subtypes are usually undefined. T-ALL constitutes 15% of pediatric and 25% of adult ALL cases [1–4].

In early stages of thymocyte differentiation, immature T cells undergo V(D)J recombination [1]. During this time, many other genes, especially the T-cell receptor (TCR) genes, are transcribed and are in "open chromatin" configuration, meaning that they are easily accessible to DNA binding proteins, like recombinases. An unusual recombinase action may lead to translocation of chromosomes [3, 5, 6]. Generally, the translocations involve abnormal juxtaposition of powerful enhancers or promoters of TCR genes with genes on other chromosomes, such as transcription factors or oncogenes [3, 7]. Translocations give rise to not only promoter exchange but also fusion genes, encoding chimeric proteins [1]. The other molecular-genetic abnormalities in T-ALL involve deletions, amplifications, and point mutations which activate oncogenes or inhibit tumor suppressors, which, in turn, cause differentiation arrest in thymocytes [1–3]. Deletions are the reason for loss of tumor suppressors. The correct diagnosis of acute leukemia requires wide-ranging diagnostic procedures together with cytochemistry, multiparameter flow

cytometry, cytogenetics, fluorescence in situ hybridization, and molecular-genetic methods [8]. Although chromosomal rearrangements are common to T-ALL, there is still a large fraction of incidents (50%) where normal karyotype is seen [1].

The precise diagnosis of a tumor type is the most significant step in cancer treatments [9]. In order to apply the appropriate therapy, with maximum efficiency and minimum toxicity, the cancer should be diagnosed and classified correctly [10]. The challenge is to identify new diagnostic biomarkers to differentiate diseased and healthy individuals properly. An optimum biomarker would be easily analyzed by a single test and measurable in body fluids (such as blood or urine). However, cancer, in our case T-ALL, is a complex disease, and it is very difficult to find a single optimal biomarker at the molecular level [11].

The analysis of genome-wide expression profiles is frequently used to discover new biomarkers [12]. Since microarray or RNAseq high-throughput experiments give information about the expression of many genes in parallel [13], they are proposed to be a robust technology for the identification of signatures or expression patterns that vary significantly between diseased and healthy samples [11, 14, 15]. The current long-lasting diagnostic procedures for leukemia (e.g., cytomorphology, immunophenotyping, and metaphase cytogenetics) might be replaced by the comprehensive microarray or RNAseq protocols which takes two or fewer days and allows the simultaneous detection of the expression of almost all genome in one experimental approach [9]. With the extensive usage of microarrays, an increasing number of methods have been developed to identify biomarkers [14, 16–20].

Along with the advantages of microarrays, there are some limitations. For instance, some important genes of cancer are not differentially expressed at the level of transcription or the fate of cancer may not be controlled at the level of expression [11]. In addition, the transcriptional level does not always correlate with the translational level; in other words, the mRNA expression is not always equal to the protein expression [21]. The gene-expression levels may vary even in the genetically identical cells with the same histories of environmental exposure. These variations, known as "noise," come from the random nature of biochemical reactions [22]. Moreover, there is also an "experimental noise" other than the "expression noise" (biological variations), in which slight unintended differences in experimental setup may lead to huge differences in hybridization of probes. This kind of technical noise is also considered as one of the restrictions for successful use of this technology [23–25]. Another limitation of microarrays might be the lack of pathway knowledge. One way to overcome this problem is to integrate gene-expression profiles with protein-protein interaction (PPI) networks [26–29]. A biological function or a phenotype is not controlled by just one gene [30]; rather pathways or cross-talks among proteins are responsible for the regulation of a function [31–33]. Thus, network information provides a functional insight when integrated with microarray data [11]. Therefore, identification of differential gene modules or subnetworks instead of individual differentially expressed genes may increase the reliability and robustness of biomarkers [11].

Traditional expression-based classification techniques identify only differentially expressed (DE) genes as signatures/markers, but a network-based approach returns subnetworks including both DE and non-DE genes [12]. The differential expression analysis of gene-expression profiles returns massive numbers of genes which makes it difficult to conclude that the differential expression of a particular gene has resulted from the disease/abnormality. Therefore, we cannot be so sure that the differential analysis identifies genes whose differential expression has only resulted from the disease: there may be another factor affecting the expression of that gene, such as experimental setup, treatment, or gender. As noted earlier, even cells with identical genome and environmental exposure history show variations in their gene expression (noise). Thus, just differential analysis alone is not reliable enough to conclude a gene as a biomarker of a disease.

However, network-based approaches do not depend on only expression data. They integrate microarray data with PPI data and return subnetworks rather than individual genes. The number of subnetworks is not as high as the number of individual genes identified by differential analysis. These subnetworks are differential as a whole, but individual genes in a subnetwork may not be differential. Although irresponsive genes cannot be regarded as biomarkers, they have very important role in interconnecting several DE genes and can help us to understand the pathogenesis of the disease. The non-DE genes are necessary to maintain the integrity of the subnetworks, meaning that they are required to interconnect highly DE genes. Furthermore, generation of a disease necessitates some mutations or polymorphisms that may lead to loss of function or gain of function of proteins without affecting the expression level of that protein. This does not mean that this particular protein is not significant for the disease just because it is not differentially expressed. The resulting subnetworks represent models of the underlying molecular mechanisms. Each module corresponds to a different functional pathway or complex. Since genes function in collaboration rather than alone, subnetworks are more rational than independent responsive genes, from a biological perspective [34].

In this study, we used PinnacleZ algorithm [12] to integrate microarray and PPI data. In addition to PinnacleZ algorithm, there are several other studies which provide different ways to integrate gene-expression data with other biological data, such as protein-protein interactions, protein-DNA interactions, molecular signatures, or hub proteins [35–42]. These integrated analyses not only improve the prediction accuracy but also shed light on the biological pathways involved in the pathogenesis of the diseases. However, there are not significant differences in their prediction accuracies among themselves [43]. Combining PPI and microarray data has led also to determining some important proteins that are highly connected in interaction networks: "party" and "date" hubs [44, 45]. Party hubs are the ones highly correlated with its interacting partner proteins (coexpressed) and date hubs are less-correlated genes with its interacting partners [44, 45]. The method of Taylor et al. [42] successfully used hub proteins to reveal the dynamic modularity in protein networks to predict breast cancer outcome.

Hierarchical clustering is one example of the expression-based classification techniques which groups together both genes and samples with similar expression patterns. It can also define subclasses of a disease, such as cancer (e.g., different stages of cancer or different types of cancer, like B-ALL versus T-ALL) [46]. Clustering algorithms try to organize genes or samples according to their similarity in expression. Genes with related pattern appear in the immediate vicinity of each other. In other words, it gathers coexpressed genes together. It also has a tendency to arrange genes with similar functions together since the functionally related genes are likely to be co-expressed [47]. Hierarchical clustering is a type of unsupervised clustering which does not use any prior knowledge regarding the sample classes.

2. Methods

2.1. Microarray Data (The MILE Study). We used a comprehensive group of Affymetrix microarray datasets to determine which genes or modules discriminate T-ALL samples from healthy individual samples. This study included bone marrow samples of 173 T-ALL patients at diagnosis (untreated patients) and 74 nonleukemia/healthy specimens (e.g., healthy, hemolysis, and iron deficiency). The patient samples were heterogeneous; that is, there were samples from different stages of T-ALL. All microarray data were obtained from the Microarray Innovations in Leukemia (MILE) study, the National Center for Biotechnology Information's Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/) under series accession no. GSE13204. The MILE study is an international standardization program and was conducted by 11 laboratories across three continents [13]. The comprehensive MILE data has a high degree of intra- and interlaboratory correlation, meaning that they tried to minimize the disadvantages of microarrays, such as noise. In order to avoid the limitations of microarrays, we used only MILE data. There were two stages in this study containing microarray samples of 17 different classes of leukemia and myelodysplastic syndromes and an 18th class of non-leukemia. We used only the first stage in which the whole genome microarray platform (HG-U133 Plus 2.0; Affymetrix, Santa Clara, CA, USA) was used [15]. In the GEO-series matrix, the microarray data were already summarized and quantile normalized as described before [48].

In order to integrate microarray data with the human PPI network, it is necessary to convert Affymetrix probeset IDs to corresponding Entrez gene IDs. The annotation data was also provided under the same accession number. It is important to note that values for multiple probes corresponding to the same Entrez ID were averaged so that a particular gene ID is seen only once throughout the microarray data.

We also used a recently published dataset to test performance of our subnetworks. This dataset includes gene-expression profiles of childhood T-ALL bone marrow samples under series accession no. GSE46170 from GEO. This dataset includes 31 patients and 7 healthy samples.

2.2. Human Protein Interaction Network. The human PPI network is obtained from Human Protein Reference Database. There are 38788 PPIs whose interactions are experimentally verified and extracted from the literature [49].

2.3. Data Integration. In order to combine human PPI data with microarray data, we used a previously described algorithm (PinnacleZ, a plugin to cytoscape) [12]. This algorithm superimposes gene-expression values with corresponding network proteins, begins from every protein (seed) in the PPI network, and greedily appends interactions to identify the subnetwork starting from each seed whose mean expression for each sample best discriminates between the two sample types (In our case, the sample types are T-ALL and nonleukemia/healthy conditions). At first, the number of resulting subnetworks is equal to the number of proteins in the PPI network. Then, nonsignificant modules are filtered out by three types of permutation testing [30]. At last, subnetworks with high discriminative potential are obtained.

To equate the number of healthy individuals (74) to the number of patients with T-ALL (173), we divided T-ALL patient samples into two groups randomly. Then, we merged each half of patient data with the healthy data, separately. For each merged data (87 + 74 and 86 + 74), we ran the algorithm 4 times. Then, we resampled our patient samples 5 times (we again divided T-ALL samples into 2 groups randomly; thus these 2 groups are different from the 2 groups created before. So, we generated 10 different combinations of patients) and repeated the merging and algorithm-running steps. We obtained results of 40 runs in total.

2.4. Classification Accuracy. The classification accuracy of a given subnetwork or the overall rate of correct predictions of the patient and healthy datasets was estimated by 10-fold cross-validation with two different classifiers (J48 and RBF network classifiers implemented in WEKA [50]). According to 10-fold cross-validation, the complete dataset was divided into 10 uniformly sized subgroups; the classifier was trained for nine subgroups, and predictions were made for the remaining subgroup. High classification accuracy, like 90%, means that a given subnetwork differentiates/predicts the patient and healthy samples correctly, 90% of the time.

As pointed out, we randomly divided diseased samples into two groups: for each half we found subnetworks and determined their classification accuracy. We used the other half as a validation set to cross-test the prediction accuracy of a given subnetwork.

We also used an independent microarray dataset (childhood T-ALL samples, GSE46170) to test the classification accuracies of our subnetworks. Again the prediction accuracies are obtained by J48 and RBF network classifiers in WEKA, by 10-fold cross-validation.

2.5. Functional Enrichments. Functional enrichment analysis was achieved by using Gene Ontology Tree Machine (GOTM) which searches for functional enrichments from Gene Ontology (GO) and Kyoto Encyclopedia of Genes and

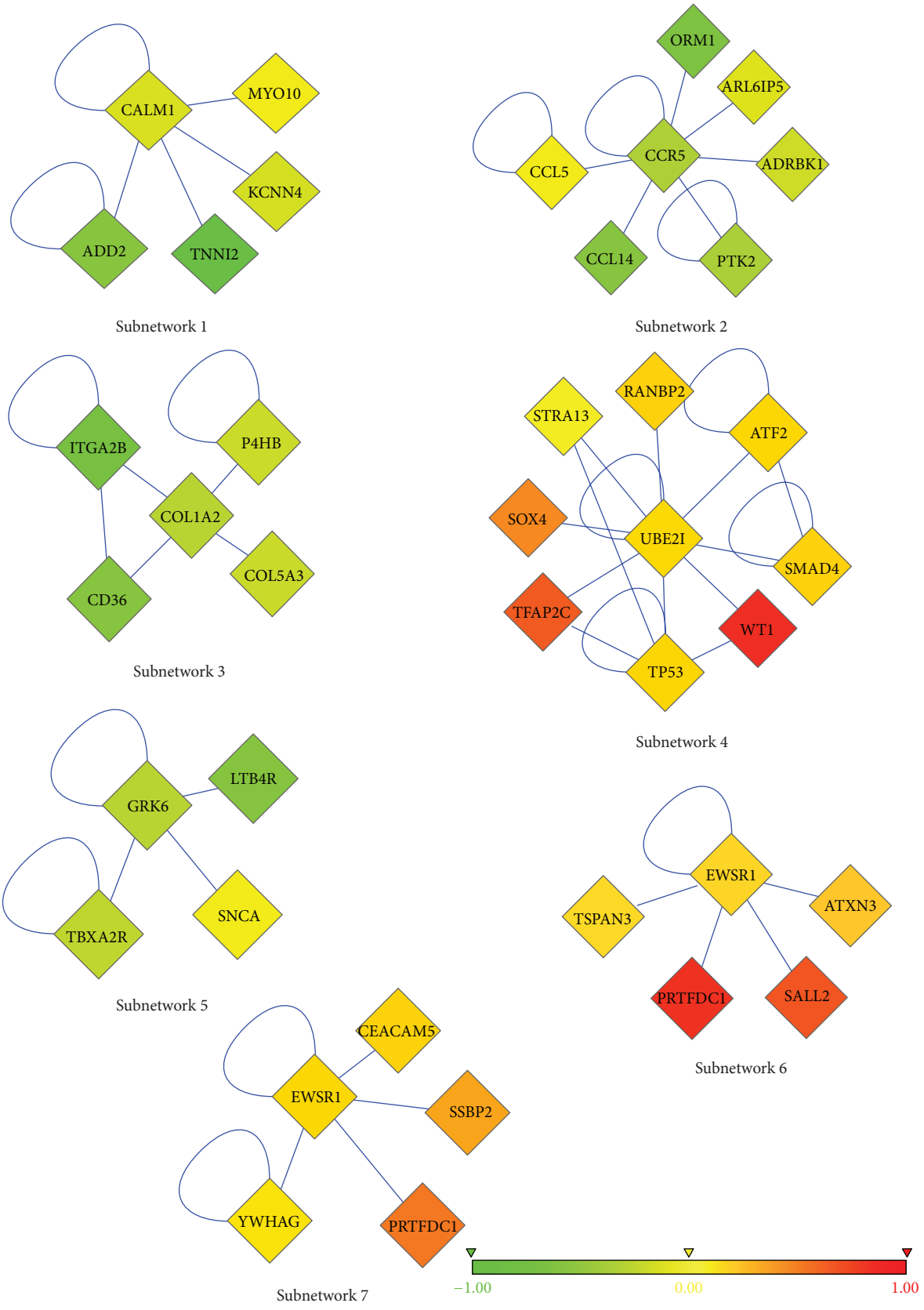


Figure 1: First seven of the most frequent subnetworks. Nodes represent proteins, and edges represent interactions. The color of each node ranges in accordance with the change in expression of the corresponding gene for T-ALL versus healthy samples. The shape of each node shows whether its gene is significantly differentially expressed (diamond; $P < 0.05$ from a two-tailed t-test) or not (circle).

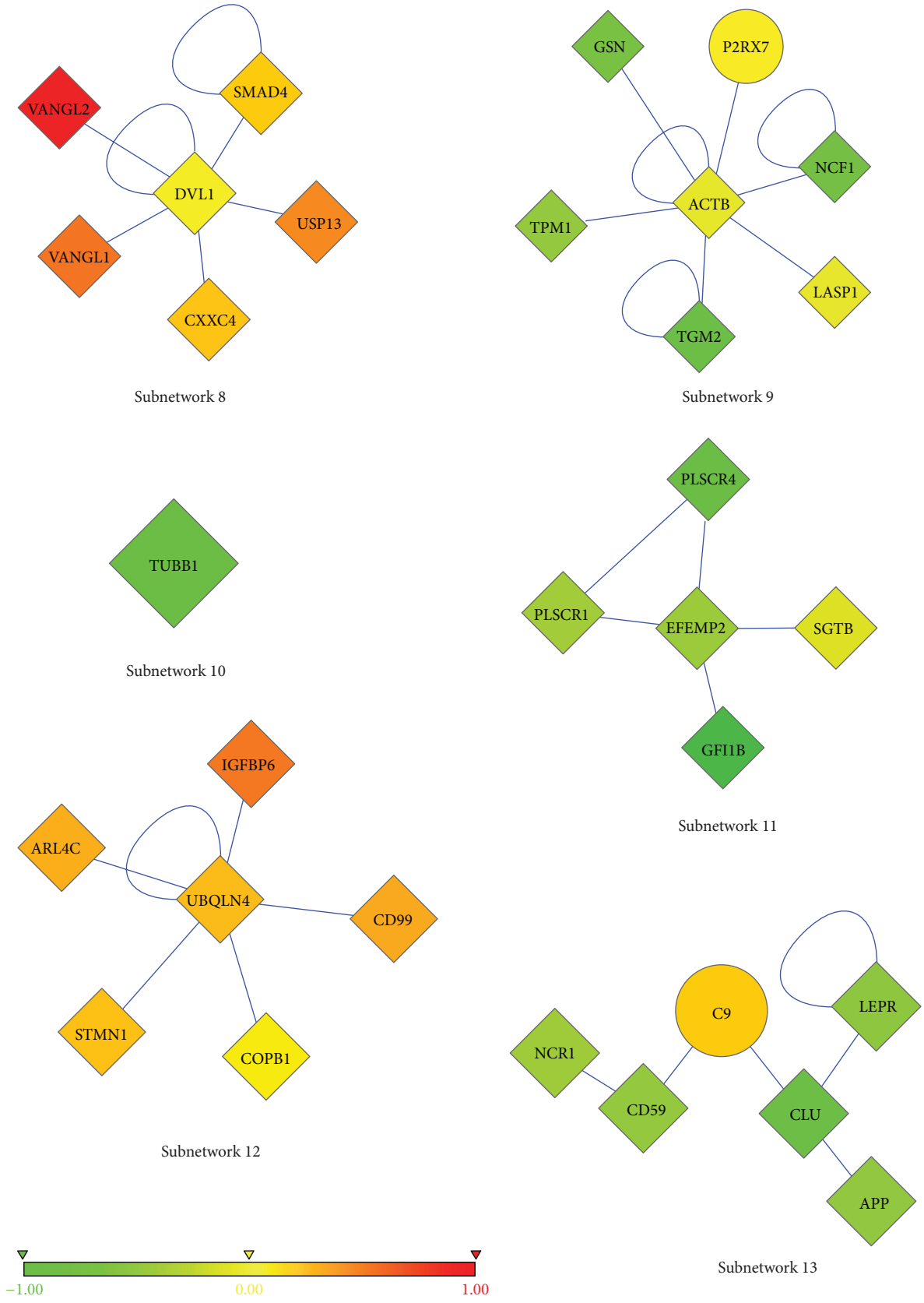


FIGURE 2: Second six of the most frequent subnetworks. Nodes represent proteins, and edges represent interactions. The color of each node ranges in accordance with the change in expression of the corresponding gene for T-ALL versus healthy samples. The shape of each node shows whether its gene is significantly differentially expressed (diamond; $P < 0.05$ from a two-tailed t-test) or not (circle).

Table 1: GO enrichments of nondifferentially expressed genes.

Gene symbol	Subnetwork	Gene title	GO: function	GO: process
P2RX7	Sub-9	Purinergic receptor P2X, ligand-gated ion channel, 7	ATP binding, ATP-gated cation channel activity, ion channel, and receptor activity	Ion transport, signal transduction
C9	Sub-13	Complement component 9	No enrichments	Caspase activation, complement activation, and induction of apoptosis
CHGA	Sub-14	Chromogranin A (parathyroid secretory protein 1)	Calcium ion binding, protein binding	Regulation of blood pressure
PLG	Sub-14	Plasminogen	Apolipoprotein binding, calcium ion binding, peptidase activity, and plasmin activity	Blood coagulation, induction of apoptosis, negative regulation of angiogenesis, cell proliferation, fibrinolysis, proteolysis, and tissue remodeling
MEP1A	Sub-14	Meprin A, alpha (PABA peptide hydrolase)	Astacin activity, metal-ion binding, metallopeptidase activity, and zinc-ion binding	Digestion, proteolysis
HN1L	Sub-16	Hematological and neurological expressed 1-like	No enrichments	No enrichments

Genomes (KEGG) categories for genes in the given subnetworks.

2.6. Hierarchical Clustering. In order to identify groups of genes that have similar expression patterns and to show the difference between expression-based and network-based classification approaches, hierarchical clustering method was applied to 173 T-ALL and 74 healthy samples together. Initially, there were 20148 probes. Then, these probes are filtered by t-test. We determined the first 100 most differential genes, and we repeated the same analysis with the first 200 most differential genes. The rationale behind this is to analyze how the results would change between these two sets. Hierarchical clustering algorithm which is implemented in Expander [51] was performed on both the 200 and 100 most differentially expressed genes. Both samples and genes are clustered with complete linkage and Pearson correlation.

3. Results and Discussion

We applied both expression-based and network-based approaches to find important genes for the generation of T-ALL and to show that network-based approaches are more successful in returning more meaningful results than expression-based approaches.

As a network-based approach, we used PinnacleZ algorithm [12] to distinguish T-ALL patients and healthy samples by integrating microarray data with the human PPI network. This approach enabled us to identify subnetworks/modules as markers which differentiate the patients from healthy individuals. The size of subnetworks or the number of genes in a subnetwork varies, ranging from 1 to 10 genes. As noted earlier, the individual genes may not be responsive, meaning that expression of a particular gene is different in patients and healthy samples, but an entire subnetwork is differential.

As explained in the Methods section, we prepared different combinations (10 different combinations) of patient-healthy merged data and ran the algorithm 4 times with each combination. Then, we focused on the most frequent subnetworks. We recovered the most repeated 19 subnetworks, out of 183 subnetworks (see Figures 1, 2, and 3. Please refer to Supplementary Material available online at http://dx.doi.org/10.1155/2013/210253 (Table S1–Table S19) to see the names and functions of these genes.). These subnetworks cover 102 genes in total, and the majority of them are differentially expressed (DE). There are only 6 genes whose expression is not differential among T-ALL and healthy samples (see Table 1). The list of genes in the subnetworks, their functions, and the pathways they are involved in can be found in Table 2 and also in the Supplementary Material.

Development of cancer requires the accumulation of several mutations in several genes in different pathways [52]. Since the cell regulation is controlled by many pathways, a mutation in one pathway may be compensated by other pathways. But if there are many mutations in numerous pathways, the harmful impacts of these mutations cannot be compensated. For instance, if a tumor suppressor gene is

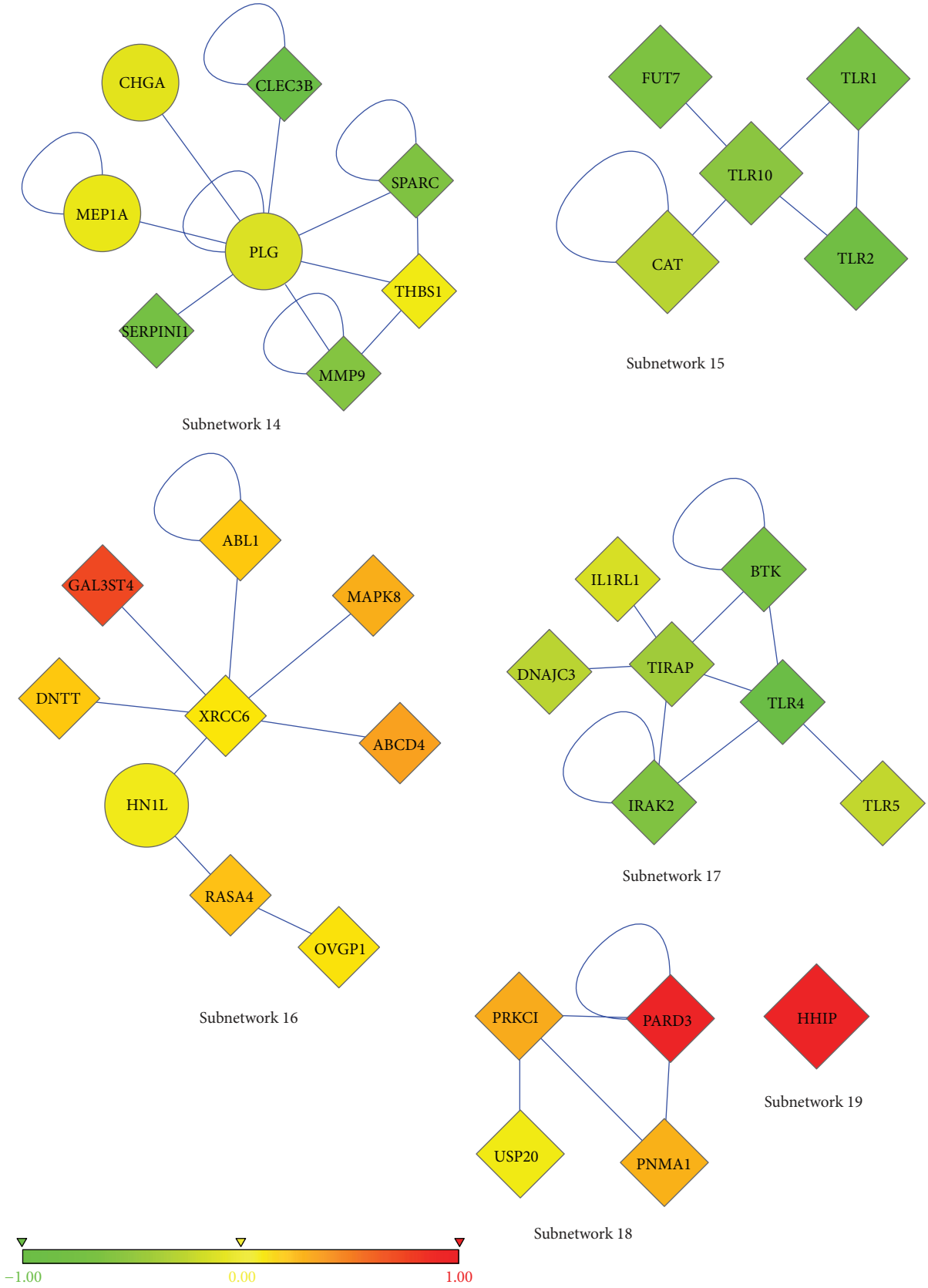


FIGURE 3: The last six of the most frequent subnetworks. Nodes represent proteins, and edges represent interactions. The color of each node ranges in accordance with the change in expression of the corresponding gene for T-ALL versus healthy samples. The shape of each node shows whether its gene is significantly differentially expressed (diamond; $P < 0.05$ from a two-tailed t-test) or not (circle).

Table 2: KEGG and GO enrichments of the most frequent 19 subnetworks.

Subnetworks	KEGG	GO: function
Sub-1	No enrichments	Actin binding, calmodulin binding
Sub-2	Chemokine-signaling pathway, Chemokine-cytokine receptor interaction, endocytosis	Signal-transducer activity, chemokine activity, and cytokine activity
Sub-3	ECM-receptor interaction, focal adhesion, and hematopoietic cell lineage	Growth factor binding, extracellular matrix structural constituent
Sub-4	Pathways in cancer, colorectal cancer, pancreatic cancer, chronic myeloid leukemia, cell cycle, Wnt signaling pathway, and MAPK signaling pathway	Promoter binding, transcription activator activity, and transcription factor activity
Sub-5	Neuroactive ligand-receptor interaction	Signal-transducer activity, G-protein coupled receptor activity
Sub-6	No enrichments	No enrichments
Sub-7	No enrichments	No enrichments
Sub-8	Pathways in cancer, colorectal cancer, and Wnt signaling pathway	No enrichments
Sub-9	Leukocyte transendothelial migration, regulation of actin cytoskeleton, dilated cardiomyopathy, Fc gamma R-mediated phagocytosis, and hypertrophic cardiomyopathy (HCM)	GTP binding, actin binding
Sub-10	Phagosome, gap junction	GTP binding, GTPase activity, nucleotide binding, and structural molecule activity
Sub-11	No enrichments	Metal-ion binding, calcium ion binding, SH3 domain binding, and phospholipid transporter activity
Sub-12	No enrichments	No enrichments
Sub-13	Complement and coagulation cascades	No enrichments
Sub-14	Bladder cancer	Calcium ion binding, metalloendopeptidase activity
Sub-15	Toll-like receptor signaling pathway	Transmembrane receptor activity
Sub-16	Pathways in cancer, neurotrophin signaling pathway, ErbB signaling pathway, and nonhomologous endjoining	ATP binding, protein C-terminus binding
Sub-17	Toll-like receptor signaling pathway, pathogenic *Escherichia coli* infection	Transmembrane receptor activity
Sub-18	Tight junction, endocytosis	No enrichments
Sub-19	No enrichments	No enrichments

inactivated, it is not enough to generate cancer; additional mutations are needed before cancer appears [53]. We found that the subnetworks correspond to different pathways, in general (Table 2). This result supports the notion that multiple cell regulatory pathways are involved in production of leukemogenesis. Targeting defective molecular pathways is more effective in getting rid of cancer cells and less destructive for rapidly dividing normal cells, which is referred to as "targeted therapy" [54, 55]. Since targeted therapy aims at destroying only tumor cells whose particular pathways are broken or malfunctioning, it might have fewer side effects than chemotherapy and radiotherapy which are cytotoxic to all fast-proliferating cells, including the healthy ones. Targeting our subnetworks or their corresponding pathways may result in development of efficient novel drugs for T-ALL treatment.

Table 3: T-ALL related genes found in subnetworks and their behavior in our samples.

T-ALL related genes	Subnetwork	Behavior
ABL1	16	↑
CCL5	2	↓
CD99	12	↑
TP53	4	↑
WT1	4	↑

Among the genes in the subnetworks, some are known to be associated with T-ALL from previous studies: ABL1 [1, 56, 57], CCL5 [58], CD99 [59], TP53 [60], and WT1 [61, 62]. Table 3 exhibits these T-ALL related genes and their behavior. Many other genes seen in subnetworks were not previously

Table 4: Cancer related genes recovered in subnetworks and their behavior in our samples.

Cancer related genes	Subnetwork	Behavior
VANGL1	8	↑
CEACAM5	7	↑
SSBP2	7	↑
LTB4R	5	↓
TFAP2C	4	↑
USP13	8	↑
CD36	3	↓
UBE2I	4	↑
EWSR1	6	↑
SOX4	4	↑
LASP1	9	↓

found to be linked to T-ALL (some are related to cancer, but not specifically to T-ALL, such as VANGL1 [63], CEACAM5 [64], SSBP2 [65], LTBR4 [66], TFAP2C [67], USP13 [68], CD36 [69], UBE2I [70], EWSR1 [71], SOX4 [72], and LASP1 [73]) (See Table 4, for corresponding subnetworks and the behavior of these genes). So, after experimental validation, these genes may serve as novel markers for T-ALL.

There are 6 non-DE genes in our subnetworks (Table 1). Although these irresponsive genes cannot be considered as markers of T-ALL, they have very important roles in interconnecting numerous DE genes, and their presence could be essential for malignant transformation of precursor T-cells. One of the 6 non-DE genes is P2RX7 (Figure 2, subnetwork 9), which is a purinergic receptor P2X, expressed in hematopoietic cells, and mediates both apoptosis and proliferation, depending on the level of activation [74–76]. Prolonged activation of this receptor by extracellular ATP is a significant mechanism to initiate apoptosis in T and B lymphocytes [76, 77]. Its loss of function by an SNP (1513A → C) has an antiapoptotic effect and is previously shown to be related to chronic lymphoblastic leukemia (CLL). Although this SNP abolishes the function of P2RX7, it does not have an impact on the expression level of the receptor [76, 78]. Thus, loss of function does not imply reduced expression. In T-ALL, it is possible seeing a similar mutation which leads to loss of function of this receptor, without affecting its expression level. Therefore, despite its nondifferential expression, it may contribute to the pathogenesis of the disease. Only differential expression analysis would not highlight this gene as important, but network-based approach detected it as a significant one. Moreover, loss of function of P2RX7 decreases the efficiency of adjuvant chemotherapy in breast cancer patients [79]. So, P2RX7 should be present to benefit from chemotherapy. Apart from its role in apoptosis, it also promotes proliferation upon weak stimulation [80]. Furthermore, P2RX7 expression (not necessarily upregulation) results in increased proliferation and reduced apoptosis [81]. When oxidized ATP, P2RX7 inhibitor, was injected into the tumor, the tumor shrank [81]. Upregulation of this gene was seen in acute lymphoblastic leukemia (ALL), acute myelogenous leukemia (AML), chronic myelogenous leukemia (CML), and myelodysplastic syndrome [82]. In addition, its high expression diminished the remission rate of AML after a dose of standard therapy [82]. Although the upregulation of P2RX7 is not observed in our T-ALL samples, these data suggest that it is worthwhile to further investigate the potential role of P2RX7 in the generation of T-ALL.

PLG and MEP1A are other examples of our non-DE genes. PLG (Figure 3, subnetwork 14) encodes plasminogen which is essential for cancer cell invasion and metastasis. Plasminogen activators convert plasminogen to active plasmin which in turn activates MEP1A [83]. MEP1A (Figure 3, subnetwork 14), meprin A, is a metalloprotease that cleaves proteins and degrades extracellular matrix, facilitating the tissue invasion and metastasis [84]. They participate in the migration of leukocytes to the sites of infection and migration of cancer cells in metastasis [85]. Meprin A is upregulated in several cancer cells [86, 87]. T-cell lymphoid tumor growth is decreased by plasmin inhibitors by suppressing metalloproteinases [88]. Even though these two genes are not DE in our T-ALL samples, there is clear evidence that these two groups of enzymes are very important candidates of disease-causing genes.

Another non-DE gene is CHGA, chromogranin A, which is an acidic glycoprotein commonly expressed by neuroendocrine cells [89]. It is widely used as a diagnostic and prognostic biomarker for neuroendocrine tumors [90]. Our fifth non-DE gene is C9, complement component 9. Tumor cells possess some protective mechanisms against complement-mediated tumor cell lysis [91]. Human leukemic cells remove membrane attack complexes from their surfaces by phosphorylating C9 [92]. Therefore extracellular phosphorylation of C9 provides a defense mechanism against complement system. In addition, complement system is defective in CLL patients [93]. T-ALL could have a similar protective method as CLL does. Further studies are necessary to elucidate the roles of these non-DE genes in the pathogenesis of T-ALL. These genes may lead to new clinical therapies for T-ALL.

Subnetworks are rich in transcription factors: there are 14 transcription factors involved in subnetworks (Table 5). This result is in accordance with the important assumption that abnormal or ectopic activation of specific transcription factor genes, with/without chromosomal rearrangements, is the main event in transformation of immature T cells [5].

There were 6 tyrosine kinases (Table 6), in our subnetworks. Tyrosine kinases have a critical role in TCR signaling, regulation of T-cell immune response, and T-cell survival and proliferation. Expression of tyrosine kinases, like ABL1, affect pre-TCR and TCR signaling and give a proliferative and survival advantage [1]. ABL1 is an oncogene and is often seen as ABL1-NUP214 fusion gene in T-ALL [1, 7, 61]. As can be seen in subnetwork 16 (Figure 3), ABL1 is upregulated in T-ALL patients compared to healthy individuals. Although YWHAG was not found to be related to T-ALL before, it interacts with ABL1. The overexpression of ABL1 might induce upregulation of the YWHAG gene (Figure 1, subnetwork 7). BTK is significantly downregulated in T lymphocytes, which is consistent with our results. PTK2, also known as FAK, Focal Adhesion Kinase, (Figure 1, subnetwork 2) has a role in growth, differentiation, tumor metastasis, and wound

Table 5: Transcription factors recovered in subnetworks.

Gene symbol	Gene title	Subnetwork	Behavior
TNNI2	Troponin I type 2 (skeletal, fast)	Sub-1	Down
SMAD4	SMAD family member 4	Sub-4	Up
TP53	Tumor protein p53	Sub-4	Up
ATF2	Activating transcription factor 2	Sub-4	Up
WT1	Wilms, tumor 1	Sub-4	Up
SOX4	SRY- (sex determining region Y-) box 4	Sub-4	Up
TFAP2C	Transcription factor AP-2 gamma (activating enhancer binding protein 2 gamma)	Sub-4	Up
ATXN3	Ataxin 3	Sub-6	Up
SALL2	Sal-like 2 (*Drosophila*)	Sub-6	Up
EWSR1	Ewing sarcoma breakpoint region 1	Sub-6, sub-7	Up
SSBP2	Single-stranded DNA binding protein 2	Sub-7	Up
GFI1B	Growth factor independent 1B transcription repressor	Sub-11	Down
ABL1	c-abl oncogene 1, receptor tyrosine kinase	Sub-16	Up
XRCC6	X-ray repair complementing defective repair in Chinese hamster cells 6 (Ku autoantigen, 70 kDa)	Sub-16	Up

Table 6: Genes involved in tyrosine-kinase signaling pathway recovered in subnetworks.

Gene symbol	Gene title	Subnetwork	Behavior
BTK	Bruton agammaglobulinemia tyrosine kinase	Sub-17	Down
ABL1	c-abl oncogene 1, receptor tyrosine kinase	Sub-16	Up
YWHAG	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, gamma polypeptide	Sub-7	Up
PTK2	PTK2 protein tyrosine kinase 2	Sub-2	Down
COL1A2	Collagen, type I, and alpha 2	Sub-3	Down
SPARC	Secreted protein, acidic, and cysteine rich (osteonectin)	Sub-14	Down

healing [94–97]. Its overexpression is associated with several types of cancer [98, 99]. But in a recent study, it has been shown that PTK2 protein is predominantly absent in both normal T cells and T-lymphoblastic leukemia/lymphoma. Although it is negative in T-cell leukemia/lymphoma, it is mostly positive in B-cell lymphomas [96]. Consistent with the literature, PTK2 gene is also downregulated in T-ALL patients used in this study. Cell adhesion molecules (CAMs) are necessary for interaction of hematopoietic cells with extracellular matrix with stromal and other cells [53]. Defects in adhesion were reported in other types of leukemia before, such as in chronic myeloid leukemia (CML) [53, 100, 101]. The failure of hematopoietic stem cells (HSCs) to express the correct or fundamental adhesion molecules may contribute to transformation of a normal HSC to leukemic cell and to get arrested at a particular step of their differentiation. The adhesion deficiency may also help leukemic cells to escape from the recognition by immune system [53].

Interestingly, the subnetworks are also abundant in zinc-ion (Zn^{2+}) binding proteins (Table 7) which are generally enzymes, including those involved in DNA repair. Zinc-finger motifs play key role in interaction of proteins with nucleic acids (DNA/RNA) [102]. They are essential for site-specific DNA recognition and transcriptional activation [103]. Zn^{2+} has both structural and regulatory roles in zinc-binding proteins, meaning that Zn^{2+} maintains the three-dimensional structure of the proteins, and it is required for the proper function of the proteins. For example, p53 needs Zn^{2+} to fold properly. Both excess and inadequate amounts of Zn^{2+} cause misfolding of p53 [103]. One of the molecular mechanisms in carcinogenesis is the deformation of zinc-finger domains in DNA repair proteins [102]. Zn^{2+} is also important for thymic immune responses [104]. Low levels of zinc are frequently reported in ALL cases. Normal lymphocytes contain more zinc than leukemic cells [105]. Treatment of ALL patients with zinc, in addition to chemotherapy, was hypothesized to increase the overall ability to recover from T-ALL permanently and to endure toxic effects of chemotherapy [106]. Some of the Zn^{2+} ion binding proteins are upregulated, but some are downregulated (see Table 7). Although zinc-ion binding proteins do not behave similarly at the level of expression (i.e., they are not all up-regulated or down-regulated: some of them are upregulated, while some are downregulated), it is evident that more attention should be paid to them.

Two subnetworks are actually individual genes rather than interconnected genes (subnetwork 10, Figure 2 and subnetwork 19, Figure 3). These genes are TUBB1 and HHIP.

Table 7: Zinc-ion binding proteins recovered in subnetworks.

Gene symbol	Gene title	Subnetwork	Behavior
TP53	Tumor protein p53	Sub-4	Up
ATF2	Activating transcription factor 2	Sub-4	Up
WT1	Wilms, tumor 1	Sub-4	Up
SALL2	Sal-like 2 (*Drosophila*)	Sub-6	Up
EWSR1	Ewing sarcoma breakpoint region 1	Sub-6, sub-7	Up
USP13	Ubiquitin specific peptidase 13 (isopeptidase T-3)	Sub-8	Up
CXXC4	CXXC finger 4	Sub-8	Up
LASP1	LIM and SH3 protein 1	Sub-9	Down
GFI1B	Growth factor independent 1B transcription repressor	Sub-11	Down
APP	Amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer's disease)	Sub-13	Down
MEP1A	Meprin A, alpha (PABA peptide hydrolase)	Sub-14	Non-DE
MMP9	Matrix metallopeptidase 9 (gelatinase B, 92 kDa gelatinase, and 92 kDa type IV collagenase)	Sub-14	Down
RASA4	RAS p21 protein activator 4	Sub-16	Up
BTK	Bruton agammaglobulinemia tyrosine kinase	Sub-17	Down
USP20	Ubiquitin specific peptidase 20	Sub-18	Up
PRKCI	Protein kinase C, iota	Sub-18	Up

HHIP stands for Hedgehog Interacting Protein, which is a negative regulator of Hedgehog signaling pathway. Overactivity of Hedgehog signaling pathway is related to many cancer types [107]. HHIP is found to be associated with lung cancer [108] and brain tumor [109]. Although it is down-regulated in several tumor types, it is up-regulated in our T-ALL patients (subnetwork 19, Figure 3) [110, 111]. The other individual gene that is found as a marker is TUBB1, Tubulin beta-1 (subnetwork 10, Figure 2), which has a role in assembly of microtubules only in hematopoietic cells. Altered expressions of beta-tubulin isotypes were observed in specific tumor types [112]. Tubulin mutations are involved in resistance to drugs that target microtubules in cancer patients [113].

CALM1 gene, a member of subnetwork 1 (Figure 1), has been shown to be involved in a translocation with AF10 gene, and this fusion gene was detected in almost 10% of immature T-ALL patients. CALM-AF10 fusion gene upregulates HOXA gene cluster and has shown to be related to bad prognosis [114]. Dik et al. studied gene-expression profiles of CALM-AF10 positive and negative T-ALL patients and revealed that the TUBB gene was 7-fold overexpressed in CALM-AF10 positive patients [115]. In this study, TUBB1 gene was detected in subnetwork 10 as an individual gene (Figure 2). TUBB polymorphisms were described in ALL patients, and they

Table 8: Classification accuracies of subnetworks found by two different classifiers in WEKA.

Subnetworks	J48	RBF network	J48 (validation set)	RBF network (validation set)
Sub-1	93	97	95	96
Sub-2	91	96	91	98
Sub-3	93	97	96	98
Sub-4	98.75	99	93	96
Sub-5	94	97	92	98
Sub-6	96	97	94	94
Sub-7	96	97.5	96	96
Sub-8	94	97	96	93.75
Sub-9	93	98	93	94
Sub-10	97	97	96	96.25
Sub-11	91	98	93	97
Sub-12	95	99	96	98
Sub-13	98	98	95	99
Sub-14	94	97.5	91	93
Sub-15	93	97	92	93
Sub-16	95	96	92.5	97
Sub-17	89	97	90	94
Sub-18	93	96	94	97
Sub-19	96.27	96.89	96.25	97

Table 9: Classification accuracies of subnetworks on an independent microarray data by two different classifiers in WEKA.

Subnetworks	J48	RBF network
Sub-1	73.68	76.31
Sub-2	94.73	92.1
Sub-3	81.57	76.31
Sub-4	68.42	94.73
Sub-5	89.47	97.36
Sub-6	78.94	78.94
Sub-7	78.94	81.57
Sub-8	89.47	84.21
Sub-9	89.47	81.57
Sub-10	81.57	73.68
Sub-11	81.57	86.84
Sub-12	68.42	89.47
Sub-13	71.05	71.05
Sub-14	81.57	89.47
Sub-15	92.1	86.84
Sub-16	86.84	86.84
Sub-17	84.21	89.47
Sub-18	92.1	81.57
Sub-19	81.57	76.31

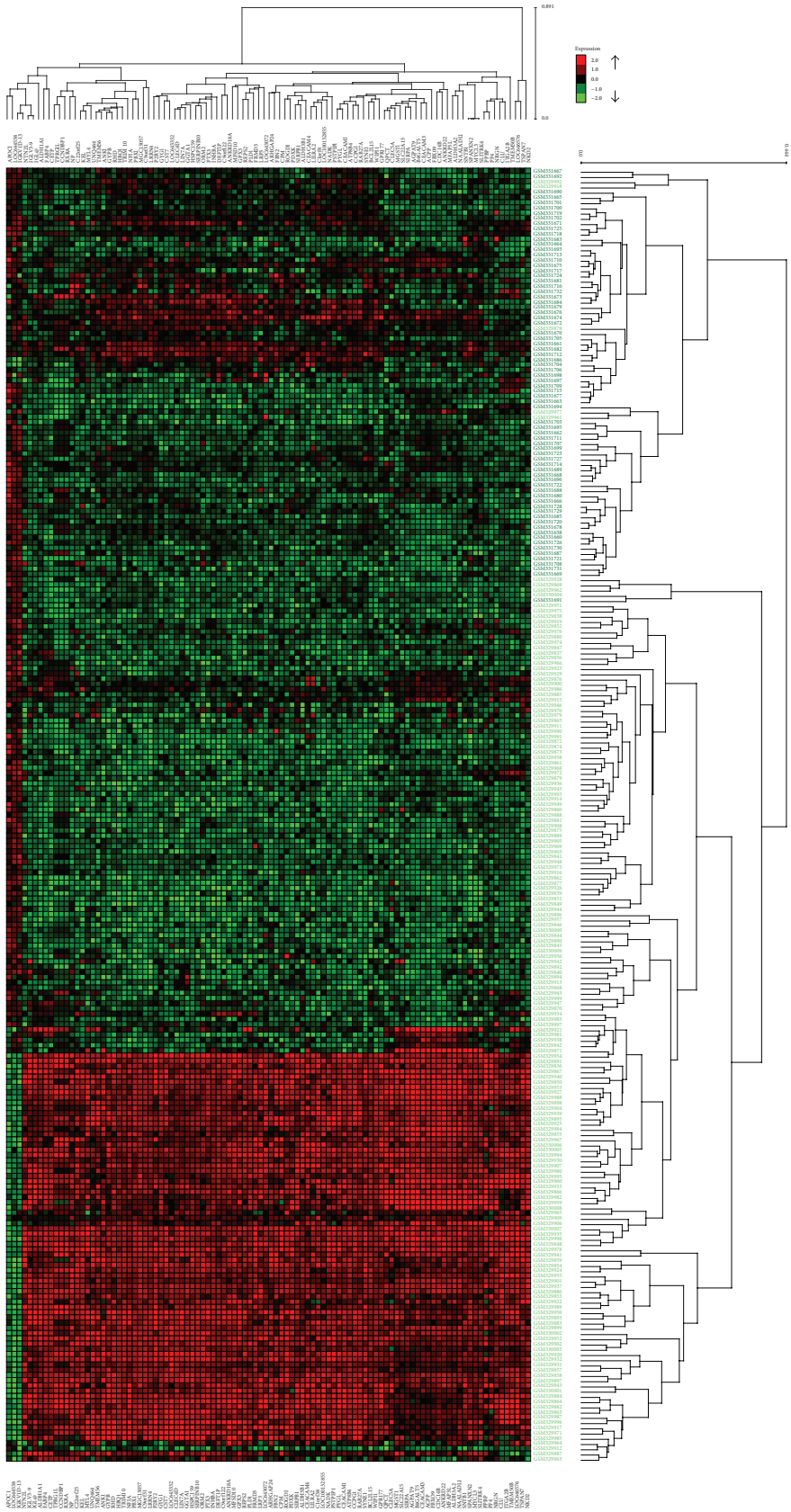


Figure 4: The heat map shows the hierarchical clustering result of the 100 most differentially expressed genes in T-ALL with respect to healthy individuals. Red and green spots represent upregulated and downregulated genes, respectively. Black spots denote equal expression. The columns labeled with light green belong to healthy individuals, and the columns labeled with black are individuals with T-ALL.

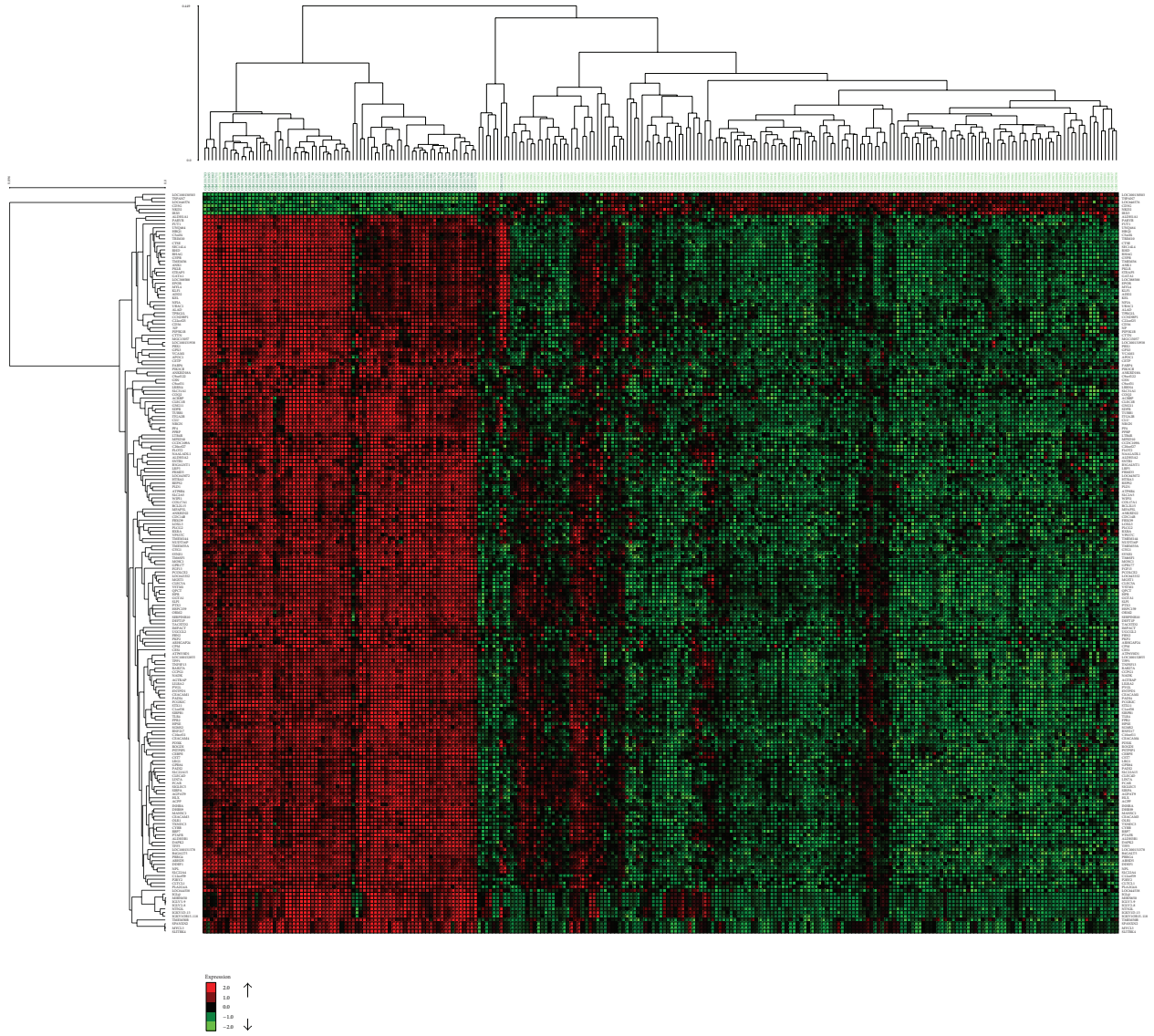


FIGURE 5: The heat map shows the hierarchical clustering result of the 200 most differentially expressed genes in T-ALL with respect to healthy individuals. Red and green spots represent up-regulated and down-regulated genes, respectively. Black spots denote equal expression. The columns labeled with light green belong to healthy individuals, and the columns labeled with black are individuals with T-ALL.

were suspected to be involved in drug resistance [116]. ADD2 gene is another gene in subnetwork 1. ADD genes are a family of cytoskeleton proteins encoded by three genes (ADD1, ADD2, and ADD3). ADD2 gene knockout mice are used as models for leukemia, and ADD3 gene was shown to have a translocation with NUP98 in T-ALL patients [117]. These two findings show that ADD gene family takes place in the hematopoiesis and also in hematologic malignancies.

In subnetwork 8 (Figure 2), two genes that take part in early developmental stages, VANGL1 and VANGL2, are found directly related to DVL gene. DVL gene negatively regulates WNT signalling pathway which plays an important role in the hematopoiesis, particularly in T-cell development [118]. These findings stress once again the importance of these networks in T-ALL pathogenesis not only on gene-expression level but also on protein level.

3.1. Classification Accuracies of Subnetworks. After finding subnetworks, we tested their classification accuracies with two different classifiers, namely, J48 and RBF-network in WEKA [50]. The prediction accuracy of each subnetwork was tested individually with 10-fold cross-validation. As discussed, the patient samples were randomly divided into 2 groups, and, for each subgroup, differential subnetworks were found. Classification accuracies of subnetworks were found by testing their original sub-group (the group for which the subnetworks were found) and by cross-testing the remaining sub-group (the other half of the patients). The cross-testing was applied to validate the prediction accuracies of modules also in different sets of patient microarray data. All of the subnetworks achieved very high accuracies in prediction, higher than 90% (Table 8). There are some subnetworks that achieved 99% prediction accuracy (Figure 1, subnetwork 4

Table 10: T-ALL related genes in KEGG pathways [119].

Gene symbol	Gene title
SIX4	SIX homeobox 4
LMO2	LIM domain only 2 (rhombotin-like 1)
HPGD	15-Hydroxyprostaglandin dehydrogenase (NAD)
GRIA3	Glutamate receptor, ionotropic, and AMPA 3
EYA1	Eyes absent homolog 1 (*Drosophila)*
FUT8	Fucosyltransferase 8 (alpha (1,6) fucosyltransferase)
MLL	Myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, *Drosophila*)
MEIS1	Meis homeobox 1
HOXA9	Homeobox A9
TLX1	T-cell leukemia homeobox 1
CCR7	Chemokine (C-C motif) receptor 7
LDB1	LIM domain binding 1
CDKN2C	Cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)
LYL1	Lymphoblastic leukemia derived sequence 1
TCF3	Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)
HHEX	Hematopoietically expressed homeobox
SIX1	SIX homeobox 1
HOXA11	Homeobox A11
PTCRA	Pre-T-cell antigen receptor alpha
MLLT1	Myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, *Drosophila*), translocated to 1
HOXA10	Homeobox A10

and Figure 2, subnetwork 12). These results also prove the success of the network-based classification approaches. Cross-comparisons between two independent halves of patient dataset revealed that subnetworks are good at distinguishing T-ALL patients from healthy individuals, regardless of the dataset in which they are found. In other words, they can classify patients in both independent datasets with similar accuracies.

To check whether our subnetworks are also applicable to other publically available microarray data, we used gene-expression profiles of childhood T-ALL samples (GSE46170). As Table 9 shows, the classification accuracies of subnetworks on this independent dataset are also relatively high, about 83% on average (ranging 71–94%) over subnetworks. Compared to original dataset (MILE) on which the subnetworks are found, the independent dataset showed lower performance on classification. This decrease in classification accuracies may stem from the fact that the independent dataset contains only childhood T-ALL samples as opposed to MILE study which has heterogeneous patients, meaning that there are patients from different stages of the disease and they are not specifically childhood T-ALL samples. Another reason may be the imbalanced number of patients (31 patients) and healthy samples (7 healthy individuals) in this independent dataset. The imbalanced numbers of healthy and patient samples may also decrease the prediction accuracy.

Table 11: Classification accuracies of 21 T-ALL related genes in KEGG pathways in MILE and independent microarray data by two different classifiers in WEKA.

Datasets	J48	RBF network
MILE study dataset	93.11	95.95
Independent dataset	84.21	86.84

Moreover, to demonstrate the accomplishment of our subnetworks, we compared them with T-ALL related genes in KEGG pathways [119], considered as a module. Table 10 displays these 21 genes, and Table 11 presents their classification accuracies. The performance of these genes is also high, but this outcome is not surprising because these 21 genes are already known to be related to T-ALL. However, our subnetworks largely consist of novel markers of T-ALL, and they do the same or better jobs than these 21 genes in KEGG pathway. Moreover, there is also a decrease in accuracies of these genes when tested on independent dataset compared to MILE dataset. So, it is normal that our subnetworks achieve higher accuracies in MILE dataset but lower in the independent dataset (GSE46170). As indicated above, the reason may be the imbalanced numbers of healthy and patient samples.

Although integrating microarray data with network information is a promising way to identify functional biomarkers, the drawback of pathway-based classifiers is that most of the human genes have not been assigned to a definitive pathway yet [12]. As pathways become more complete, the classification performances of pathway-based approaches will increase [18].

3.2. Hierarchical Clustering Results. After filtering 20148 probes in microarray data with t-test and obtaining the 100 and 200 most DE genes, hierarchical clustering was applied with complete linkage and Pearson correlation. The resulting clusters are shown in Figures 4 and 5. Since there are too many samples and genes, the gene names are not visible on these figures. Please refer to Supplementary Material (Table S20 and Table S21) to see the names and functions of these genes.

In Figure 4, cluster results of the 100 most DE genes, 5 T-ALL samples, and 1 healthy sample are misclassified, meaning that 5 T-ALL samples were grouped in healthy samples and 1 healthy sample was grouped in patient samples. Moreover, there are 1 T-ALL and 1 healthy sample, which are misclassified in clusters of the 200 most DE (Figure 5). This result is expected since 200 genes provide more information. However, the number of genes is much larger than the one we obtained in subnetworks.

A far more striking result is that 102 genes in our subnetworks and the 100 most DE genes have only 2 genes in common (Table 12). Furthermore, subnetworks and the 200 most DE genes have only 8 genes in common (Table 13). This result shows that 94 of our subnetwork genes are less

Table 12: The genes common to subnetwork genes and the 100 most differentially expressed genes.

Gene symbol	Gene title	GO: function	GO: process
CLU	Clusterin	Protein binding	Antiapoptosis, apoptosis, cell death, complement activation, classical pathway, endocrine pancreas development, innate immune response, lipid metabolic process, neurite morphogenesis, positive regulation of cell differentiation, positive regulation of cell proliferation, and response to oxidative stress
ITGA2B	Integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41)	Calcium ion binding, identical protein binding, protein binding, and receptor activity	Cell adhesion, cell adhesion, and integrin-mediated signaling pathway

Table 13: The genes common to subnetwork genes and the 200 most differentially expressed genes.

Gene symbol	Gene title	GO: function	GO: process
ADD2	Adducin 2 (beta)	Actin binding, calmodulin binding, and metal-ion binding	No enrichments
CD36	CD36 molecule (thrombospondin receptor)	Lipoprotein binding, low-density lipoprotein receptor activity, and receptor activity	Blood coagulation, cell adhesion, lipid metabolic process, lipoprotein transport, and transport
CLU	Clusterin	Protein binding	Antiapoptosis, apoptosis, cell death, complement activation, classical pathway, endocrine pancreas development, innate immune response, lipid metabolic process, neurite morphogenesis, positive regulation of cell differentiation, positive regulation of cell proliferation, and response to oxidative stress
GSN	Gelsolin (amyloidosis, Finnish type)	Actin binding, calcium ion binding, and protein binding	Actin filament polymerization, actin filament severing, and barbed-end actin filament capping
ITGA2B	Integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41)	Calcium ion binding, identical protein binding, protein binding, and receptor activity	Cell adhesion, cell adhesion, and integrin-mediated signaling pathway
LTB4R	Leukotriene B4 receptor	Leukotriene B4 receptor activity, nucleotide binding, and receptor activity	G-protein signaling, coupled to IP3 second messenger (phospholipase C activating), cell motility, immune response, inflammatory response, muscle contraction, and signal transduction
TLR4	Toll-like receptor 4	Lipopolysaccharide binding, protein binding, and transmembrane receptor activity	I-kappaB kinase/NF-kappaB cascade, T-helper 1 type immune response, detection of fungus, inflammatory response, innate immune response, macrophage activation, negative regulation of osteoclast differentiation, positive regulation of interleukin-12 biosynthetic process, positive regulation of interleukin-12 biosynthetic process, positive regulation of interleukin-8 biosynthetic process, positive regulation of tumor necrosis factor biosynthetic process, and signal transduction
TUBB1	Tubulin, beta-1	GTP binding, GTPase activity, nucleotide binding, and structural molecule activity	Microtubule-based movement, protein polymerization

Table 14: The classification accuracies of the 100 and 200 most differential genes between T-ALL and healthy samples.

Genes	J48	RBF network
100 most DE	95	98
200 most DE	95	98

differential than the 200 most DE genes. There are also 6 non-DE genes in our subnetworks (in the remaining 94 genes). Therefore, it would not be wrong to expect much higher classification accuracies from the 100 and 200 most DE genes than that of 102 subnetwork genes. But the classification accuracies of the 100 and 200 most DE genes are not very different from those of subnetworks. Actually, 2 subnetworks (subnetworks 12 and 13, Figure 2) performed even higher classification accuracies than the 100 and 200 most DE genes. Table 14 displays the classification accuracies of the 100 and 200 most DE genes. Thus, we can safely conclude that our subnetworks (even they are less differential than the 100 and 200 most DE genes and contain non-DE genes) can do the same or better job in distinguishing diseased samples from healthy samples. Doing the same job with 10 genes, instead of 100 or 200 genes, might be regarded as an accomplishment.

It is interesting that very well-known cancer genes such as TP53 and MAPK8 were not included in the 100/200 most DE genes, but we were able to detect them by network-based approach.

In conclusion, each subnetwork with high prediction accuracy provides a new suggestion for pathways and molecular mechanisms involved in the pathogenesis of T-ALL. All subnetworks serve as a biomarker which can be helpful in diagnosis and in identifying potential drug targets for T-ALL, in the near feature. The accomplishment of network-based classification and subnetwork/pathway detection is in line with the idea that cancer is not a result of solely one pathway, but instead it is a "disease of pathways" [12, 52, 120]. Unlike conventional differential expression analysis, network-based approach allowed us to identify potential disease-causing non-DE genes. According to our results, we conclude that transcription factors, tyrosine kinases, and zinc-ion binding proteins are the most important protein groups involved in generation of T-ALL.

The goal of this study is to highlight potential disease-causing genes for further experimental validation. It is beyond the scope of this study to verify all genes that appear in subnetworks; experimental proof is vital. We recommend investigators with an interest in a subnetwork/pathway to validate them with experimental techniques, like RT-PCR and Western blot. The important point of this work is that a combination of bioinformatic methods and high-throughput gene expression profiles and interactomics provide a promising way of identifying T-ALL specific modules and reveal pathways involved in T-ALL.

References

[1] C. Graux, J. Cools, L. Michaux, P. Vandenberghe, and A. Hagemeijer, "Cytogenetics and molecular genetics of T-cell acute lymphoblastic leukemia: from thymocyte to lymphoblast," *Leukemia*, vol. 20, no. 9, pp. 1496–1510, 2006.

[2] P. Van Vlierberghe, R. Pieters, H. B. Beverloo, and J. P. P. Meijerink, "Molecular-genetic insights in paediatric T-cell acute lymphoblastic leukaemia," *British Journal of Haematology*, vol. 143, no. 2, pp. 153–168, 2008.

[3] I. Aifantis, E. Raetz, and S. Buonamici, "Molecular pathogenesis of T-cell leukaemia and lymphoma," *Nature Reviews Immunology*, vol. 8, no. 5, pp. 380–390, 2008.

[4] F. J. T. Staal and A. W. Langerak, "Signaling pathways involved in the development of T-cell acute lymphoblastic leukemia," *Haematologica*, vol. 93, no. 4, pp. 493–497, 2008.

[5] A. A. Ferrando, D. S. Neuberg, J. Staunton et al., "Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia," *Cancer Cell*, vol. 1, no. 1, pp. 75–87, 2002.

[6] L. Espinosa, S. Cathelin, T. D'Altri et al., "The Notch/Hes1 pathway sustains NF-κB activation through *CYLD* repression in T cell leukemia," *Cancer Cell*, vol. 18, no. 3, pp. 268–281, 2010.

[7] T. Hoang and T. Hoang, "The T-ALL paradox in cancer," *Nature Medicine*, vol. 16, no. 11, pp. 1185–1186, 2010.

[8] T. Haferlach, W. Kern, S. Schnittger, and C. Schoch, "Modern diagnostics in acute leukemias," *Critical Reviews in Oncology/Hematology*, vol. 56, no. 2, pp. 223–234, 2005.

[9] T. Haferlach, A. Kohlmann, S. Schnittger et al., "Global approach to the diagnosis of leukemia using gene expression profiling," *Blood*, vol. 106, no. 4, pp. 1189–1198, 2005.

[10] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[11] R. K. Nibbe and M. R. Chance, "Approaches to biomarkers in human colorectal cancer: looking back, to go forward," *Biomarkers in Medicine*, vol. 3, no. 4, pp. 385–396, 2009.

[12] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, article 140, 2007.

[13] A. Kohlmann, T. J. Kipps, L. Z. Rassenti et al., "An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the microarray Innovations in Leukemia study prephase," *British Journal of Haematology*, vol. 142, no. 5, pp. 802–807, 2008.

[14] B. J. Wouters, B. Löwenberg, and R. Delwel, "A decade of genome-wide gene expression profiling in acute myeloid leukemia: flashback and prospects," *Blood*, vol. 113, no. 2, pp. 291–298, 2009.

[15] T. Haferlach, A. Kohlmann, L. Wieczorek et al., "Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the international microarray innovations in leukemia study group," *Journal of Clinical Oncology*, vol. 28, no. 15, pp. 2529–2537, 2010.

[16] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue classification with gene expression profiles," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 559–583, 2000.

[17] A. A. Alizadeh, M. B. Elsen, R. E. Davis et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[18] J. Su, B.-J. Yoon, and E. R. Dougherty, "Accurate and reliable cancer classification based on probabilistic inference of pathway activity," *PLoS One*, vol. 4, no. 12, p. e8161, 2009.

[19] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub, "A molecular signature of metastasis in primary solid tumors," *Nature Genetics*, vol. 33, no. 1, pp. 49–54, 2003.

[20] T. Ideker, V. Thorsson, A. F. Siegel, and L. E. Hood, "Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data," *Journal of Computational Biology*, vol. 7, no. 6, pp. 805–817, 2001.

[21] R. Jansen, D. Greenbaum, and M. Gerstein, "Relating whole-genome expression data with protein-protein interactions," *Genome Research*, vol. 12, no. 1, pp. 37–46, 2002.

[22] J. M. Raser and E. K. O'Shea, "Molecular biology—noise in gene expression: origins, consequences, and control," *Science*, vol. 309, no. 5743, pp. 2010–2013, 2005.

[23] V. M. Aris, M. J. Cody, J. Cheng et al., "Noise filtering and non-parametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer," *BMC Bioinformatics*, vol. 5, article 185, 2004.

[24] L. Klebanov and A. Yakovlev, "How high is the level of technical noise in microarray data?" *Biology Direct*, vol. 2, article 9, 2007.

[25] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 22, pp. 14031–14036, 2002.

[26] J. Chen and B. Yuan, "Detecting functional modules in the yeast protein-protein interaction network," *Bioinformatics*, vol. 22, no. 18, pp. 2283–2290, 2006.

[27] M. Gersten, M. Alirezaei, M. C. G. Marcondes et al., "An integrated systems analysis implicates EGR1 downregulation in simian immunodeficiency virus encephalitis-induced neural dysfunction," *Journal of Neuroscience*, vol. 29, no. 40, pp. 12467–12476, 2009.

[28] E. Segal, H. Wang, and D. Koller, "Discovering molecular pathways from protein interaction and gene expression data," *Bioinformatics*, vol. 19, no. 1, pp. i264–i272, 2003.

[29] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18, no. 1, pp. S233–S240, 2002.

[30] Y.-Q. Qiu, S. Zhang, X.-S. Zhang, and L. Chen, "Detecting disease associated modules and prioritizing active genes based on high throughput data," *BMC Bioinformatics*, vol. 11, article 26, 2010.

[31] A. Aderem, "Systems biology: its practice and challenges," *Cell*, vol. 121, no. 4, pp. 511–513, 2005.

[32] J. Gu, Y. Chen, S. Li, and Y. Li, "Identification of responsive gene modules by network-based gene clustering and extending: application to inflammation and angiogenesis," *BMC Systems Biology*, vol. 4, article 47, 2010.

[33] M. Vidal, "A biological atlas of functional maps," *Cell*, vol. 104, no. 3, pp. 333–339, 2001.

[34] Ş. Nacu, R. Critchley-Thorne, P. Lee, and S. Holmes, "Gene expression network analysis and applications to immunology," *Bioinformatics*, vol. 23, no. 7, pp. 850–858, 2007.

[35] L. Chen, J. Xuan, R. B. Riggins, Y. Wang, and R. Clarke, "Identifying protein interaction subnetworks by a bagging Markov random field-based method," *Nucleic Acids Research*, vol. 41, article e42, 2013.

[36] A. Ergün, C. A. Lawrence, M. A. Kohanski, T. A. Brennan, and J. J. Collins, "A network biology approach to prostate cancer," *Molecular Systems Biology*, vol. 3, article 82, 2007.

[37] T. Hwang, Z. Tian, R. Kuang, and J.-P. Kocher, "Learning on weighted hypergraphs to integrate protein interactions and gene expressions for cancer outcome prediction," in *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM '08)*, pp. 293–302, Pisa, Italy, December 2008.

[38] K. Lage, E. O. Karlberg, Z. M. Størling et al., "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nature Biotechnology*, vol. 25, no. 3, pp. 309–316, 2007.

[39] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS Computational Biology*, vol. 4, no. 11, Article ID e1000217, 2008.

[40] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J.-P. Vert, "Classification of microarray data using gene networks," *BMC Bioinformatics*, vol. 8, article 35, 2007.

[41] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[42] I. W. Taylor, R. Linding, D. Warde-Farley et al., "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nature Biotechnology*, vol. 27, no. 2, pp. 199–204, 2009.

[43] C. Staiger, S. Cadot, R. Kooter et al., "A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer," *PLoS One*, vol. 7, no. 4, Article ID e34796, 2012.

[44] X. Chang, T. Xu, Y. Li, and K. Wang, "Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of "date" and "party" hubs," *Scientific Reports*, vol. 3, article 1691, 2013.

[45] J. D. Han, N. Bertin, T. Hao et al., "Evidence for dynamically organized modularity in the yeast protein-protein interaction network," *Nature*, vol. 430, pp. 88–93, 2004.

[46] A. Bhattacharjee, W. G. Richards, J. Staunton et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13790–13795, 2001.

[47] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[48] W.-M. Liu, R. Li, J. Z. Sun et al., "PQN and DQN: algorithms for expression microarrays," *Journal of Theoretical Biology*, vol. 243, no. 2, pp. 273–278, 2006.

[49] T. S. Keshava Prasad, R. Goel, K. Kandasamy et al., "Human protein reference database—2009 update," *Nucleic Acids Research*, vol. 37, no. 1, pp. D767–D772, 2009.

[50] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[51] R. Shamir, A. Maron-Katz, A. Tanay et al., "EXPANDER—an integrative program suite for microarray data analysis," *BMC Bioinformatics*, vol. 6, article 232, 2005.

[52] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, 2000.

[53] P. H. Wiernik and American Cancer Society, *Adult Leukemias*, BC Decker, Lewiston, NY, USA, 2001.

[54] C. Sawyers, "Targeted cancer therapy," *Nature*, vol. 432, no. 7015, pp. 294–297, 2004.

[55] I. B. Weinstein and A. K. Joe, "Mechanisms of disease: oncogene addiction—a rationale for molecular targeting in cancer therapy," *Nature Clinical Practice Oncology*, vol. 3, no. 8, pp. 448–457, 2006.

[56] K. De Keersmaecker, C. Graux, M. D. Odero et al., "Fusion of EML1 to ABL1 in T-cell acute lymphoblastic leukemia with cryptic t(9;14)(q34;q32)," *Blood*, vol. 105, no. 12, pp. 4849–4852, 2005.

[57] C. Graux, J. Cools, C. Melotte et al., "Fusion of *NUP214* to *ABL1* on amplified episomes in T-cell acute lymphoblastic leukemia," *Nature Genetics*, vol. 36, no. 10, pp. 1084–1089, 2004.

[58] N. Mori, A. M. Krensky, K. Ohshima et al., "Elevated expression of CCL5/RANTES in adult T-cell leukemia cells: possible transactivation of the CCL5 gene by human T-cell leukemia virus type I tax," *International Journal of Cancer*, vol. 111, no. 4, pp. 548–557, 2004.

[59] M. N. Dworzak, G. Fröschl, D. Printz et al., "CD99 expression in T-lineage ALL: implications for flow cytometric detection of minimal residual disease," *Leukemia*, vol. 18, no. 4, pp. 703–708, 2004.

[60] K. De Keersmaecker, P. Marynen, and J. Cools, "Genetic insights in the pathogenesis of T-cell acute lymphoblastic leukemia," *Haematologica*, vol. 90, no. 8, pp. 1116–1127, 2005.

[61] S. Chiaretti, M. Messina, S. Tavolaro, and R. Foa, "Myeloid/T-cell acute lymphoblastic leukemia in children and adults," *Pediatric Reports*, vol. 3, supplement 2, article e3, 2011.

[62] B. Xu, X. Song, N. C. Yip et al., "Simultaneous detection of MDR1 and WT1 gene expression to predict the prognosis of adult acute lymphoblastic leukemia," *Hematology*, vol. 15, no. 2, pp. 74–80, 2010.

[63] R. Yagyu, R. Hamamoto, Y. Furukawa, H. Okabe, T. Yamamura, and Y. Nakamura, "Isolation and characterization of a novel human gene, VANGL1, as a therapeutic target for hepatocellular carcinoma," *International journal of oncology*, vol. 20, no. 6, pp. 1173–1178, 2002.

[64] R. D. Blumenthal, E. Leon, H. J. Hansen, and D. M. Goldenberg, "Expression patterns of CEACAM5 and CEACAM6 in primary and metastatic cancers," *BMC Cancer*, vol. 7, article 2, 2007.

[65] Y. Wang, S. Klumpp, H. M. Amin et al., "SSBP2 is an in vivo tumor suppressor and regulator of LDB1 stability," *Oncogene*, vol. 29, no. 21, pp. 3044–3053, 2010.

[66] V. G. Oehler, Y. Y. Ka, Y. E. Choi, R. E. Bumgarner, A. E. Raftery, and J. P. Radich, "The derivation of diagnostic markers of chronic myeloid leukemia progression from microarray data," *Blood*, vol. 114, no. 15, pp. 3292–3298, 2009.

[67] J. Soulier, E. Clappier, J.-M. Cayuela et al., "*HOXA* genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL)," *Blood*, vol. 106, no. 1, pp. 274–286, 2005.

[68] H. Ovaa, B. M. Kessler, U. Rolén, H. L. Ploegh, and M. G. Masucci, "Activity-based ubiquitin-specific protease (USP), profiling of virus-infected and malignant human cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 8, pp. 2253–2258, 2004.

[69] Y. Ge and M. T. Elghetany, "CD36: a multiligand molecule," *Laboratory Hematology*, vol. 11, no. 1, pp. 31–37, 2005.

[70] Y.-Y. Mo and S. J. Moschos, "Targeting Ubc9 for cancer therapy," *Expert Opinion on Therapeutic Targets*, vol. 9, no. 6, pp. 1203–1216, 2005.

[71] A. Martini, R. La Starza, H. Janssen et al., "Recurrent rearrangement of the Ewing's sarcoma gene, *EWSR1*, or its homologue, *TAF15*, with the transcription factor *CIZ/NMP4* in acute leukemia," *Cancer Research*, vol. 62, no. 19, pp. 5408–5412, 2002.

[72] C. S. Moreno, "The sex-determining region Y-Box 4 and homeobox C6 transcriptional networks in prostate cancer progression: crosstalk with the Wnt, Notch, and PI3K pathways," *American Journal of Pathology*, vol. 176, no. 2, pp. 518–527, 2010.

[73] S. Strehl, A. Borkhardt, R. Slany, U. E. Fuchs, M. König, and O. A. Haas, "The human LASP1 gene is fused to MLL in an acute myeloid leukemia with t(11;17)(q23;q21)," *Oncogene*, vol. 22, no. 1, pp. 157–160, 2003.

[74] E. Adinolfi, L. Melchiorri, S. Falzoni et al., "$P2X_7$ receptor expression in evolutive and indolent forms of chronic B lymphocytic leukemia," *Blood*, vol. 99, no. 2, pp. 706–708, 2002.

[75] G. Cabrini, S. Falzoni, S. L. Forchap et al., "A His-155 to tyr polymorphism confers gain-of-function to the human $P2X_7$ receptor of human leukemic lymphocytes," *Journal of Immunology*, vol. 175, no. 1, pp. 82–89, 2005.

[76] J. S. Wiley, L. P. Dao-Ung, B. J. Gu et al., "A loss-of-function polymorphic mutation in the cytolytic $P2X_7$ receptor gene and chronic lymphocytic leukaemia: a molecular study," *Lancet*, vol. 359, no. 9312, pp. 1114–1119, 2002.

[77] C. O. Souza, G. F. Santoro, V. R. Figliuolo et al., "Extracellular ATP induces cell death in human intestinal epithelial cells," *Biochimica et Biophysica Acta*, vol. 1820, pp. 1867–1878, 2012.

[78] U. Thunberg, G. Tobin, A. Johnson et al., "Polymorphism in the $P2X_7$ receptor gene and survival in chronic lymphocytic leukaemia," *Lancet*, vol. 360, no. 9349, pp. 1935–1939, 2002.

[79] E. Vacchelli, L. Galluzzi, V. Rousseau et al., "Loss-of-function alleles of P2RX7 and TLR4 fail to affect the response to chemotherapy in non-small cell lung cancer," *Oncoimmunology*, vol. 1, pp. 271–278, 2012.

[80] E. Adinolfi, M. G. Callegari, M. Cirillo et al., "Expression of the $P2X_7$ receptor increases the Ca^{2+} content of the endoplasmic reticulum, activates NFATc1, and protects from apoptosis," *Journal of Biological Chemistry*, vol. 284, no. 15, pp. 10120–10128, 2009.

[81] E. Adinolfi, L. Raffaghello, A. L. Giuliani et al., "Expression of $P2X_7$ receptor increases in vivo tumor growth," *Cancer Research*, vol. 72, pp. 2957–2969, 2012.

[82] X.-J. Zhang, G.-G. Zheng, X.-T. Ma et al., "Expression of $P2X_7$ in human hematopoietic cell lines and leukemia patients," *Leukemia Research*, vol. 28, no. 12, pp. 1313–1322, 2004.

[83] S. Rösmann, D. Hahn, D. Lottaz, M. N. Kruse, W. Stöcker, and E. E. Sterchi, "Activation of human meprin-α in a cell culture model of colorectal cancer is triggered by the plasminogen-activating system," *Journal of Biological Chemistry*, vol. 277, no. 43, pp. 40650–40658, 2002.

[84] M. Ranson and N. M. Andronicos, "Plasminogen binding and cancer: promises and pitfalls," *Frontiers in Bioscience*, vol. 8, pp. s294–s304, 2003.

[85] J. S. Bond, G. L. Matters, S. Banerjee, and R. E. Dusheck, "Meprin metalloprotease expression and regulation in kidney, intestine, urinary tract infections and cancer," *FEBS Letters*, vol. 579, no. 15, pp. 3317–3322, 2005.

[86] D. Lottaz, C. A. Maurer, A. Noël et al., "Enhanced activity of meprin-α, a pro-migratory and pro-angiogenic protease, in colorectal cancer," *PLoS One*, vol. 6, no. 11, Article ID e26450, 2011.

[87] E. E. Sterchi, W. Stöcker, and J. S. Bond, "Meprins, membrane-bound and secreted astacin metalloproteinases," *Molecular Aspects of Medicine*, vol. 29, no. 5, pp. 309–328, 2009.

[88] M. Ishihara, C. Nishida, Y. Tashiro et al., "Plasmin inhibitor reduces T-cell lymphoid tumor growth by suppressing matrix metalloproteinase-9-dependent $CD11b^+$ /$F4/80^+$ myeloid cell recruitment," *Leukemia*, vol. 26, no. 2, pp. 332–339, 2012.

[89] M. O. Khan and M. H. Ather, "Chromogranin A—serum marker for prostate cancer," *Journal of the Pakistan Medical Association*, vol. 61, no. 1, pp. 108–111, 2011.

[90] O. Louthan, "Chromogranin a in physiology and oncology," *Folia Biologica*, vol. 57, pp. 173–181, 2011.

[91] K. Jurianz, S. Ziegler, H. Garcia-Schüler et al., "Complement resistance of tumor cells: basal and induced mechanisms," *Molecular Immunology*, vol. 36, no. 13-14, pp. 929–939, 1999.

[92] Y. Paas, O. Bohana-Kashtan, and Z. Fishelson, "Phosphorylation of the complement component, C9, by an ecto-protein kinase of human leukemic cells," *Immunopharmacology*, vol. 42, no. 1–3, pp. 175–185, 1999.

[93] M. Schlesinger, I. Broman, and G. Lugassy, "The complement system is defective in chronic lymphatic leukemia patients and in their healthy relatives," *Leukemia*, vol. 10, no. 9, pp. 1509–1513, 1996.

[94] A. P. Gilmore and L. H. Romer, "Inhibition of focal adhesion kinase (FAK) signaling in focal adhesions decreases cell motility and proliferation," *Molecular Biology of the Cell*, vol. 7, no. 8, pp. 1209–1224, 1996.

[95] L. J. Jarvis, J. E. Maguire, and T. W. LeBien, "Contact between human bone marrow stromal cells and B lymphocytes enhances very late antigen-4/vascular cell adhesion molecule-1-independent tyrosine phosphorylation of focal adhesion kinase, paxillin, and ERK2 in stromal cells," *Blood*, vol. 90, no. 4, pp. 1626–1635, 1997.

[96] S. Ozkal, J. C. Paterson, S. Tedoldi et al., "Focal adhesion kinase (FAK) expression in normal and neoplastic lymphoid tissues," *Pathology Research and Practice*, vol. 205, no. 11, pp. 781–788, 2009.

[97] P. J. Reddig and R. L. Juliano, "Clinging to life: cell to matrix adhesion and cell survival," *Cancer and Metastasis Reviews*, vol. 24, no. 3, pp. 425–439, 2005.

[98] V. Gabarra-Niecko, M. D. Schaller, and J. M. Dunty, "FAK regulates biological processes important for the pathogenesis of cancer," *Cancer and Metastasis Reviews*, vol. 22, no. 4, pp. 359–374, 2003.

[99] G. W. McLean, N. O. Carragher, E. Avizienyte, J. Evans, V. G. Brunton, and M. C. Frame, "The role of focal-adhesion kinase in cancer—a new therapeutic opportunity," *Nature Reviews Cancer*, vol. 5, no. 7, pp. 505–515, 2005.

[100] M. Y. Gordon, C. R. Dowding, G. P. Riley, J. M. Goldman, and M. F. Greaves, "Altered adhesive interactions with marrow stroma of haematopoietic progenitor cells in chronic myeloid leukemia," *Nature*, vol. 328, no. 6128, pp. 342–344, 1987.

[101] M. Takahashi, A. Keating, and J. W. Singer, "A functional defect in irradiated adherent layers from chronic myelogenous leukemia long-term marrow cultures," *Experimental Hematology*, vol. 13, no. 9, pp. 926–931, 1985.

[102] A. Witkiewicz-Kucharczyk and W. Bal, "Damage of zinc fingers in DNA repair proteins, a novel molecular mechanism in carcinogenesis," *Toxicology Letters*, vol. 162, no. 1, pp. 29–42, 2006.

[103] S. N. Loh, "The missing Zinc: P53 misfolding and cancer," *Metallomics*, vol. 2, no. 7, pp. 442–449, 2010.

[104] E. Mocchegiani, L. Costarelli, R. Giacconi, C. Cipriano, E. Muti, and M. Malavolta, "Zinc-binding proteins (metallothionein and α-2 macroglobulin) and immunosenescence," *Experimental Gerontology*, vol. 41, no. 11, pp. 1094–1107, 2006.

[105] G. Şahin, U. Ertem, F. Duru, D. Birgen, and N. Yüksek, "High prevalence of chronic magnesium deficiency in T cell lymphoblastic leukemia and chronic zinc deficiency in children with acute lymphoblastic leukemia and malignant lymphoma," *Leukemia and Lymphoma*, vol. 39, no. 5-6, pp. 555–562, 2000.

[106] G. A. Eby, "Treatment of acute lymphocytic leukemia using zinc adjuvant with chemotherapy and radiation—a case history and hypothesis," *Medical Hypotheses*, vol. 64, no. 6, pp. 1124–1126, 2005.

[107] S. T. Martin, N. Sato, S. Dhara et al., "Aberrant methylation of the Human Hedgehog interacting protein (HHIP) gene in pancreatic neoplasms," *Cancer Biology and Therapy*, vol. 4, no. 7, pp. 728–733, 2005.

[108] R. P. Young, C. F. Whittington, R. J. Hopkins et al., "Chromosome 4q31 locus in COPD is also associated with lung cancer," *European Respiratory Journal*, vol. 36, no. 6, pp. 1375–1382, 2010.

[109] M. H. Shahi, M. Afzal, S. Sinha et al., "Human hedgehog interacting protein expression and promoter methylation in medulloblastoma cell lines and primary tumor samples," *Journal of Neuro-Oncology*, vol. 103, no. 2, pp. 287–296, 2011.

[110] C. L. Olsen, P.-P. Hsu, J. Glienke, G. M. Rubanyi, and A. R. Brooks, "Hedgehog-interacting protein is highly expressed in endothelial cells but down-regulated during angiogenesis and in several human tumors," *BMC Cancer*, vol. 4, article 43, 2004.

[111] M. Tada, F. Kanai, Y. Tanaka et al., "Down-regulation of hedgehog-interacting protein through genetic and epigenetic alterations in human hepatocellular carcinoma," *Clinical Cancer Research*, vol. 14, no. 12, pp. 3768–3776, 2008.

[112] L. J. Leandro-García, S. Leskelä, I. Landa et al., "Tumoral and tissue-specific expression of the major human β-tubulin isotypes," *Cytoskeleton*, vol. 67, no. 4, pp. 214–223, 2010.

[113] S. Yin, R. Bhattacharya, and F. Cabral, "Human mutations that confer paclitaxel resistance," *Molecular Cancer Therapeutics*, vol. 9, no. 2, pp. 327–335, 2010.

[114] Y. Okada, Q. Jiang, M. Lemieux, L. Jeannotte, L. Su, and Y. Zhang, "Leukaemic transformation by *CALM-AF10* involves upregulation of *HOXA5* by hDOT1L," *Nature Cell Biology*, vol. 8, no. 9, pp. 1017–1024, 2006.

[115] W. A. Dik, W. Brahim, C. Braun et al., "*CALM-AF10*+ T-ALL expression profiles are characterized by overexpression of *HOXA* and *BMI1* oncogenes," *Leukemia*, vol. 19, no. 11, pp. 1948–1957, 2005.

[116] K. W. L. Yee, A. Hagey, S. Verstovsek et al., "Phase 1 study of ABT-751, a novel microtubule inhibitor, in patients with refractory hematologic malignancies," *Clinical Cancer Research*, vol. 11, no. 18, pp. 6615–6624, 2005.

[117] I. Lahortiga, J. L. Vizmanos, X. Agirre et al., "NUP98 is fused to Adducin 3 in a patient with T-cell acute lymphoblastic leukemia and myeloid markers, with a new translocation t(10;11)(q25;p15)," *Cancer Research*, vol. 63, no. 12, pp. 3079–3083, 2003.

[118] F. J. T. Staal and J. M. Sen, "The canonical Wnt signaling pathway plays an important role in lymphopoiesis and hematopoiesis,"

European Journal of Immunology, vol. 38, no. 7, pp. 1788–1794, 2008.

[119] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[120] E. F. Petricoin III, V. E. Bichsel, V. S. Calvert et al., "Mapping molecular networks using proteomics: a vision for patient-tailored combination therapy," *Journal of Clinical Oncology*, vol. 23, no. 15, pp. 3614–3621, 2005.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 910941, 9 pages
http://dx.doi.org/10.1155/2013/910941

Research Article

Robust Cell Size Checkpoint from Spatiotemporal Positive Feedback Loop in Fission Yeast

Jie Yan,[1] Xin Ni,[2] and Ling Yang[1,2]

[1] *School of Mathematical Sciences, Soochow University, Suzhou 215006, China*
[2] *Center for Systems Biology, Soochow University, Suzhou 215006, China*

Correspondence should be addressed to Ling Yang; lyang@suda.edu.cn

Received 29 April 2013; Accepted 7 June 2013

Academic Editor: Tao Huang

Cells must maintain appropriate cell size during proliferation. Size control may be regulated by a size checkpoint that couples cell size to cell division. Biological experimental data suggests that the cell size is coupled to the cell cycle in two ways: the rates of protein synthesis and the cell polarity protein kinase Pom1 provide spatial information that is used to regulate mitosis inhibitor Wee1. Here a mathematical model involving these spatiotemporal regulations was developed and used to explore the mechanisms underlying the size checkpoint in fission yeast. Bifurcation analysis shows that when the spatiotemporal regulation is coupled to the positive feedback loops (active Cdc2 promotes its activator, Cdc25, and suppress its inhibitor, Wee1), the mitosis-promoting factor (MPF) exhibits a bistable steady-state relationship with the cell size. The switch-like response from the positive feedback loops naturally generates the cell size checkpoint. Further analysis indicated that the spatial regulation provided by Pom1 enhances the robustness of the size checkpoint in fission yeast. This was consistent with experimental data.

1. Introduction

In order to maintain proper size, dividing cells need to time mitosis carefully. Previous analyses performed in fission yeast suggested that there is a homeostatic mechanism that can maintain the appropriate cell size [1–3]. The cell is allowed to enter mitosis only after it reaches a critical size (size checkpoint). Experimental data also showed that cells smaller than critical size had to grow until they reached the threshold value. This period is called the size-dependent phase, or sizer. Then, after a fixed period, called as timer, the cells completed mitosis. Daughter cells that are larger than critical size when produced can undergo mitosis without going through the sizer phase [3]. Some higher eukaryotes such as *Xenopus laevis* [4, 5], Drosophila [6], animal cells [7], and HeLa cells [8] also have similar methods of size control.

Biological experimental data indicate that the rate of cyclin protein synthesis may increase as the cell grows [9]. This may be one mechanism underlying size control. Previous mathematical models have explored the nonlinear dynamic properties of the temporal regulation of cell cycle events [10–12]. The cyclin protein synthesis rate is assumed to increase as the cell grows, and it exhibits a bistable relationship with MPF. This bistability, which is generated by the positive feedback loops in the cell cycle, is responsible for the mitosis initialization [9, 13]. In this way, cell size is linked to entry into mitosis.

Recent evidence has shown that the cell polarity protein kinase Pom1 forms a polar gradient from the ends of the cylindrical cell to its center [14, 15]. In this way, it can provide spatial information that can be used to regulate the mitosis inhibitor Wee1. This spatial regulation links cell size directly to mitosis, and it may play a critical role in size control.

In summary, cell size is coupled to the progression of the cell cycle through the rates of synthesis of cyclin proteins and the direct spatial information provided by Pom1. The results of the present study show that when spatial regulation and the rate of synthesis rate are both coupled to temporal positive feedback loops, a bistable response generates the cell size checkpoint. Bifurcation analysis shows that the concentration of MPF can exhibit a bistable steady-state relationship with the rate of synthesis of cyclin proteins or the concentration of Cdr (downstream of Pom1) alone. The size checkpoint is naturally built into the system in the form of dual regulations of the rate of synthesis and the Pom1 gradient. Stochastic analysis then showed that the direct spatial regulation can allow

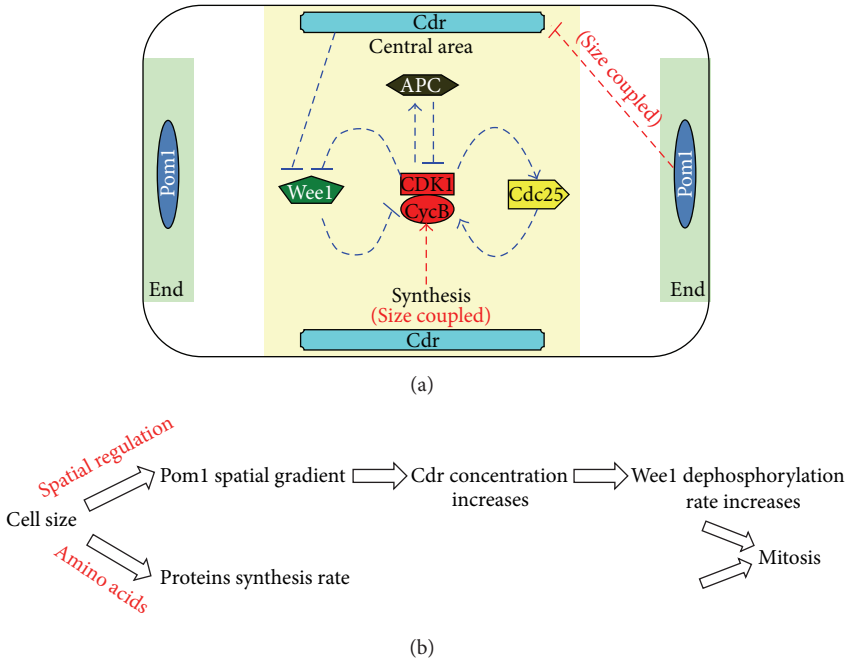


Figure 1: (a) Regulatory network of the cell cycle in fission yeast. (b) Two ways in which cell growth is coupled to cell division: the rate of synthesis of Cdc13 and the direct spatial regulation provided by Pom1.

temporal positive feedback to enhance the robustness of the cell size checkpoint in fission yeast, which is consistent with the experimental data.

2. Results and Discussion

2.1. Mathematical Modeling. The upper panel of Figure 1(a) shows a schematic diagram of the protein interaction in G2-M phase. The regulatory network includes a negative feedback loop: active Cdc2/Cdc13 dimer (MPF) inhibits itself by promoting the production of APC complexes and thus promotes cyclin ubiquitination and degradation. In addition to this negative feedback loop, regulation during the G2/M phase also involves two positive feedback loops: there is a phosphorylatable tyrosine residue (the Tyr-15 residue) at the active site of Cdc2. If the active site is phosphorylated, MPF is inactive. Wee1, a kind of tyrosine kinases, can inactivate Cdc2/Cdc13 (MPF) in this way. MPF can also phosphorylate Wee1 to repress its activity. On the other side, tyrosine phosphatases Cdc25 can remove the inhibitory phosphate group on Tyr-15 to activate MPF. In return, MPF can promote the activity of Cdc25 by increasing the phosphorylation rate of Cdc25. In summary, active Cdc2/Cdc13 activates its activator Cdc25 and inactivates its inhibitor Wee1. During the cell cycle, Cdc13 is continuously synthesized from amino acids. The rate of synthesis of Cdc13 increases as the cell grows (Figure 1(b)).

Besides, Cdr proteins also couple cell growth to cell division through a size sensing mechanism involving Pom1. Several previous works have identified the function of the Pom1 pathway [14–16]. The cell polarity protein kinase Pom1 is a cell polarity protein kinase, which can form a spatial gradient that is greatest at the ends of the cylindrical cell and least in the middle of the cell. Cdr which locates near the center of the cell can suppress the activity of Wee1 and so promote mitosis. Pom1 phosphorylates Cdr to inhibit its activity. The size-dependent relief of this inhibition can repress Wee1 to promote the initialization of mitosis.

The network was then transferred into a set of ordinary differential equations using the principles of biochemical kinetics. The initial size of a WT daughter cell was normalized to 1. The model was adapted from the models constructed by James Ferrell's group [17] and Novak-Tyson's group [11]. However, different from their models, we also took the spatial information provided by Pom1 into consideration.

A detailed mathematical model is presented in Section 4.

2.2. Bifurcation Analysis. Experimental observations have provided some evidences of size checkpoint [2]. If the initial size of a *cdc2-33* fission yeast cell was smaller than 12 μm, a marked negative relationship was observed between the extension length and the initial size. However, the extension length was not found to be significantly related to cell size at initial sizes larger than 12 μm. This critical size that determined whether the cell could begin mitosis was the size checkpoint. Besides, Rupeš and colleagues also showed that cells smaller than critical size had to grow until they reached the threshold value. If the birth size of the fission yeast is larger than the critical size, the cell can undergo mitosis without additional time delay [3]. This critical size also indicates the existence of the size checkpoint.

Earlier experimental studies have revealed that the steady state of MPF shows a hysteretic steady-state response relationship with the concentration of cyclin B [13]. Mathematical

models have established that the concentration of MPF has a bistable relationship with the rate of synthesis of the cyclin proteins. This bistability is attributable to the positive feedback loops (active Cdc2 promotes its activator Cdc25 and suppresses its inhibitor Wee1) [18, 19]. In our model, the positive feedback loops rely both on the rate of synthesis and on the spatial regulation involving Wee1. In this way, the coupling between cell size and cell division is more realistic in this model.

To demonstrate how the rate of synthesis of Cdc13 and the concentration of Cdr both affect the activation of MPF, we first calculated the steady state of MPF for a given rate of Cdc13 synthesis and a given concentration of Cdr, when Cdc13 was made nondegradable (during the period prior to mitosis, the concentration of APC remains at a low and constant level) [20]. The results are presented in three-dimensional space (the green surfaces in Figures 2(a), 2(c), and 2(e)).

If the regulation related to the rate of synthesis is solely considered, then the vertical plane Cdr = 0.2 (which represents a fixed concentration of Cdr, which occurs when Pom1 spatial regulation is blocked) intersects with the surface at an S-shaped curve (Figure 2(a)). The bifurcation analysis shows that the steady state of MPF has a bistable relationship with the rate of synthesis (Figure 2(b)). As the cell grows, the rate of synthesis of Cdc13 increases and the concentration of MPF accumulates in turn. When the rate of synthesis passes point K2 in Figure 2(b) as the cell grows, the low stable branch disappears and the MPF has to jump to the upper stable branch (arrow (1)). And the mitosis begins. In this way, the rate of synthesis of Cdc13 contributes to the function of the size checkpoint.

Similarly, if the regulation related to the Pom1 pathway is solely considered, the vertical plane synthesis rate = 0.009 (which represents a fixed synthesis rate of Cdc13) intersects with the bent surface along an S-shaped line (Figure 2(c)). The bifurcation analysis shows that the steady state of MPF also has a bistable relationship with Cdr (the downstream of Pom1) when the direct spatial regulation is linked to the positive feedback loops (Figure 2(d)). As the cell grows, the concentration of Cdr increases (due to a reduction in regulation provided by Pom1). Then the concentration of MPF also accumulates along the lower branch. When the concentration of Cdr passes point C2 as the cell grows (Figure 2(d)), the low stable branch disappears and MPF has to jump to the upper stable branch (arrow (2)). Then mitosis begins. In this way, the direct spatial regulation provided by the Pom1 pathway also contributes to the function of the size checkpoint.

In the real-world cell cycle, these two previous regulations both contribute to the coupling of the cell size and cell division. The steady state of MPF in real-world systems was assessed as follows. First, the relationship between the rate of synthesis of Cdc13 and the concentration of Cdr as the cell grows was calculated. Then the vertical surface, which represents the variation in the rate of synthesis rate and the concentration of Cdr as the cell grows, was intersected with the steady state surface (Figure 2(e)). As in the sole regulation scenarios, the line of intersection is S-shaped. This means that the steady state of MPF continues to exhibit bistability with cell size when the spatial regulation and the rate of synthesis are both involved in the positive feedback loops. After that we directly linked the steady state of MPF to the cell size through bifurcation analysis (Figure 2(f)). Figure 2(f) shows that the concentration of MPF increases as the cell grows. When the cell size reaches the threshold, about 1.5 (size checkpoint, S2 in Figure 2(f)), MPF switches to the upper branch (arrow (3)). Then the cell undergoes mitosis.

Then we further summarized the relationship between the bifurcation analysis and the size checkpoint. The bifurcation analysis shows that MPF exhibits the bistability with the cell size. There is a critical cell size S2 (corresponding to saddle node point SN2): if the cell size is smaller than S2, MPF stays in low level; if the cell size passes S2 point, the low stable branch disappears and MPF has to jump to the upper stable branch. As we mentioned previously, experimental studies [1, 2, 21, 22] have shown that a cell will not begin mitosis until it grows to a critical size. Therefore, this saddle node point SN2 naturally performs the role of a check point: before the size reaches S2, cell remains in G2 state (low MPF); once the cell passes S2 point, MPF can jump to the upper branch to trigger mitosis.

After that, a numerical simulation was used to check the size checkpoint (Figure 3). The initial size of the model varied from 0.25 to 4. The result shows that when the initial cell size is smaller than 1.5, the cycle time shows a significantly negative relationship with the initial size. However, if the initial size exceeded 1.5, then the cycle time was mostly independent of the initial cell size. This result accords with the previous experimental data in yeast (the inserted figure in Figure 3 [2]). In this way, 1.5 is established as the size checkpoint. The result of the simulation is also consistent with the bifurcation analysis shown in Figure 2(f), where the critical size for the mitosis initialization is about 1.5.

In summary, the concentration of MPF exhibits a bistable steady-state relationship with cell size, which depends on the spatiotemporal positive feedbacks. This bistability naturally produces the size checkpoint.

2.3. Stochastic Analysis. Experimental evidence showed that some intrinsic stochastic noise (such as random cell production and collisions between molecules) and extrinsic stochastic noise (such as variations in the environment) will result in fluctuations in gene expression [23]. In this way, processes related to the cell cycle may vary from cell to cell within a population, over time, and even within a single cell. The present study not only coupled cell size to the rate of synthesis of Cdc13 but also to the direct spatial regulation provided by the Pom1 pathway. This direct spatial regulation may help the size checkpoint resist interference from different sources and keep cell size coupled to cell division.

To evaluate the impact of random fluctuation on the cell cycle, some stochastic noise was introduced to the present model: (1) each parameter in the deterministic model was multiplied by a stochastic factor, which was randomly chosen from the normal distribution with $\mu = 1$ and $\sigma = 0.016$ (μ represents the mean value and σ represents the variance of the distribution). In this way, the cell cycle can fluctuate near the deterministic value. (2) After mitosis, the cell did not divide

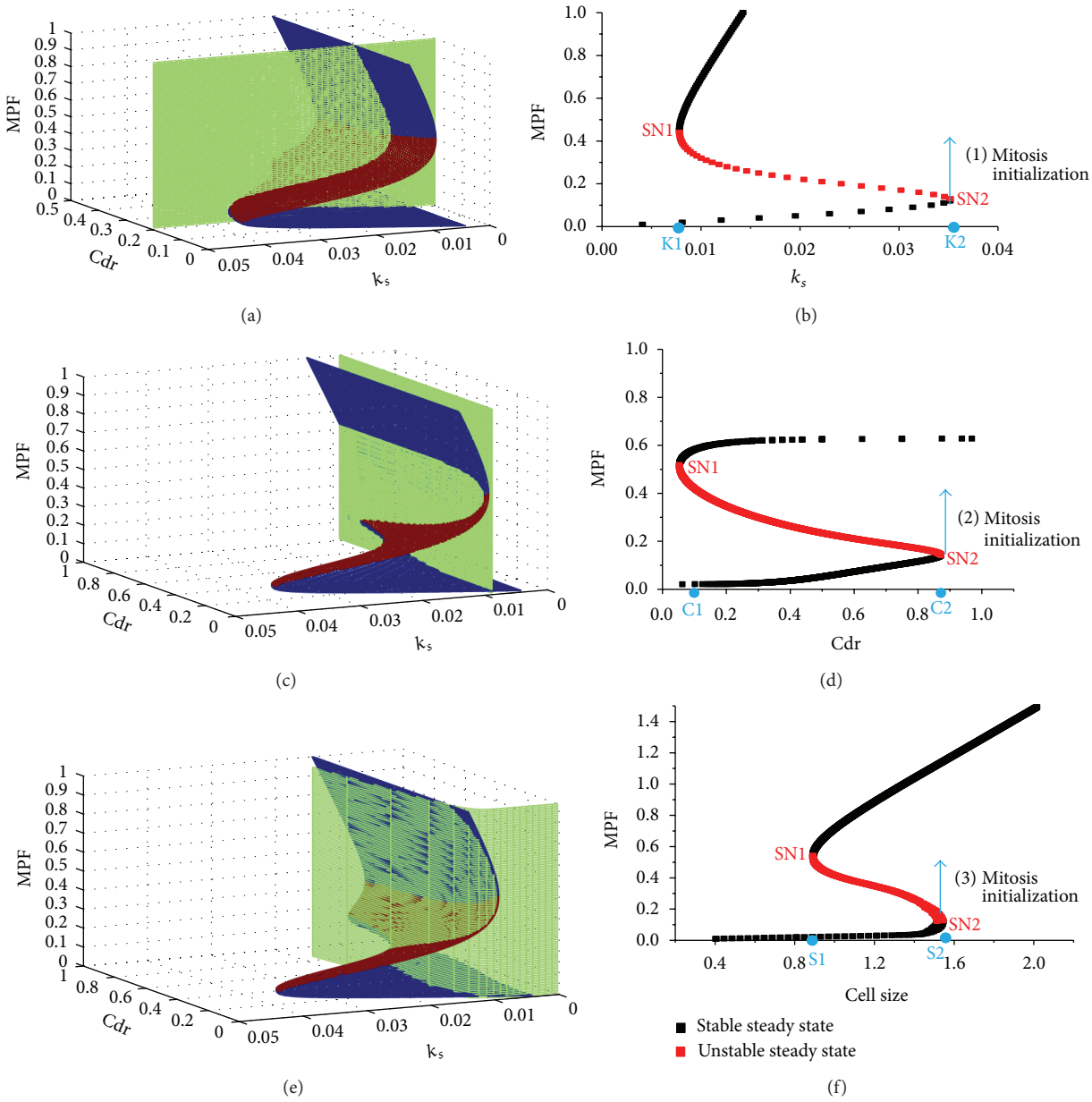


FIGURE 2: (a)(c)(e) The steady state of MPF with different given rates of synthesis (k_s) and different given concentrations of Cdr (cdr). The dark blue areas indicate the stable steady state of MPF, and the dark red areas indicate the unstable steady state of MPF. (a) The green vertical plane represents the variation in the rates of Cdc13 synthesis as the cell grows for a fixed concentration of Cdr. The plane intersects with the bent surface, forming an S-shaped curve. (b) When this curve is displayed in two-dimensional space, it represents the relationship between MPF and the rate of synthesis. (c) The green vertical plane represents the variation in the concentration of Cdr as the cell grows for a fixed rate of Cdc13 synthesis. The plane intersects with the bent surface, forming an S-shaped curve. (d) When this curve is displayed in two-dimensional space, it represents the relationship between MPF and the concentration of Cdr. (e) The vertical green surface represents the relationship between the concentration of Cdr and the rate of synthesis of Cdc13 as the cell grows. The vertical green surface intersects with the bent surface, forming an S-shaped curve. (f) When this curve is displayed in two-dimensional space, it represents the relationship between MPF and cell size.

into two identical daughter cells. Asymmetrical division was characterized by a normal distribution with $\mu = 0.5$ and $\sigma = 0.016$.

Then the model was used to determine if the spatial regulation can help the size checkpoint resist the fluctuations of the system. When the stochastic factor was disturbed, the size check point was calculated 100 times with and without the Pom1 spatial regulation. Results are shown in Figure 4.

Figure 4(a) shows that if there is no spatial regulation in the system (i.e., if cell size is linked to cell division only through the rate of synthesis and if the concentration of Cdr is fixed at 0.5), the size checkpoint varies from 1.15 to 1.55 in

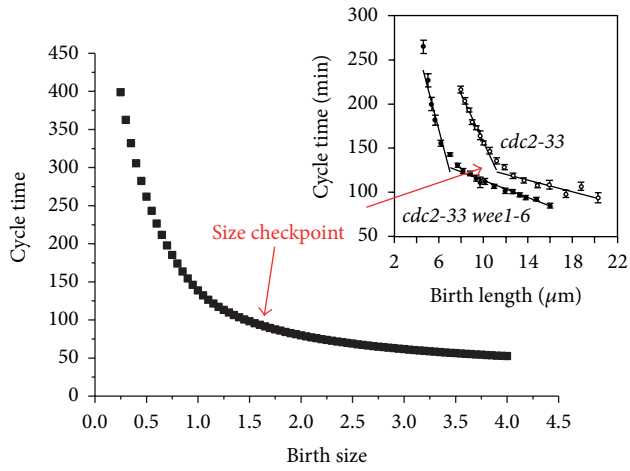


Figure 3: The numerical simulation was used to confirm the size checkpoint. The x-axis represents the initial size of the fission yeast in the model. The y-axis represents the points in the cell cycle when the birth size changes from 0.25 to 4. The red arrow indicates the size checkpoint. A previous experimental data [2] is inserted to compare with the simulation.

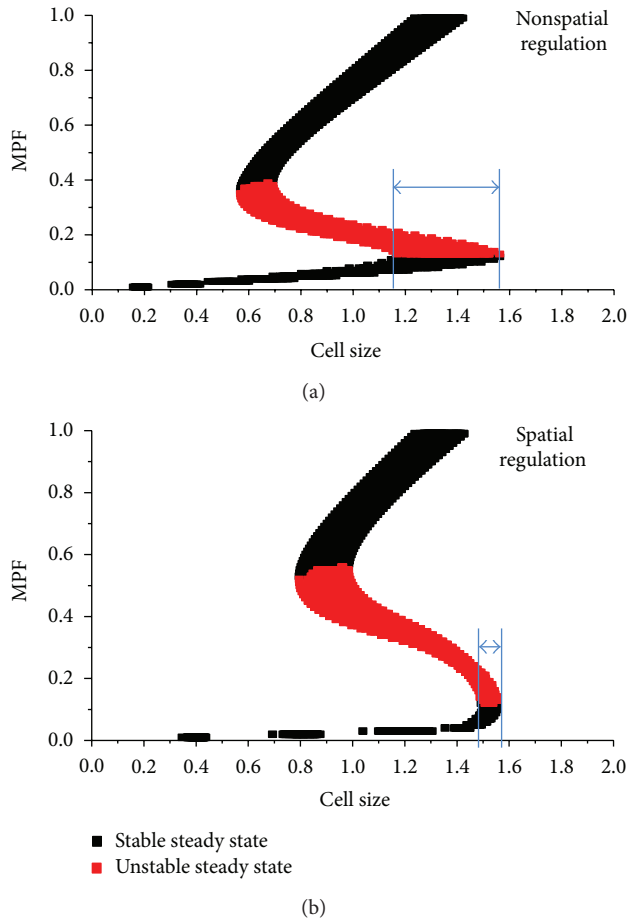


Figure 4: The relationship between the steady state of MPF and cell size in (a) a system without spatial regulation and (b) a system with spatial regulation. The red points represent the unstable steady state of MPF, and the black points represent the stable steady state of MPF. The blue bar indicates the range of the size checkpoint.

the presence of stochastic noise. However, if spatial regulation is taking place in the system, the size checkpoint changes from 1.48 to 1.56, which is much narrower than the one shown in Figure 4(a). In this way, even with the interference produced by stochastic noise, the cell must still exceed a strict size checkpoint. The comparison indicates that the direct spatial regulation provided by the Pom1 pathway can ensure tight coupling between cell size and the cell division.

Because direct spatial regulation works through the mitosis inhibitor Wee1, *wee1*Δ may show a weak ability to resist interference. This mathematical model was used to assess the size checkpoint in *wee1*Δ ($k5 = 0.15$, decreases from 2 in WT, the stochastic noise presents as mentioned previously). Bifurcation analysis shows the size at which cells undergo mitosis in *wee1*Δ to be about half of that in WT (Figure 5(a)). This result accords with the previous experimental data that the *wee1*Δ cell divided at a half size of WT [2]. However, in the presence of random disturbances, the width of the size checkpoint in *wee1*Δ was found to be twice of that of WT. These theoretical results are consistent with observations made in earlier experiments [2]. Experimental observation showed that *wee1*Δ fission yeast exhibited larger variance in the duration of the cell cycle for any given initial size (Figure 1(b) of a previous study) [2]. Because the duration of the cell cycle includes the time required for the cell to reach the size checkpoint (sizer phase) and the fixed time, which is independent of other factors (timer phase) [24]. This indicates that the variation of the size checkpoint is larger in *wee1*Δ than in WT. Furthermore, Table 1 in [2] summarized that the variation in the division length was about twice as large in *wee1*Δ as in WT, and the variation in cycle time was increased in the similar way. Therefore, it means that the size control in *wee1*Δ is not as strict as that in WT. In this way, the size checkpoint in *wee1*Δ is less robust than in WT.

Then the numerical simulation results of *wee1*Δ and WT fission yeast were compared to those produced in earlier experiments. During the simulation, stochastic factors continued to act on the cell. And the initial concentrations of proteins and initial size were given in the deterministic model. After every division, the system listed the initial cell size and initial concentrations of relevant proteins for the next cell cycle. The results of the simulation are shown in Figure 5(b): the overall range of the duration of the cell cycle was similar in *wee1*Δ and WT. However, for a given initial size, the range of the duration of the cell cycle was always larger in *wee1*Δ than in WT. The large variation in the length of the cycle time can be attributed to weakness in the size checkpoint control. These results are consistent with those of a previous experiment published by Novak and Tyson [2].

In conclusion, the direct spatial regulation provided by Pom1 can enhance the robustness of the size checkpoint and couple cell size to cell division.

3. Discussion

Although a group of models have investigated the temporal regulation of cell cycle [10, 25–27], most of them did not

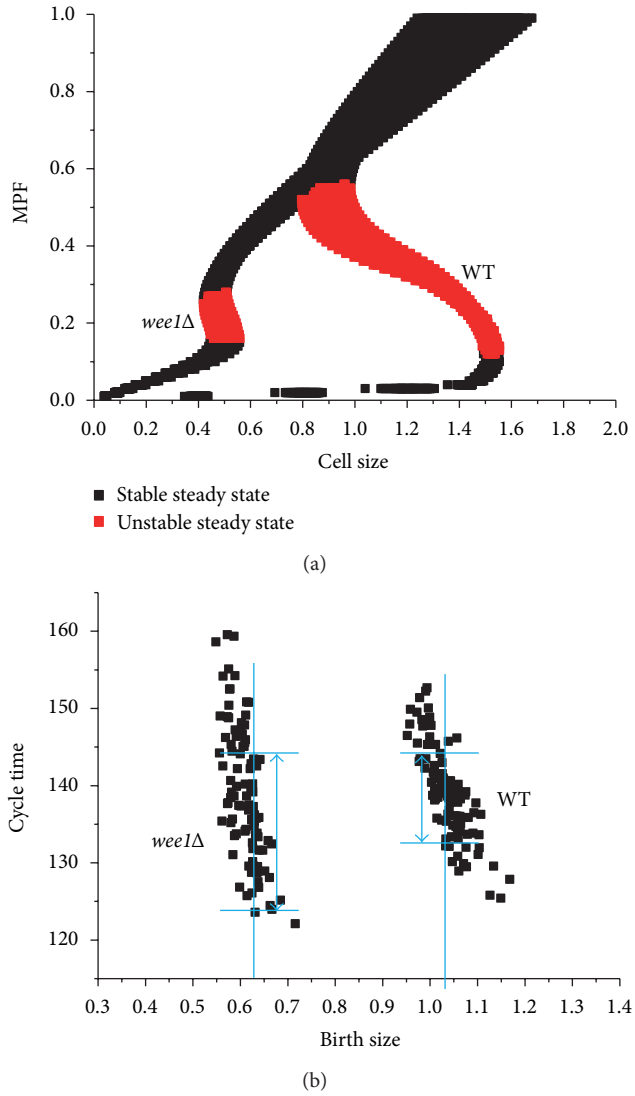


FIGURE 5: (a) Bifurcation analysis in *wee1*Δ and in WT fission yeast. The red points represent the unstable steady state of MPF, and the black points represent the stable steady state of MPF. The blue bar indicates the range of the size checkpoint. (b) Stochastic numerical simulation in *wee1*Δ and WT. The x-axis and y-axis represent the initial size and the duration of the cell cycle of the fission yeast. The blue bars represent the range of the cycle time for a given birth size in *wee1*Δ and in WT fission yeast.

consider the direct spatial regulation provided by Pom1. In our model, we take this regulation into account. Vilela and colleagues built a mathematical model incorporating the Pom1 pathway [16]. However, they paid more attention on the formation of the Pom1 gradient and overpassed the link between bistability and size checkpoint. In our model, we specified that the critical size S2 (corresponding to saddle node of the lower branch SN2) is the cell size checkpoint and focused on the robustness of the size checkpoint.

Since the function of the Pom1 pathway has not been understood until 2009, our previous work related the size checkpoint to the cytoplasmic-to-nuclear size ratio. In the present work, we revealed that the underlying mechanism of size checkpoint is the saddle node bifurcation.

Bifurcation analysis of *wee1*Δ (Figure 5(a)) showed that although the range of the cell size checkpoint is larger than that of WT, it is still narrower than that of systems without spatial regulation (Figure 4(a)). This is because the spatial regulation is still assumed to work in *wee1*Δ ($k5$ remains 7.5% of WT, not 0 in *wee1*Δ). Even this weak spatial regulation can enhance the robustness of size checkpoint significantly. Therefore, the direct spatial regulation provided by Pom1 is thought to play a more important role in coupling cell growth to cell division.

Cell size checkpoints are present in many kinds of cells. A robust cell size checkpoint is required for the maintenance of appropriate cell size during proliferation. Although only spatial regulation was reflected in the present model of fission yeast, other cells, such as frog eggs, may also have similar ways of transferring spatial information directly, but this has not been experimentally established. Unlike that of fission yeast, the spatial regulation of oocytes takes place in a spherical space [12].

In the present study, a mathematical model was used to investigate the manner in which cell size can be coupled to the cell division in fission yeast. As the cells grow, the rate of synthesis of Cdc13 increases. However, the relief gradient offered by Pom1 can reduce the concentration of Cdr, which reduces the ability of Cdr to inhibit Wee1. The novel dynamics shown in the present model can be used to evaluate the direct spatial regulation provided by Pom1 and to examine its impact on cell checkpoints. The positive feedback loops were found to depend on spatial regulation and to generate a switch-like MPF response, which naturally produces the endogenetic size checkpoint. This direct spatial relation was found to protect the size checkpoint from fluctuations in gene expression.

4. Methods and Materials

The mathematical models of the cell cycle have been extensively studied. We adapted the parameters from the classic models of Ferrell's and Tyson's group. Besides, we also added the effect of the Pom1 pathway on Wee1 regulation. In other words, we introduced the spatial regulation to the system. The ordinary differential equations for this mathematical model are as follows:

$$V = V_0 * e^{\mu t} \tag{1}$$

$$\frac{d\mathrm{Cdc13}\,(t)}{dt} = k1 * V - k3 * \left(\mathrm{Cdc2_{tot}} - \mathrm{Cc}\,(t) - \mathrm{Ccp}\,(t) - \mathrm{Cptp}\,(t) - \mathrm{MPF}\,(t)\right) * \mathrm{Cdc13}\,(t) + k4 * \mathrm{CC}\,(t) - k2 * \left(\mathrm{Apc}\,(t) + \mathrm{Apc\_basal}\right) * \mathrm{Cdc13}\,(t), \tag{2}$$

$$\frac{d\mathrm{Cc}(t)}{dt} = k3 * \left(\mathrm{Cdc2}_{\mathrm{tot}} - \mathrm{Cc}(t) - \mathrm{Ccp}(t) - \mathrm{Cptp}(t) - \mathrm{MPF}(t)\right) * \mathrm{Cdc13}(t) - k4 * \mathrm{CC}(t) + k8 * \mathrm{Ccp}(t) * \mathrm{Cdc25}(t) + k9 * \mathrm{Ccp}(t) * (k10 - \mathrm{Cdc25}(t)) - k5 * \mathrm{Wee1}(t)\,\mathrm{Cc}(t) - k6 * (k7 - \mathrm{Wee1}(t)) * \mathrm{Cc}(t), \quad (3)$$

$$\frac{d\mathrm{Ccp}(t)}{dt} = k5 * \mathrm{Wee1}(t)\,\mathrm{Cc}(t) + k6 * (k7 - \mathrm{Wee1}(t)) * \mathrm{Cc}(t) - k8 * \mathrm{Ccp}(t) * \mathrm{Cdc25}(t) - k9 * \mathrm{Ccp}(t) * (k10 - \mathrm{Cdc25}(t)) + k12 * y4 - k11 * y3 - k2 * \left(y9 + \mathrm{Apc\_basal}\right) * y3, \quad (4)$$

$$\frac{d\mathrm{Ccpt}(t)}{dt} = k11 * \mathrm{Ccp}(t) - k12 * \mathrm{Ccpt}(t) - k8 * \mathrm{Cdc25}(t) * \mathrm{Ccptp} - k9 * (k10 - \mathrm{Cdc25}(t)) * \mathrm{Ccptp}(t) + k5 * \mathrm{Wee1}(t) * \mathrm{MPF}(t) + k6 * (k7 - \mathrm{Wee1}(t)) * y5 - k2 * \left(\mathrm{Apc}(t) + \mathrm{Apc\_basal}\right) * \mathrm{Ccptp}(t), \quad (5)$$

$$\frac{d\mathrm{MPF}(t)}{dt} = k8 * \mathrm{Cdc25}(t) * \mathrm{Ccpt}(t) + k9 * (k10 - \mathrm{Cdc25}(t)) * \mathrm{Ccpt}(t) - k5 * \mathrm{Wee1}(t) * \mathrm{MPF}(t) - k6 * (k7 - \mathrm{Wee1}(t)) * \mathrm{MPF}(t) - k2 * \left(\mathrm{APC}(t) + \mathrm{Apc\_basal}\right) * \mathrm{MPF}, \quad (6)$$

$$\frac{d\mathrm{Cdc25}(t)}{dt} = \frac{k13 * \mathrm{MPF}(t)^{k14}}{k27^{k14} + \mathrm{MPF}(t)^{k14}} * (k10 - \mathrm{Cdc25}(t)) - k15 * \mathrm{Cdc25}(t), \quad (7)$$

$$\mathrm{rel} = V * 7, \quad (8)$$

$$kd1 = \frac{\exp(-A * (\mathrm{rel}/2))}{\exp(A * (\mathrm{rel}/2)) + \exp(-A * (\mathrm{rel}/2))},$$
$$kd2 = \frac{\exp(A * (\mathrm{rel}/2))}{\exp(A * (\mathrm{rel}/2)) + \exp(-A * (\mathrm{rel}/2))},$$
$$\mathrm{pom1} = kd1 * \exp\left(A * \frac{\mathrm{rel}}{2}\right) + kd2 * \exp\left(-A * \frac{\mathrm{rel}}{2}\right), \quad (9)$$

$$\mathrm{Cdr} = \frac{k\mathrm{cdr\ on}}{k\mathrm{cdr\ on} + k\mathrm{cdr\ off} * \mathrm{pom1}^n / \left(\mathrm{pom1}^n + K^n\right)} \quad (10)$$

$$\frac{d\mathrm{Wee1}(t)}{dt} = -k16 * \frac{(\mathrm{MPF}(t) + \mathrm{cdr})^{k17}}{k28^{k17} + (\mathrm{MPF}(t) + \mathrm{cdr})^{k17}} * \mathrm{Wee1}(t) + k18 * (k7 - \mathrm{Wee1}(t)), \quad (11)$$

$$\frac{d\mathrm{Plo1}(t)}{dt} = k19 * \frac{\mathrm{MPF}(t)^{k20}}{k30^{k20} + \mathrm{MPF}(t)^{k20}} * (k21 - \mathrm{Plo1}(t)) - k22 * \mathrm{Plo1}(t), \quad (12)$$

$$\frac{d\mathrm{Apc}(t)}{dt} = k23 * \frac{\mathrm{Plo1}(t)^{k24}}{k30^{k24} + \mathrm{Plo1}(t)^{k24}} * \left(k25 - \mathrm{Apc}(t)\right) - k26 * \mathrm{Apc}(t). \quad (13)$$

In our model, the change rate of Pom1 at location x is given by a kinetic equation, where the first term is the rate of diffusion and the second term is its rate of degradation.

$$\frac{\partial \mathrm{Pom1}(x,t)}{\partial t} = m * \frac{\partial^2 \mathrm{Pom1}(x,t)}{\partial x^2} - k * \mathrm{Pom1}(x,t). \quad (14)$$

Here m represents the diffusion coefficient of Pom1 in fission yeast. And k represents the degradation coefficient of Pom1.

To make the system simpler, we supposed that the gradient of Pom1 can be formed rapidly. Therefore the concentration of Pom1 at x is set in quasi-steady state as follows:

$$m * \frac{d^2 \mathrm{Pom1}(x)}{dx^2} = k * \mathrm{Pom1}(x),$$
$$\mathrm{Pom1}(x) = kd1 * e^{x*A} + kd2 * e^{-x*A}, \quad (15)$$

where $A = \sqrt{k/m}$

$$kd1 = \frac{\exp(-A * (\mathrm{rel}/2))}{\exp(A * (\mathrm{rel}/2)) + \exp(-A * (\mathrm{rel}/2))},$$
$$kd2 = \frac{\exp(A * (\mathrm{rel}/2))}{\exp(A * (\mathrm{rel}/2)) + \exp(-A * (\mathrm{rel}/2))}. \quad (16)$$

Here rel represents the amplified size of the fission yeast. These parameters are estimated from the model of Vilela and colleagues [16]. It is notable that the birth size of the fission yeast is normalized to 1 in our model. However, the birth size of fission yeast is 7 μm in Vilela's model. Therefore, we amplified the cell size to 7-fold when applying the parameters of Vilela's model:

$$\mathrm{rel} = V * 7. \quad (17)$$

The activity of Cdr is repressed by the Pom1 at the central zone of the fission yeast. Thus we only need to consider the concentration of Pom1 at the center of the fission yeast:

$$\text{pom1} = kd1 * \exp\left(A * \frac{\text{rel}}{2}\right) + kd2 * \exp\left(-A * \frac{\text{rel}}{2}\right). \tag{18}$$

The change rate of Cdr is formulated as follows:

$$\frac{d\text{Cdr}(t)}{dt} = -k\text{cdr off} * \frac{\text{Pom1}^n}{K^n + \text{Pom1}^n} * \text{Cdr}(t) + k\text{cdr on} * (1 - \text{Cdr}(t)). \tag{19}$$

Similarly, to simplify the model, we assumed that the concentration of Cdr is set in steady state. Thus the ordinary equation can be transformed into an algebraic equation:

$$\text{Cdr} = \frac{k\text{cdr on}}{k\text{cdr on} + k\text{cdr off} * \text{pom1}^n / (\text{pom1}^n + K^n)}. \tag{20}$$

The parameters are as follows:

$$\begin{gathered} k1 = 0.01056; \quad k2 = 1; \quad k3 = 10; \quad k4 = 0.1; \\ k5 = 2; \quad k6 = 0.1; \quad k8 = 2; \quad k9 = 0.05; \\ k7 = 1; \quad k10 = 1; \quad k11 = 0.4; \quad k12 = 0.002; \\ k13 = 10; \quad k14 = 4; \quad k15 = 1; \quad k16 = 10; \\ k17 = 4; \quad k18 = 1; \quad k24 = 4; \quad k20 = 4; \\ k21 = 1; \quad k25 = 1; \quad k22 = 0.2; \quad k23 = 10; \\ k26 = 0.2; \quad k19 = 10; \quad k27 = 0.8; \quad k28 = 0.8; \\ k29 = 1.3; \quad k30 = 1.3; \quad \text{cdk tot} = 20; \\ \text{apc\_basal} = 0.01; \quad k\text{cdr on} = 5; \quad k\text{cdr off} = 497; \\ K = 0.5; \quad A = \sqrt{0.12}; \quad n = 9. \end{gathered} \tag{21}$$

Acknowledgments

This work was funded by the National Science Foundation of China (Grant no. 61271358).

References

[1] P. Fantes and P. Nurse, "Control of cell size at division in fission yeast by a growth modulated size control over nuclear division," *Experimental Cell Research*, vol. 107, no. 2, pp. 377–386, 1977.

[2] A. Sveiczer, B. Novak, and J. M. Mitchison, "The size control of fission yeast revisited," *Journal of Cell Science*, vol. 109, part 12, pp. 2947–2957, 1996.

[3] I. Rupeš, B. A. Webb, A. Mak, and P. G. Young, "G2/M arrest caused by actin disruption is a manifestation of the cell size checkpoint in fission yeast," *Molecular Biology of the Cell*, vol. 12, no. 12, pp. 3892–3903, 2001.

[4] Y. Masui and P. Wang, "Cell cycle transition in early embryonic development of *Xenopus laevis*," *Biology of the Cell*, vol. 90, no. 8, pp. 537–548, 1998.

[5] P. Wang, S. Hayden, and Y. Masui, "Transition of the blastomere cell cycle from cell size-independent to size-dependent control at the midblastula stage in *Xenopus laevis*," *Journal of Experimental Zoology*, vol. 287, no. 2, pp. 128–144, 2000.

[6] B. A. Edgar and C. F. Lehner, "Developmental control of cell cycle regulators: a fly's perspective," *Science*, vol. 274, no. 5293, pp. 1646–1652, 1996.

[7] H. Dolznig, F. Grebien, T. Sauer, H. Beug, and E. W. Müllner, "Evidence for a size-sensing mechanism in animal cells," *Nature Cell Biology*, vol. 6, no. 9, pp. 899–905, 2004.

[8] T. Takahashi, P. G. Bhide, T. Goto, S. Miyama, and V. S. Caviness Jr., "Proliferative behavior of the murine cerebral wall in tissue culture: cell cycle kinetics and checkpoints," *Experimental Neurology*, vol. 156, no. 2, pp. 407–417, 1999.

[9] A. Sveiczer, J. J. Tyson, and B. Novak, "Modelling the fission yeast cell cycle," *Briefings in Functional Genomics and Proteomics*, vol. 2, no. 4, pp. 298–307, 2004.

[10] A. Sveiczer, A. Csikasz-Nagy, B. Gyorffy, J. J. Tyson, and B. Novak, "Modeling the fission yeast cell cycle: quantized cycle times in wee1- cdc25Δ mutant cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 14, pp. 7865–7870, 2000.

[11] B. Novak, Z. Pataki, A. Ciliberto, and J. J. Tyson, "Mathematical model of the cell division cycle of fission yeast," *Chaos*, vol. 11, no. 1, pp. 277–286, 2001.

[12] Z. Qu, J. N. Weiss, and W. R. MacLellan, "Coordination of cell growth and cell division: a mathematical modeling study," *Journal of Cell Science*, vol. 117, no. 18, pp. 4199–4207, 2004.

[13] J. R. Pomerening, Y. K. Sun, and J. E. Ferrell Jr., "Systems-level dissection of the cell-cycle oscillator: bypassing positive feedback produces damped oscillations," *Cell*, vol. 122, no. 4, pp. 565–578, 2005.

[14] S. G. Martin and M. Berthelot-Grosjean, "Polar gradients of the DYRK-family kinase Pom1 couple cell length with the cell cycle," *Nature*, vol. 459, no. 7248, pp. 852–856, 2009.

[15] J. B. Moseley, A. Mayeux, A. Paoletti, and P. Nurse, "A spatial gradient coordinates cell size and mitotic entry in fission yeast," *Nature*, vol. 459, no. 7248, pp. 857–860, 2009.

[16] M. Vilela, J. J. Morgan, and P. A. Lindahl, "Mathematical model of a cell size checkpoint," *PLoS Computational Biology*, vol. 6, no. 12, Article ID e1001036, 2010.

[17] T. Y. Tsai, S. C. Yoon, W. Ma, J. R. Pomerening, C. Tang, and J. E. Ferrell Jr., "Robust, tunable biological oscillations from interlinked positive and negative feedback loops," *Science*, vol. 321, no. 5885, pp. 126–139, 2008.

[18] B. Novak and J. J. Tyson, "Numerical analysis of a comprehensive model of M-phase control in *Xenopus oocyte* extracts and intact embryos," *Journal of Cell Science*, vol. 106, part 4, pp. 1153–1168, 1993.

[19] J. E. Ferrell Jr., "Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability," *Current Opinion in Cell Biology*, vol. 14, no. 2, pp. 140–148, 2002.

[20] F. R. Cross, V. Archambault, M. Miller, and M. Klovstad, "Testing a mathematical model of the yeast cell cycle," *Molecular Biology of the Cell*, vol. 13, no. 1, pp. 52–70, 2002.

[21] P. A. Fantes, "Control of cell size and cycle time in *Schizosaccharomyces pombe*," *Journal of Cell Science*, vol. 24, pp. 51–67, 1977.

[22] P. Nurse and P. Thuriaux, "Controls over the timing of DNA replication during the cell cycle of fission yeast," *Experimental Cell Research*, vol. 107, no. 2, pp. 365–375, 1977.

[23] J. Lei, "Stochasticity in single gene expression with both intrinsic noise and fluctuation in kinetic parameters," *Journal of Theoretical Biology*, vol. 256, no. 4, pp. 485–492, 2009.

[24] L. Yang, Z. Han, W. R. MacLellan, J. N. Weiss, and Z. Qu, "Linking cell division to cell growth in a spatiotemporal model of the cell cycle," *Journal of Theoretical Biology*, vol. 241, no. 1, pp. 120–133, 2006.

[25] B. Novak, A. Csikasz-Nagy, B. Gyorffy, K. Chen, and J. J. Tyson, "Mathematical model of the fission yeast cell cycle with checkpoint controls at the G1/S, G2/M and metaphase/anaphase transitions," *Biophysical Chemistry*, vol. 72, no. 1-2, pp. 185–200, 1998.

[26] A. Sveiczer, J. J. Tyson, and B. Novak, "A stochastic, molecular model of the fission yeast cell cycle: role of the nucleocytoplasmic ratio in cycle time regulation," *Biophysical Chemistry*, vol. 92, no. 1-2, pp. 1–15, 2001.

[27] Z. Qu, W. R. MacLellan, and J. N. Weiss, "Dynamics of the cell cycle: checkpoints, sizers, and timers," *Biophysical Journal*, vol. 85, no. 6, pp. 3600–3611, 2003.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 628536, 7 pages
http://dx.doi.org/10.1155/2013/628536

Research Article

Molecular Dynamics Studies on the Conformational Transitions of Adenylate Kinase: A Computational Evidence for the Conformational Selection Mechanism

Jie Ping,[1] Pei Hao,[1,2] Yi-Xue Li,[2,3] and Jing-Fang Wang[2,4]

[1] *Pathogen Diagnostic Center, Institut Pasteur of Shanghai Chinese Academy of Sciences, Shanghai 200025, China*
[2] *Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai 200235, China*
[3] *Bioinformatics Center, Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[4] *Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai 200240, China*

Correspondence should be addressed to Jing-Fang Wang; jfwang8113@gmail.com

Received 5 April 2013; Accepted 13 June 2013

Academic Editor: Yudong Cai

Escherichia coli adenylate kinase (ADK) is a monomeric phosphotransferase enzyme that catalyzes reversible transfer of phosphoryl group from ATP to AMP with a large-scale domain motion. The detailed mechanism for this conformational transition remains unknown. In the current study, we performed long time-scale molecular dynamics simulations on both open and closed states of ADK. Based on the structural analyses of the simulation trajectories, we detected over 20 times conformational transitions between the open and closed states of ADK and identified two novel conformations as intermediate states in the catalytic processes. With these findings, we proposed a possible mechanism for the large-scale domain motion of *Escherichia coli* ADK and its catalytic process: (1) the substrate free ADK adopted an open conformation; (2) ATP bound with LID domain closure; (3) AMP bound with NMP domain closure; (4) phosphoryl transfer occurred with ATP, and AMP converted into two ADPs, and no conformational transition was detected in the enzyme; (5) LID domain opened with one ADP released; (6) another ADP released with NMP domain open. As both open and closed states sampled a wide range of conformation transitions, our simulation strongly supported the conformational selection mechanism for *Escherichia coli* ADK.

1. Introduction

Escherichia coli adenylate kinase (ADK) is a monomeric phosphotransferase enzyme regulating energy homeostasis in cells by catalyzing reversible transfer of a phosphoryl group from ATP to AMP. The structure of ADK is well studied in the past few years, and by now more than 20 crystal structures of ADK from *Escherichia coli* and other organisms in the absence and presence of substrates have been released in the protein structure databases. According to these crystal studies, the structure of ADK is mainly composed of three major components (Figure 1): a core domain (residues 1–29, 68–117, and 161–214), an AMP binding lid domain (also called NMP domain, residues 30–67), as well as an ATP binding lid domain (also called LID domain, residues 118–167). These basic components are shared by many other kinases and ATPases. As characterized by structural [1–4], biophysical [5, 6], and computational studies [7–10], ADK is believed to adopt an open conformation in the absence of substrates (Figure 1(a)), and with ATP or AMP binding the LID and NMP domains of this enzyme undergo large conformational transitions, leading to a closed conformation (Figure 1(b)). However, during the large conformational transitions, the core domain is comparatively rigid with the structural motions mainly located on the LID and NMP domains. This large conformational transition, especially opening of the

nucleotide binding lids required for the catalysis and product release, occur on the microsecond-millisecond times scale [11].

Large-scale conformational alterations are thought to mediate allosteric regulation, which are related to the protein function in signal transduction, immune response, and enzymatic activity. Thus, a fundamental problem is to understand the mechanism for the large-scale conformational transitions. To answer this question, two models are proposed: the induced-fit model and the population-shift model. In the induced-fit model, substrate binding is believed to induce conformational alterations in the active site to cause a new conformation for the entire enzyme. Contrarily, in the population-shift model, the enzyme is thought to adopt a conformational equilibrium among many native states, and substrates are able to selectively bind to a suitable native conformation, shifting the equilibrium toward the binding conformation.

Several computational works have been done in recent years to study the conformational transitions of ADK. All atomic molecular dynamics simulations have been successfully used to reveal the structural trajectories of ADK during the large conformational transitions [7–9, 12–21]. However, these studies still remain on the nanosecond times scale much smaller than the times scale on which the large conformational transitions of ADK occur. Additionally, coarse-grained simulations [22–26], normal mode analysis [27–29], and plastic/elastic network approaches [30–33] are also employed to generate transition pathways of ADK. However, these theoretical studies are controversial for adopting the simple harmonic potential approximation instead of all-atom models, which is less convincing due to the intermediate structures far away from the native states. To this end, we performed long time-scale molecular dynamics simulations on both open and closed state of ADK to study the large conformational transitions.

2. Materials and Methods

2.1. Initial Structures. Owing to playing an important role in cellular energy homeostasis, many good attempts have been made to study the three-dimensional (3D) structure of ADK. By now, 29 crystal structures of ADK and its homolog, in the absence and presence of substrates, are available in the protein structure databases [34–45]. In this study, the crystal structures (PDB ID: 4AKE [2] and 1AKE [46]) were selected as the initial structures for the further molecular dynamics simulations. The former was released in 1996 with 2.2 Å resolutions and treated as an open state of ADK in the current case. The latter was obtained in the presence of Ap_5A with 1.9 Å resolutions and considered as a closed state of ADK in this study. Except for the polar hydrogen and heavy atoms of the enzyme and Ap_5A, all other atoms including nonpolar hydrogen in both crystal structures were removed. The pKa values for each residue in ADK were calculated by Delphi [47, 48] as a Poisson-Boltzmann solver with a dielectric constant of 4. Hydrogen atoms were subsequently added to the enzyme with t-Leap procedure of Amber 11 package [49] based on the computational pKa values in the last step to give a total charge of −4. Then, the enzyme (together with Ap_5A in the closed structure) was solvated in a simulation box with explicit TIP3P water molecules. To neutralize the system, 4 sodium ions were added to random place 4 water molecules in the simulation box. The atoms of the enzyme were parameterized by Amber force field parameters [50], while Ap_5A was done by Antechamber module [51] in Amber 11 package.

2.2. Molecular Dynamics Simulations. Two simulation systems were involved in this study, with the crystal structures (PDB ID: 4AKE and 1AKE) as initial structures for the open and closed conformations, respectively. Both systems were subjected to the steepest descent energy minimization (~ 5000 steps) followed by conjugate gradient energy minimization for the next 5000 steps and subsequently equilibrated with the enzyme atoms (or ligand in the closed system) fixed for a short-time molecular dynamics simulations at 300 K to reduce the van der Waals conflicts. Finally, 10-nanosecond (ns) molecular dynamics simulations were performed for both systems under a constant temperature (300 K) by Amber 11 package [49] with periodic boundary conditions and NPT ensemble. Tweenty five frames were randomly selected from the first 10 ns simulation trajectories for both systems, and 10 independent molecular dynamics simulations (10 ns) were launched for each frame. For the simulations that were detected to have large scale conformational transitions, 200 ns molecular dynamics simulations were added using a GPU-accelerated approach. SHAKE algorithm with a tolerance of 10^{-6} was applied to constrain all bonds in both simulation systems [52, 53], and atom velocities for start-up runs were obtained according to the Maxwell distribution at 300 K [54, 55]. The isothermal compressibility was set to 4.5×10^{-5}/bar for solvent simulations [56, 57]. The electrostatic interactions were treated by the particle mesh Ewald (PME) algorithm with interpolation order of 4 and a grid spacing of 0.12 nanometers (nm) [58, 59]. The van der Waals interactions were calculated by using a cut-off of 12 Å. All the molecular dynamics simulations were performed with a time step of 2 femtoseconds (fs), and coordinates for both systems were saved every 1 picosecond (ps).

3. Results and Discussion

3.1. Conformational Transitions. As ADK has been well studied, a lot of kinetic and thermodynamic experimental data were released [60–64], based on which the large-scale conformational transitions for the NMP and LID domain opening are considered to be rate limiting for the catalytic reactions of the enzyme. However, it is really a challenging task to capture large-scale and long-time conformational transitions of proteins or enzymes for both experimental and computational methods. To check whether our simulations involved the large-scale conformational transitions of ADK, we calculated the RMS deviations of the Cα atoms from both open (4ake.pdb) and closed (1ake.pdb) conformations

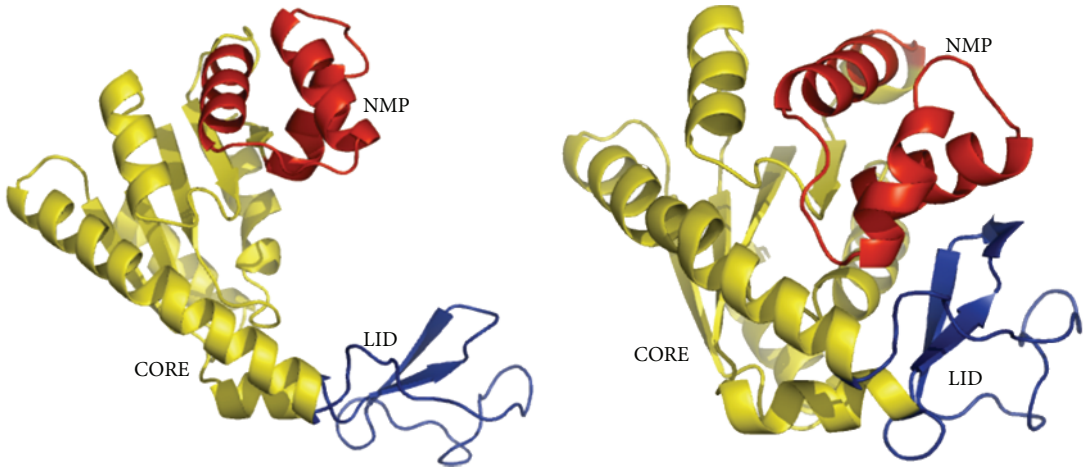


Figure 1: The three-dimensional structures of *Escherichia coli* adenylate kinase in the open (a) and closed (b) conformations. The crystal structures 4ake.pdb and 1ake.pdb are selected as the open and closed state of ADK, respectively. NMP (residues 30–67), CORE (residues 1–29, 68–117, and 161–214), and LID (residues 118–167) domains are colored in red, yellow, and blue, respectively.

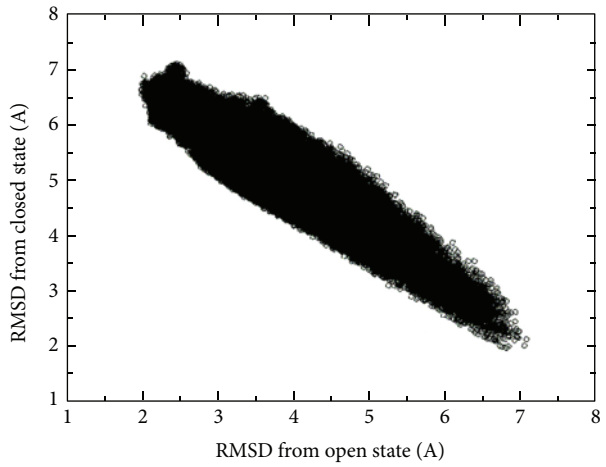


Figure 2: The RMS deviations of Cα atoms from both open and closed conformation along the molecular dynamics trajectories. The open (4ake.pdb) and closed (1ake.pdb) conformations of ADK were specially labeled. The ribbon RMS deviation distribution indicated that large-scale conformational transitions were detected during our molecular dynamics simulations.

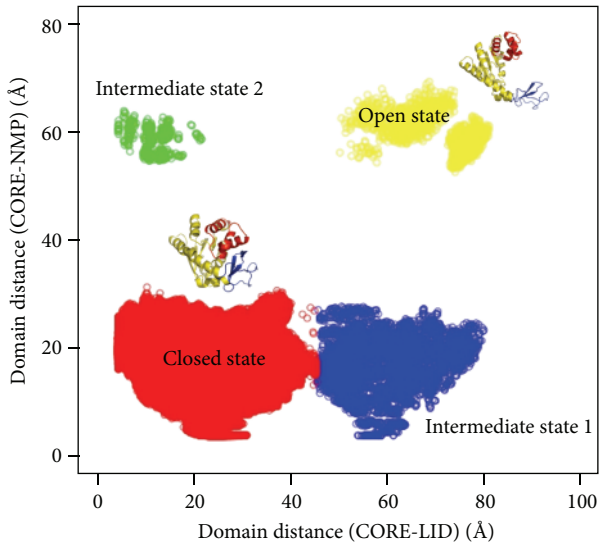


Figure 3: The mass center distances of NMP and LID domain with LID domain. The geometric center distances of NMP and CORE domains in the open and closed conformations were 62.7 and 18.4 Å, while those of LID and CORE domain for the open and closed conformations were 70.1 and 21.0 Å, respectively.

of ADK along all the molecular dynamics simulation trajectories. As shown in Figure 2, the RMS deviation employed a ribbon distribution, indicating that the large-scale conformational transitions occurred in our molecular dynamics simulations. If no conformational transitions occurred, the 2D-RMS deviations should be divided into two major parts. Additionally, the 2D-RMS deviations located in the middle of the ribbon distribution might stand for the intermediate states. As reported by Shapiro and Meirovitch [64], the rate for ADK domain motions was about 52 ns. While we employed a series of simulations with a time scale of 200 ns, almost 4 times larger than the measured rates for ADK domain motions, thus it was expected that the conformational transitions should be detected in our molecular dynamics simulations.

To confirm the conformational transitions for *Escherichia coli* ADK, we also measured the domain motion of ADK using the geometric center distances of NMP and LID domains with CORE domain along the molecular dynamics simulation trajectories. In the crystal structures, the geometric center distances of NMP and CORE domains in the open (4ake.pdb) and closed (1ake.pdb) conformations were 62.7 Å and 18.4 Å, while those for LID and CORE domain were 70.1 Å (open

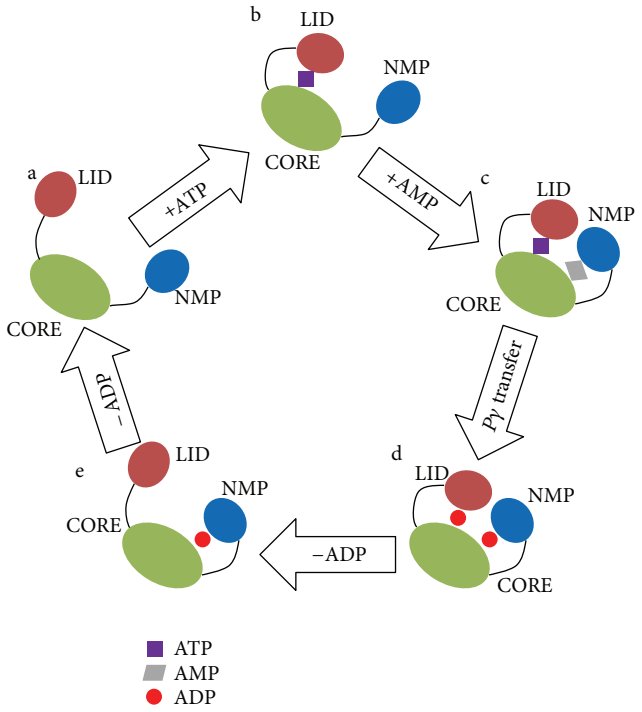


Figure 4: Conformational transition pathway and proposed catalytic mechanism of ADK. Model a, substrate free ADK with an open conformation. Model b, ATP bound form of ADK with a closed LID domain. Model c, ATP and AMP bound form of ADK with a closed conformation. Model d, two ADP bound forms of ADK with a closed conformation. Model e, one ADP bound form of ADK with a closed NMP domain.

state) and 21.0 Å (closed state), respectively. As shown in Figure 3, the domain distances were clustered into four major parts. The ones colored in yellow and red, respectively, represented the open and closed conformations defined by the crystal structures, whereas the ones colored in blue and green were identified as novel configurations from the open and closed states. The blue ones in Figure 3 employed an open LID domain with a closed NMP domain, while the green ones had an open NMP domain with a closed LID domain. We believed that the blue and green parts in Figure 3 might be the intermediate states of ADK domain motions.

3.2. Conformational Transition Pathway. The large-scale domain motions were detected in the catalytic conversion of Mg^{2+}–ATP + AMP → Mg^{2+}–ADP + ADP by *Escherichia coli* ADK [65, 66]. Thus, the possible conformational transition pathway was thought to be associated with the catalytic mechanism for *Escherichia coli* ADK. According to the previous theoretical studies, LID domain motion was believed to precede NMP domain motion [6, 26, 33, 66]. One possible reason was that there was a stable salt bridge D118-K136 between LID and CORE, which could connect both domains by the strong contributions to the total enthalpic interactions. This salt bridge, also detected in several adenylate kinase structures of different species, was thought to have stabilizing function for the open conformation [9]. Another possible reason was that ATP binding in LID domain was found to have ability to assist NMP domain motion [66]. Based on these points, we proposed a possible conformational transition pathway and catalytic mechanism for *Escherichia coli* ADK (Figure 4). The substrate free enzyme adopted an open state. After ATP binding, ADK would close its LID domain first to form a novel configuration (model b in Figure 4 and intermediate state 2 in Figure 3). The closure motion of LID domain allowed AMP to bind in NMP domain binding site, resulting in the NMP domain closure (model c in Figure 4). With ATP and AMP binding, the enzyme would adopt a complete closed conformation. After phosphoryl transfer, ATP and AMP were converted into two ADP molecules. In this step, no conformational transition occurred, and thus the enzyme remained in its closed state (model d in Figure 4). When an ADP molecule was released from the ATP binding site, LID domain altered its closed configuration into an open state (model e in Figure 4 and intermediate state 1 in Figure 3). Subsequently, NMP domain opened via the bending of $\alpha 2$ helix toward $\alpha 4$ helix of CORE domain by nearly 15 degrees with respect to $\alpha 3$ helix, and the enzyme would adopt a complete open conformation (model a in Figure 4). This opening process of NMP domain (also AMP binding cleft) was considered to be involved in facilitating an efficient release of the formed product after catalysis.

Our proposed mechanism was also supported by some experimental results. Firstly, the complete closed conformation of ADK was found to be the only product-forming state [24]. In other words, to achieve its biological functions, ADK had to convert the open state (substrate free state) into the closed state. Secondly, The ATP binding site in LID domain was able to accommodate ATP, ADP, as well as AMP. ATP employed the highest binding affinity, whereas AMP had the lowest binding affinity [26]. Thirdly, the AMP binding site in NMP domain could only accommodate ADP and AMP [26]. Finally, the AMP binding site could only accommodate AMP in the intermediate state 2 with a closed LID domain and an open NMP domain [18]. Along the aforementioned pathway, a secondary structure analysis of the enzyme structure was performed using DSSP package to check the structural stability during the large-scale conformational transitions. As expected, no significant alterations in secondary structures were detected in CORE and NMP domain during the large-scale conformational transitions. For LID domain, some residues were found to bend and turn with few alterations in secondary structures. Thus, it was believed that ADK maintained its integrate structure with minute changes in its secondary structures, indicating that this enzyme behaved as a rigid body with flexible domains, and pathway detected in our study was reasonable.

Furthermore, our proposed mechanism was also in good agreement with previous theoretical studies. In 2008, Lu and Wang developed a coarse grained two-well model to study the conformational dynamics of ADK in microscopic detail [19]. They identified the LID-closing and NMP-closing pathways, providing theoretical evidence to our model (especially the transitions between model a and b and between model b and c). In 2008, Kubitzki and de Groot used TEE-REX molecular dynamics to study the conformational transitions

of ADK and detected a pathway for the open-to-closed transitions [9]. Although a complete transition was not observed, the pathway they found was in accord with the transitions from model a to model c. In 2010, Prof. Head-Gordon and her co-workers used normal modes analyses to study the transition pathway [27]. Although a complete transition pathway was identified, they did not provide any explanation for the catalytic mechanism of ADK. Besides the aforementioned studies, many other works were involved in the free energy calculations during the conformational transitions [19–22], giving an indication that there were intermediate states during the conformational transitions.

3.3. Conformational Transition Mechanism. For the conformational transitions in proteins or enzymes, two different substrate binding mechanisms, induced-fit model, and conformational selection model can be proposed. The former is a model for enzyme-substrate interactions to describe that only the proper substrate is capable of inducing the conformational changes in the binding site, allowing the enzyme to perform its catalytic function [67]. The latter believes that the proteins or enzymes exist in the multiple conformations in the vicinity of its native state, and the ligand selectively binds to an active conformation to shift the equilibrium toward the binding conformation [68, 69]. In the current case, long time-scale molecular dynamics simulations were performed on *Escherichia coli* ADK, which can sample a large set of conformations between the open and closed states. Our computational results strongly supported the conformational selection model proposed for ADK [5, 70–74].

4. Conclusion

In conclusion, to study the large-scale domain motions in *Escherichia coli* ADK, we performed long time-scale molecular dynamics on both open and closed states of ADK. The two-dimensional RMS deviations of the Cα atoms from both open and closed conformations along simulation trajectories confirmed the fact that conformational transitions between the open and closed conformation occurred during our simulations. Additionally, two significant intermediate states were identified by monitoring the domain distances between LID/NMP and CORE domains, one of which adopted an open LID domain with a closed NMP domain, and the other one employed an open NMP domain with a closed LID domain. Based on these computational results, we proposed a possible mechanism for the large scale conformational transitions and the catalytic function (Figure 4). The proposed mechanism was in good accordance with the previous experimental and theoretical studies, providing strong support to the conformational selection mechanism for *Escherichia coli* ADK.

Acknowledgments

This work was supported by Grants from the National Basic Research Program of China (973 Program, no. 2012CB517900), National Key Basic Research Program (no. 30800210), National Natural Science Foundation of China (nos. 31200547 and 90913009), Shanghai Pujiang Scholarship Program (no. 10PJ1408000), Doctoral Program Foundation of Institutions of Higher Education of China (no. 20110073120078), and China Postdoctoral Science Foundation (no. 2012M520949). The authors gratefully acknowledge the support of SA-SIBS scholarship Program.

References

[1] C. Vonrhein, G. J. Schlauderer, and G. E. Schulz, "Movie of the structural changes during a catalytic cycle of nucleoside monophosphate kinases," *Structure*, vol. 3, no. 5, pp. 483–490, 1995.

[2] C. W. Müller, G. J. Schlauderer, J. Reinstein, and G. E. Schulz, "Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding," *Structure*, vol. 4, no. 2, pp. 147–156, 1996.

[3] G. J. Schlauderer, K. Proba, and G. E. Schulz, "Structure of a mutant adenylate kinase ligated with an ATP-analogue showing domain closure over ATP," *Journal of Molecular Biology*, vol. 256, no. 2, pp. 223–227, 1996.

[4] G. J. Schlauderer and G. E. Schulz, "The structure of bovine mitochondrial adenylate kinase: comparison with isoenzymes in other compartments," *Protein Science*, vol. 5, no. 3, pp. 434–441, 1996.

[5] K. A. Henzler-Wildman, V. Thai, M. Lei et al., "Intrinsic motions along an enzymatic reaction trajectory," *Nature*, vol. 450, no. 7171, pp. 838–844, 2007.

[6] J. A. Hanson, K. Duderstadt, L. P. Watkins et al., "Illuminating the mechanistic roles of enzyme conformational dynamics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 46, pp. 18055–18060, 2007.

[7] K. Arora and C. L. Brooks, "Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 47, pp. 18496–18501, 2007.

[8] C. Snow, G. Y. Qi, and S. Hayward, "Essential dynamics sampling study of adenylate kinase: comparison to citrate synthase and implication for the hinge and shear mechanisms of domain motions," *Proteins*, vol. 67, no. 2, pp. 325–337, 2007.

[9] M. B. Kubitzki and B. L. de Groot, "The atomistic mechanism of conformational transition in adenylate kinase: a TEE-REX molecular dynamics study," *Structure*, vol. 16, no. 8, pp. 1175–1182, 2008.

[10] K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. Kern, "A hierarchy of timescales in protein dynamics is linked to enzyme catalysis," *Nature*, vol. 450, no. 7171, pp. 913–916, 2007.

[11] M. Wolf-Watz, V. Thai, K. Henzler-Wildman, G. Hadjipavlou, E. Z. Eisenmesser, and D. Kern, "Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair," *Nature Structural and Molecular Biology*, vol. 11, no. 10, pp. 945–949, 2004.

[12] B. Jana, B. V. Adkar, R. Biswas, and B. Bagchi, "Dynamic coupling between the LID and NMP domain motions in the catalytic conversion of ATP and AMP to ADP by adenylate kinase," *Journal of Chemical Physics*, vol. 134, no. 3, Article ID 035101, 10 pages, 2011.

[13] J. B. Brokaw and J. W. Chu, "On the roles of substrate binding and hinge unfolding in conformational changes of adenylate kinase," *Biophysical Journal*, vol. 99, no. 10, pp. 3420–3429, 2010.

[14] F. Pontiggia, A. Zen, and C. Micheletti, "Small- and large-scale conformational changes of adenylate kinase: a molecular dynamics study of the subdomain motion and mechanics," *Biophysical Journal*, vol. 95, no. 12, pp. 5901–5912, 2008.

[15] H. Dong, S. Qin, and H. X. Zhou, "Effects of macromolecular crowding on protein conformational changes," *PLoS Computational Biology*, vol. 6, Article ID e1000833, 2010.

[16] H. Lou and R. I. Cukier, "Molecular dynamics of apo-adenylate kinase: a principal component analysis," *Journal of Physical Chemistry B*, vol. 110, no. 25, pp. 12796–12808, 2006.

[17] E. Bae and G. N. Phillips Jr., "Identifying and engineering ion pairs in adenylate kinases: Insights from molecular dynamics simulations of thermophilic and mesophilic homologues," *The Journal of Biological Chemistry*, vol. 280, no. 35, pp. 30943–30948, 2005.

[18] H. Krishnamurthy, H. Lou, A. Kimple, C. Vieille, and R. I. Cukier, "Associative mechanism for phosphoryl transfer: a molecular dynamics simulation of *Escherichia coli* adenylate kinase complexed with its substrates," *Proteins*, vol. 58, no. 1, pp. 88–100, 2005.

[19] Q. Lu and J. Wang, "Single molecule conformational dynamics of adenylate kinase: energy landscape, structural correlations, and transition state ensembles," *Journal of the American Chemical Society*, vol. 130, no. 14, pp. 4772–4783, 2008.

[20] O. Beckstein, E. J. Denning, J. R. Perilla, and T. B. Woolf, "Zipping and unzipping of adenylate kinase: atomistic insights into the ensemble of open ↔ closed transitions," *Journal of Molecular Biology*, vol. 394, no. 1, pp. 160–176, 2009.

[21] M. D. Daily, G. N. Phillips Jr., and Q. Cui, "Many local motions cooperate to produce the adenylate kinase conformational transition," *Journal of Molecular Biology*, vol. 400, no. 3, pp. 618–631, 2010.

[22] R. Potestio, F. Pontiggia, and C. Micheletti, "Coarse-grained description of protein internal dynamics: an optimal strategy for decomposing proteins in rigid subunits," *Biophysical Journal*, vol. 96, no. 12, pp. 4993–5002, 2009.

[23] R. D. Hills Jr., L. Lu, and G. A. Voth, "Multiscale coarse-graining of the protein energy landscape," *PLoS Computational Biology*, vol. 6, no. 6, Article ID e1000827, 2010.

[24] O. Miyashita, P. G. Wolynes, and J. N. Onuchic, "Simple energy landscape model for the kinetics of functional transitions in proteins," *Journal of Physical Chemistry B*, vol. 109, no. 5, pp. 1959–1969, 2005.

[25] Q. Lu and J. Wang, "Kinetics and statistical distributions of single-molecule conformational dynamics," *Journal of Physical Chemistry B*, vol. 113, no. 5, pp. 1517–1521, 2009.

[26] P. C. Whitford, O. Miyashita, Y. Levy, and J. N. Onuchic, "Conformational transitions of adenylate kinase: switching by cracking," *Journal of Molecular Biology*, vol. 366, no. 5, pp. 1661–1671, 2007.

[27] C. Peng, L. Zhang, and T. Head-Gordon, "Instantaneous normal modes as an unforced reaction coordinate for protein conformational transitions," *Biophysical Journal*, vol. 98, no. 10, pp. 2356–2364, 2010.

[28] A. Ahmed, F. Rippmann, G. Barnickel, and H. Gohlke, "A normal mode-based geometric simulation approach for exploring biologically relevant conformational transitions in proteins," *Journal of Chemical Information and Modeling*, vol. 51, no. 7, pp. 1604–1622, 2011.

[29] S. Kirillova, J. Cortés, A. Stefaniu, and T. Siméon, "An NMA-guided path planning approach for computing large-amplitude conformational changes in proteins," *Proteins*, vol. 70, no. 1, pp. 131–143, 2008.

[30] N. A. Temiz, E. Meirovitch, and I. Bahar, "*Escherichia coli* adenylate kinase dynamics: comparison of elastic network model modes with mode-coupling ^{15}N-NMR relaxation data," *Proteins*, vol. 57, no. 3, pp. 468–480, 2004.

[31] J. N. Stember and W. Wriggers, "Bend-twist-stretch model for coarse elastic network simulation of biomolecular motion," *Journal of Chemical Physics*, vol. 131, no. 7, Article ID 074112, 9 pages, 2009.

[32] W. J. Zheng, B. R. Brooks, and G. Hummer, "Protein conformational transitions explored by mixed elastic network models," *Proteins*, vol. 69, no. 1, pp. 43–57, 2007.

[33] P. Maragakis and M. Karplus, "Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase," *Journal of Molecular Biology*, vol. 352, no. 4, pp. 807–822, 2005.

[34] M. B. Berry and G. N. J. Philips, "Crystal structures of *Bacillus stearothermophilus* adenylate kinase with bound Ap_5A, Mg^{2+} Ap_5A, and Mn^{2+} Ap_5A reveal an intermediate lid position and six coordinate octahedral geometry for bound Mg^{2+} and Mn^{2+}," *Proteins*, vol. 32, no. 3, pp. 276–288, 1998.

[35] C. W. Muller and G. E. Schulz, "Crystal structures of two mutants of adenylate kinase from *Escherichia coli* that modify the Gly-loop," *Proteins*, vol. 15, no. 1, pp. 42–49, 1993.

[36] M. B. Berry, E. Bae, T. R. Bilderback, M. Glaser, and G. N. Phillips Jr., "Crystal structure of ADP/AMP complex of *Escherichia coli* adenylate kinase," *Proteins*, vol. 62, no. 2, pp. 555–556, 2006.

[37] M. B. Berry, B. Meador, T. Bilderback, P. Liang, M. Glaser, and G. N. Phillips Jr., "The closed conformation of a highly flexible protein: the structure of *E. coli* adenylate kinase with bound AMP and AMPPNP," *Proteins*, vol. 19, no. 3, pp. 183–198, 1994.

[38] E. Bae and G. N. Phillips Jr., "Structures and analysis of highly homologous psychrophilic, mesophilic, and thermophilic adenylate kinases," *The Journal of Biological Chemistry*, vol. 279, no. 27, pp. 28202–28208, 2004.

[39] R. Couñago, S. Chen, and Y. Shamoo, "In vivo molecular evolution reveals biophysical origins of organismal fitness," *Molecular Cell*, vol. 22, no. 4, pp. 441–449, 2006.

[40] U. Abele and G. E. Schulz, "High-resolution structures of adenylate kinase from yeast ligated with inhibitor Ap_5A, showing the pathway of phosphoryl transfer," *Protein Science*, vol. 4, no. 7, pp. 1262–1271, 1995.

[41] P. Spuergin, U. Abele, and G. E. Schulz, "Stability, activity and structure of adenylate kinase mutants," *European Journal of Biochemistry*, vol. 231, no. 2, pp. 405–413, 1995.

[42] H. Gu, H. F. Chen, D. Q. Wei, and J. F. Wang, "Molecular dynamics simulations exploring drug resistance in HIV-1 proteases," *Chinese Science Bulletin*, vol. 55, no. 24, pp. 2677–2683, 2010.

[43] K. Wild, R. Grafmüller, E. Wagner, and G. E. Schulz, "Structure, catalytsis and supramolecular assembly of adenylate kinase from maize," *European Journal of Biochemistry*, vol. 250, no. 2, pp. 326–331, 1997.

[44] G. J. Schlauderer and G. E. Schulz, "The structure of bovine mitochondrial adenylate kinase: comparison with isoenzymes in other compartments," *Protein Science*, vol. 5, no. 3, pp. 434–441, 1996.

[45] K. Diederichs and G. E. Schulz, "The refined structure of the complex between adenylate kinase from beef heart mitochondrial matrix and its substrate AMP at 1.85 Å resolution," *Journal of Molecular Biology*, vol. 217, no. 3, pp. 541–549, 1991.

[46] C. W. Müller and G. E. Schulz, "Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap_5A refined at 1.9 Å resolution. A model for a catalytic transition state," *Journal of Molecular Biology*, vol. 224, no. 1, pp. 159–177, 1992.

[47] R. E. Georgescu, E. G. Alexov, and M. R. Gunner, "Combining conformational flexibility and continuum electrostatics for calculating pKas in proteins," *Biophysical Journal*, vol. 83, no. 4, pp. 1731–1748, 2002.

[48] W. Rocchia, E. Alexov, and B. Honig, "Extending the applicability of the nonlinear Poisson-Boltzmann equation: multiple dielectric constants and multivalent ions," *Journal of Physical Chemistry B*, vol. 105, no. 28, pp. 6507–6514, 2001.

[49] D. A. Case, T. E. Cheatham III, T. Darden et al., "The Amber biomolecular simulation programs," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1668–1688, 2005.

[50] J. W. Ponder and D. A. Case, "Force fields for protein simulations," *Advances in Protein Chemistry*, vol. 66, pp. 27–85, 2003.

[51] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, "Antechamber, an accessory software package for molecular mechanical calculations," *Journal of Computational Chemistry*, vol. 25, pp. 1157–1174, 2005.

[52] J. F. Wang and K. C. Chou, "Insights from modeling the 3D structure of New Delhi metallo-β-lactamse and its binding interactions with antibiotic drugs," *PLoS ONE*, vol. 6, no. 4, Article ID e18414, 2011.

[53] P. Lian, D. Q. Wei, J. F. Wang, and K. C. Chou, "An allosteric mechanism inferred from molecular dynamics simulations on phospholamban pentamer in lipid membranes," *PLoS ONE*, vol. 6, no. 4, Article ID e18587, 2011.

[54] Y. Wang, D. Q. Wei, and J. F. Wang, "Molecular dynamics studies on T1 lipase: insight into a double-flap mechanism," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 875–878, 2010.

[55] J. F. Wang, P. Hao, Y. X. Li, J. L. Dai, and X. Li, "Exploration of conformational transition in the aryl-binding site of human FXa using molecular dynamics simulations," *Journal of Molecular Modeling*, vol. 18, no. 6, pp. 2717–2725, 2012.

[56] J. Li, D. Q. Wei, J. F. Wang, and Y. X. Li, "A negative cooperativity mechanism of human CYP2E1 inferred from molecular dynamics simulations and free energy calculations," *Journal of Chemical Information and Modeling*, vol. 51, no. 12, pp. 3217–3225, 2011.

[57] J. F. Wang and K. C. Chou, "Insights into the mutation-induced HHH syndrome from modeling human mitochondrial ornithine transporter-1," *PLoS ONE*, vol. 7, no. 1, Article ID e31048, 2012.

[58] J. Li, D. Q. Wei, J. F. Wang, Z. T. Yu, and K. C. Chou, "Molecular dynamics simulations of CYP2E1," *Medicinal Chemistry*, vol. 8, no. 2, pp. 208–221, 2012.

[59] J. He, D. Q. Wei, J. F. Wang, and K. C. Chou, "Predicting protein-ligand binding sites based on an improved geometric algorithm," *Protein and Peptide Letters*, vol. 18, no. 10, pp. 997–1001, 2011.

[60] T. Dahnke, Z. Shi, H. Yan, R. T. Jiang, and M. D. Tsai, "Mechanism of adenylate kinase. Structural and functional roles of the conserved arginine-97 and arginine-132," *Biochemistry*, vol. 31, no. 27, pp. 6318–6328, 1992.

[61] M. A. Sinev, E. V. Sineva, V. Ittah, and E. Haas, "Domain closure in adenylate kinase," *Biochemistry*, vol. 35, no. 20, pp. 6425–6437, 1996.

[62] Y. E. Shapiro, M. A. Sinev, E. V. Sineva, V. Tugarinov, and E. Meirovitch, "Backbone dynamics of *Escherichia coli* adenylate kinase at the extreme stages of the catalytic cycle studied by ^{15}N NMR relaxation," *Biochemistry*, vol. 39, no. 22, pp. 6634–6644, 2000.

[63] M. Wolf-Watz, V. Thai, K. Henzler-Wildman, G. Hadjipavlou, E. Z. Eisenmesser, and D. Kern, "Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair," *Nature Structural and Molecular Biology*, vol. 11, no. 10, pp. 945–949, 2004.

[64] Y. E. Shapiro and E. Meirovitch, "Activation energy of catalysis-related domain motion in *E. coli* adenylate kinase," *Journal of Physical Chemistry B*, vol. 110, no. 23, pp. 11519–11524, 2006.

[65] O. Miyashita, J. N. Onuchic, and P. G. Wolynes, "Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 22, pp. 12570–12575, 2003.

[66] P. C. Whitford, S. Gosavi, and J. N. Onuchic, "Conformational transitions in adenylate kinase: allosteric communication reduces misligation," *The Journal of Biological Chemistry*, vol. 283, no. 4, pp. 2042–2048, 2008.

[67] D. E. Koshland Jr., "Application of a theory of enzyme specificity to protein synthesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 44, no. 2, pp. 98–104, 1958.

[68] B. Ma, S. Kumar, C. Tsai, and R. Nussinov, "Folding funnels and binding mechanisms," *Protein Engineering*, vol. 12, no. 9, pp. 713–720, 1999.

[69] C. Tsai, S. Kumar, B. Ma, and R. Nussinov, "Folding funnels, binding funnels, and protein function," *Protein Science*, vol. 8, no. 6, pp. 1181–1190, 1999.

[70] H. J. Zhang, X. R. Sheng, W. D. Niu, X. M. Pan, and J. M. Zhou, "Evidence for at least two native forms of rabbit muscle adenylate kinase in equilibrium in aqueous solution," *The Journal of Biological Chemistry*, vol. 273, no. 13, pp. 7448–7456, 1998.

[71] Y. Han, X. Li, and X. M. Pan, "Native states of adenylate kinase are two active sub-ensembles," *FEBS Letters*, vol. 528, no. 1–3, pp. 161–165, 2002.

[72] E. Z. Eisenmesser, O. Millet, W. Labeikovsky et al., "Intrinsic dynamics of an enzyme underlies catalysis," *Nature*, vol. 438, no. 7064, pp. 117–121, 2005.

[73] D. Tobi and I. Bahar, "Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 52, pp. 18908–18913, 2005.

[74] P. C. Whitford, O. Miyashita, Y. Levy, and J. N. Onuchic, "Conformational transitions of adenylate kinase: switching by cracking," *Journal of Molecular Biology*, vol. 366, no. 5, pp. 1661–1671, 2007.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 625403, 8 pages
http://dx.doi.org/10.1155/2013/625403

Research Article

Application of Improved Three-Dimensional Kernel Approach to Prediction of Protein Structural Class

Xu Liu,[1] Yuchao Zhang,[2,3] Hua Yang,[1] Lisheng Wang,[1] and Shuaibing Liu[1,4]

[1] *School of Chemistry & Chemical Engineering, Guangxi University, Guangxi Province, Nanning 530004, China*
[2] *State Key Laboratory of Medical Genomics, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai Jiao Tong University School of Medicine, Shanghai 200240, China*
[3] *Graduate School of the Chinese Academy of Sciences, Beijing 100049, China*
[4] *College of Pharmacy, Guangxi University of Chinese Medicine, Nanning 530001, China*

Correspondence should be addressed to Xu Liu; wendaoliuxu@hotmail.com

Received 25 March 2013; Revised 4 May 2013; Accepted 10 May 2013

Academic Editor: Bing Niu

Kernel methods, such as kernel PCA, kernel PLS, and support vector machines, are widely known machine learning techniques in biology, medicine, chemistry, and material science. Based on nonlinear mapping and Coulomb function, two 3D kernel approaches were improved and applied to predictions of the four protein tertiary structural classes of domains (all-α, all-β, α/β, and $\alpha+\beta$) and five membrane protein types with satisfactory results. In a benchmark test, the performances of improved 3D kernel approach were compared with those of neural networks, support vector machines, and ensemble algorithm. Demonstration through leave-one-out cross-validation on working datasets constructed by investigators indicated that new kernel approaches outperformed other predictors. It has not escaped our notice that 3D kernel approaches may hold a high potential for improving the quality in predicting the other protein features as well. Or at the very least, it will play a complementary role to many of the existing algorithms in this regard.

1. Introduction

Due to the rapid development of genome and protein science, the biological information has expanded dramatically. Therefore, it is very important and highly desirable for computers to manage, organize, and interpret the information. As a part of biochemistry, study of protein structure classes has become a hot topic, because of experimental and theoretical purposes. Artificial neural networks, support vector machines, kernel methods, and ensemble algorithms are widely known machine learning techniques in biology, medicine, chemistry, and material science [1–10]. In this work, two classification problems, protein's tertiary structure classes of domains and membrane protein types, were researched with some machine learning techniques.

Several motifs pack together to form compact, local, and semiindependent units called domains. The details of proteins domains structures are extremely complicated and irregular. But their overall structural frames are simple, regular, and truly elegant [11–13]. Many protein domains often have similar or identical folding patterns even if they are quite different according to their sequences [14–16]. The overall 3D structure of the polypeptide chain is referred to as the protein's tertiary structure. Levitt and Chothia proposed to classify protein tertiary structures into the following four structural classes based on the secondary structural content of the domains. (1) All-α: it is formed essentially by α-helices. This class is dominated by small folds, many of which form a simple bundle with helices running up and down. (2) All-β: this class has a core composed of antiparallel β-sheets, usually two sheets pack against each other. (3) α/β: this class contains both α-helices and β-strands that are largely interspersed in forming mainly parallel β-sheet; (4) $\alpha+\beta$: this class also contains both of the two secondary structure elements that, however, are largely segregated in forming mainly antiparallel β-sheets.

This concept of structural class has ever since been widely used as an important attribute for characterizing the overall folding type of proteins domains. Lots of methods have been

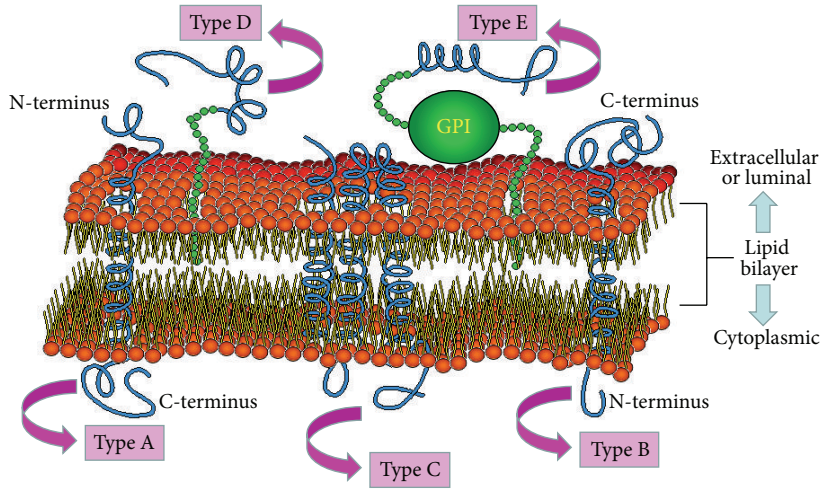


Figure 1: Five types of membrane proteins.

made to predict the structural classes based on the knowledge of protein sequences [17].

The research of membrane protein type is also important because of the special biological functions. The biomembrane usually contains some specific proteins and lipid components that enable it to perform its unique roles in the cell and organelle.

Furthermore, several studies show that many membrane proteins are also the key targets of drug discovery, particularly membrane channel proteins [18–20]. Membrane proteins can be further classified into the five types [21–23]: (a) type A membrane protein is single-pass transmembrane protein which has an extracellular (or luminal) N-terminus and cytoplasmic C-terminus for a cell (or organelle) membrane; (b) type B membrane protein is single-pass transmembrane protein which has an extracellular (or luminal) C-terminus and cytoplasmic N-terminus for a cell (or organelle) membrane; (c) type C is multipass transmembrane protein: the polypeptide crosses the lipid bilayer multiple times; (d) type D membrane proteins are lipid chain-anchored membrane proteins: they are bound to the membrane by one or more covalently attached fatty acid chains or other types of lipid chains called prenyl groups; (e) type E is GPI-anchored membrane protein which is bound to the membrane by a glycosylphosphatidylinositol (GPI) anchor.

Researchers have applied classification algorithm to predict the types of membrane proteins based on their amino acid composition [24, 25]. Figure 1 shows the forms and the locations of different membrane proteins.

The first goal of this paper is to illustrate the application of 3D kernel approach as a relatively new tool in proteins domains field for classification purposes. And the second goal is to show that the new approach can be applied to analysis of membrane protein types.

2. Materials and Methods

2.1. Kernel Function. Kernel function was originally a kind of functions used in integral operator research. However, Vapnik implemented this function in his newly invented SVMs method [26]. The use of kernel function makes SVMs able to treat nonlinear data processing problems by using linear algorithms. The basic idea of kernel function is to map the data $\mathbf{X}$ into a higher-dimensional feature space $\mathbf{F}$ via a nonlinear mapping Φ and then to do classification and regression in this space. There are four commonly used kernel functions:

linear kernel

$$K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x} \cdot \mathbf{y} \rangle + \theta. \tag{1}$$

polynomial kernel

$$K(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x} \cdot \mathbf{y} \rangle + \theta)^d. \tag{2}$$

Gaussian (RBF) kernel

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{y}\|^2}{\sigma^2}\right). \tag{3}$$

sigmoid kernel

$$K(\mathbf{x}, \mathbf{y}) = \tanh(\nu \langle \mathbf{x} \cdot \mathbf{y} \rangle + r). \tag{4}$$

The elegance of using kernel function lied in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(\mathbf{x})$. Any function that satisfies Mercer's condition can be used as kernel function.

2.2. Kernel PCA. Principal component analysis (PCA) is a versatile and easy-to-use multivariate mathematical-statistical method in multivariate data analysis and the extraction of maximal information [27, 28]. It is a linear transformation approach that compresses high-dimensional data with minimum loss of data information. PCA is performed in the original sample space, whereas kernel PCA (KPCA) applies

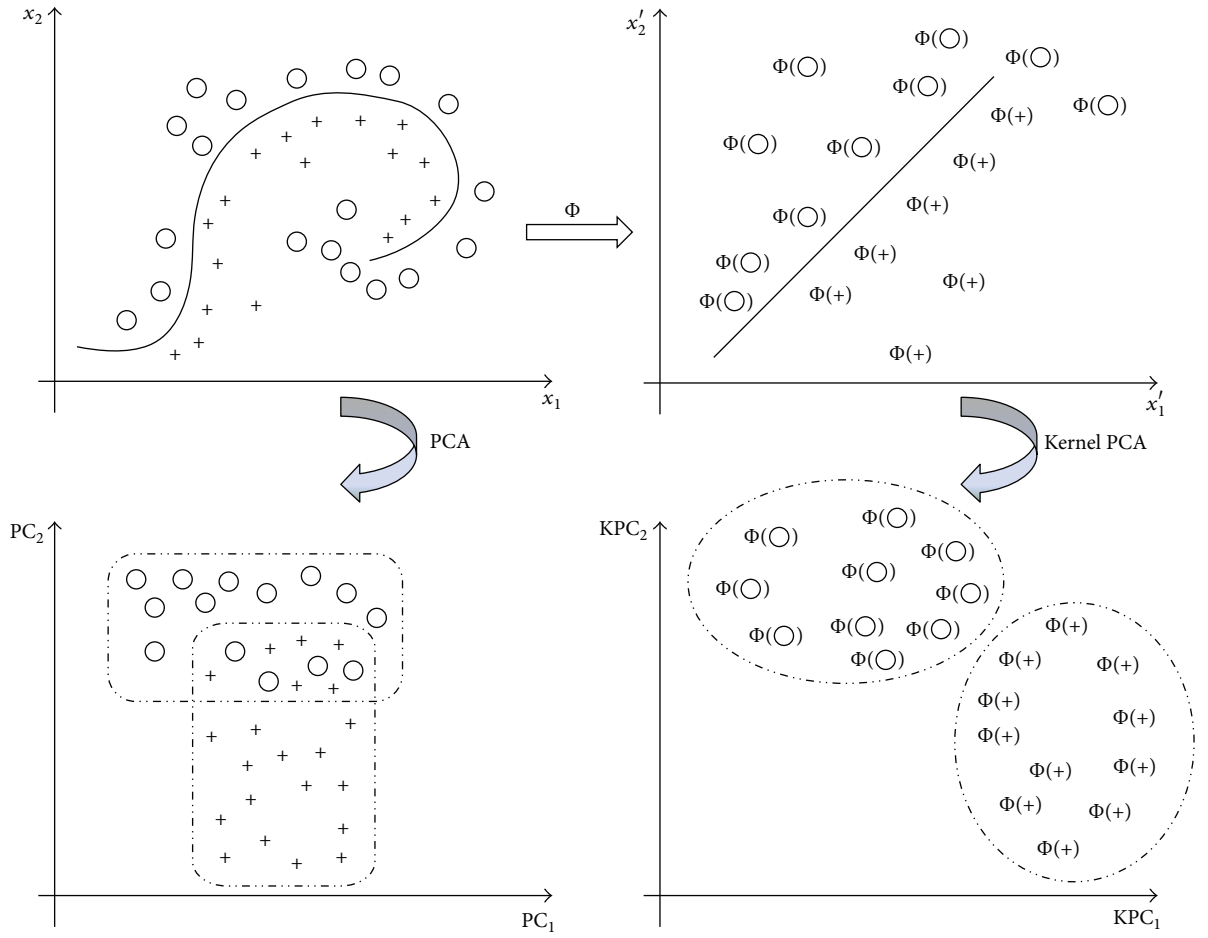


Figure 2: The mapping Φ embeds the data points in a feature space.

kernel functions in the input space to achieve the same effect of the expensive nonlinear mapping.

From Figure 2, it is found that the basic idea of KPCA is to map the original dataset into some higher dimensional feature space. In this complex space, PCA can be applied to establish a linear relationship which is nonlinear in the original input space [29, 30]. For the special case in which $\Phi(\mathbf{x}) = \mathbf{x}$, KPCA is equivalent to linear PCA. From this viewpoint, KPCA can be regarded as a generalized version of linear PCA.

For PCA, with data $\mathbf{X} = [x_1, x_2, \ldots, x_n]^T \in R^p$, one can first compute the covariance matrix $\mathbf{C}$:

$$\mathbf{C} = \frac{1}{n}\mathbf{X}^{\mathbf{T}}\mathbf{X}. \tag{5}$$

A principal component $\mathbf{v}$ is computed by solving the following eigenvalue problem:

$$\boldsymbol{\lambda}\mathbf{v} = \mathbf{C}\mathbf{v} = \frac{1}{n}\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{v}. \tag{6}$$

Thus, the eigenvectors can be written as ($\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^T$)

$$\mathbf{v} = \sum_{i=1}^{n} \alpha_i x_i = \mathbf{X}^{\mathbf{T}}\boldsymbol{\alpha}. \tag{7}$$

Then, the eigen value problem can be represented by the following simple form:

$$\boldsymbol{\lambda}\boldsymbol{\alpha} = \frac{1}{n}\mathbf{K}\boldsymbol{\alpha}, \tag{8}$$

where $\mathbf{K} = \mathbf{X}\mathbf{X}^{\mathbf{T}} \in R^{n\times n}$ is a linear kernel matrix. To derive KPCA, one firstly needs to map the data $\mathbf{X}$ into a feature space $\mathbf{F}$ (i.e., $\mathbf{M} = [\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \ldots, \Phi(\mathbf{x}_n)]^T$). Hence, a nonlinear kernel matrix $\mathbf{K}$ ($\mathbf{K} = \mathbf{M}\mathbf{M}^{\mathbf{T}} \in R^{n\times n}$) can be directly generated by means of specific kernel function ((1), (2), (3), and (4)). For extracting features of a new sample x with KPCA, one simply projects the mapped sample $\Phi(\mathbf{x})$ onto the first k projections $\mathbf{V}_k$,

$$\mathbf{V}_k \cdot \Phi(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^k \left\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) \right\rangle. \tag{9}$$

KPCA is to map the original data (in the input space) with nonlinear features into kernel feature space in which the linear PCA algorithm is then performed. Therefore, KPCA, being suitable to describe the nonlinear structure of data set, can be regarded as a generalized version of linear PCA.

2.3. GDA. Generalized discriminant analysis (GDA) is a method designed for nonlinear classification [31–33]. It is a nonlinear extension of linear discriminant analysis (LDA) based on a kernel function Φ which transforms the original space $\mathbf{X}$ to a new high dimensional feature space $\mathbf{F}$. The within-class (or total) scatter ($\mathbf{W}^{\Phi}$) and between-class scatter ($\mathbf{B}^{\Phi}$) matrixes of the nonlinearly mapped data are as follows:

$$\mathbf{W}^{\Phi} = \sum_{c=1}^{C} \sum_{\mathbf{x} \in \mathbf{X}_c} \Phi(\mathbf{x}) \Phi(\mathbf{x})^T, \tag{10}$$

$$\mathbf{B}^{\Phi} = \sum_{c=1}^{C} M_c \mathbf{m}_c^{\Phi} \left(\mathbf{m}_c^{\Phi}\right)^T. \tag{11}$$

In (11), $\mathbf{m}_c$ is the mean of class $\mathbf{X}_c$ and M_c is the number of samples belonging to $\mathbf{X}_c$. The aim of the GDA is to find such projection matrix $\mathbf{U}^{\Phi}$ that maximizes the following Fisher criterion:

$$\mathbf{U}_{\text{opt}}^{\Phi} = \arg\max \frac{\left|\left(\mathbf{U}^{\Phi}\right)^T \mathbf{B}^{\Phi} \mathbf{U}^{\Phi}\right|}{\left|\left(\mathbf{U}^{\Phi}\right)^T \mathbf{W}^{\Phi} \mathbf{U}^{\Phi}\right|} = \left[\mathbf{u}_1^{\Phi}, \ldots, \mathbf{u}_N^{\Phi}\right]. \tag{12}$$

From the theory of reproducing kernels, any solution $\mathbf{u}^{\Phi} \in \mathbf{F}$ must lie in the span of all training samples in $\mathbf{F}$:

$$\mathbf{u}^{\Phi} = \sum_{c=1}^{C} \sum_{i=1}^{M_c} \alpha_{ci} \Phi(\mathbf{x}_{ci}), \tag{13}$$

where α_{ci} are some real weights and x_{ci} is the ith sample of the class c. The solution is obtained by solving ($\boldsymbol{\alpha} = [\boldsymbol{\alpha}_c]$, $c = 1, 2, \ldots, C$; $\boldsymbol{\alpha}_c = [\boldsymbol{\alpha}_{ci}]$, $i = 1, 2, \ldots, M_c$):

$$\boldsymbol{\lambda} = \frac{\boldsymbol{\alpha}^T \mathbf{KDK} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \mathbf{KK} \boldsymbol{\alpha}}. \tag{14}$$

$\mathbf{K}$ is the $n \times n$ kernel matrix composed of the dot products of nonlinearly mapped data. And $\mathbf{D} = \text{diag}(\mathbf{D}_1, \ldots, \mathbf{D}_c)$, where $\mathbf{D}_i$ is a $n_i \times n_i$ matrix with entries all equal to $1/n_i$.

2.4. New Improved 3D Kernel Approach: 3D KPCA and 3D GDA. Traditional KPCA and GDA are typical multivariate two-dimension statistical methods. In this work, KPCA and GDA are improved with three-dimensional projection and the concept of electric field intensity.

Firstly, the data of training samples are projected onto three-dimensional space by KPCA or GDA algorithm with satisfactory classification effect. The three-dimensional coordinate axes are, respectively, the first kernel principal component, second kernel principal component, and third kernel principal component or the direction vectors of generalized discriminant analysis.

Secondly, we need to estimate the class (unknown) of new projection points, such as membrane protein types of test sample data. There are two estimation methods in this work: K-Nearest Neighbor algorithm (KNN) [34] and class intensity model.

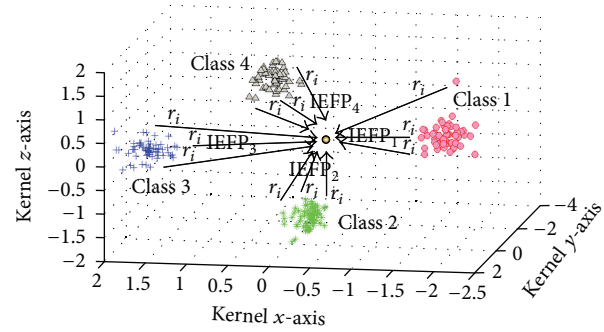


Figure 3: IEFP with different classes in 3D kernel space.

KNN algorithm estimation: new projection point (test sample) is classified by a majority vote of its neighbors (training samples in kernel three-dimensional space).

Class intensity model estimation: the projection point of one training data can be considered as point charge. The species of charge is related to the class of sample. And the Electric Quantity of Point Charge (EQPC) is related to the number of samples (n_C) which belongs to some class:

$$\text{EQPC}_C = \frac{1}{n_C}. \tag{15}$$

The value of EQPC is negative related with the sample amount of same class. Based on the Coulomb law and formula of intensity of electric field, the Intensity of Electric Field of one Point (IEFP) in 3D space is

$$\text{IEFP}_C = \sum_{i=1}^{n_C} \frac{\text{EQPC}_C}{r_i^2}, \tag{16}$$

where r is distance between point charge and the space point.

Therefore, in class intensity model, IEFP is a criterion of classification. For example, there are four classes in training data: class 1, class 2, class 3, and class 4 in Figure 3. After projecting with kernel methods, all projection class charge points of training data can form a space electric field. The test sample can be projected onto this space with the same kernel methods. Figure 3 illustrates the relationship between point charge of different class and corresponding IEFP. To project position of test sample, if there exist $\text{IEFP}_1 > \text{IEFP}_2$, $\text{IEFP}_1 > \text{IEFP}_3$ and $\text{IEFP}_1 > \text{IEFP}_4$, test sample should belong to class 1.

3. Results and Discussion

3.1. System and Software Used for Data Analysis. The calculations were carried out using the Intel(R) Core(TM) Duo CPU T5870 GHz computer running Windows XP operating system. All the learning input data were range-scaled to [0~1] in this work. The improved 3D kernel approach software package including 3D kernel PCA and 3D GDA was programmed in our laboratory referring to the literature [29, 31] based on statistical pattern recognition toolbox for MATLAB [35].

3.2. Application of Improved 3D Kernel Approach to Protein's Tertiary Structure Classes of Domains. The protein datasets

Table 1: LOOCV success rates by component-coupled, neural network, SVMs, AdaBoost, and improved 3D kernel approach.

Dataset	Algorithm	All-α	All-β	α/β	$\alpha+\beta$	Overall
Dataset A (277 domains)	Component-coupled	84.3%	82.0%	81.5%	67.7%	79.1%
	Neural networks	68.6%	85.2%	86.4%	56.9%	74.7%
	SVMs	74.3%	82.0%	87.7%	72.3%	79.4%
	AdaBoost	87.1%	95.1%	98.7%	81.5%	90.9%
	3D kernel	88.6%	85.3%	93.8%	77.0%	86.6%
Dataset B (498 domains)	Component-coupled	93.5%	88.9%	90.4%	84.5%	89.2%
	Neural networks	86.0%	96.0%	88.2%	86.0%	89.2%
	SVMs	88.8%	95.2%	96.3%	91.5%	93.2%
	AdaBoost	96.2%	92.1%	98.5%	89.9%	94.2%
	3D kernel	91.6%	95.3%	99.3%	92.3%	95.0%

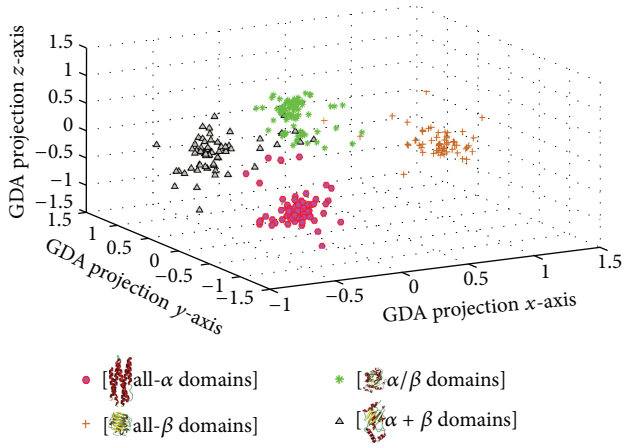


Figure 4: Distribution of different protein's tertiary structure classes data in 3D kernel space.

studied here were taken from Niu and his coworkers [17]. In dataset A, there are 277 protein domains, of which 70 are all-α domains, 61 all-β, 81 α/β, and 65 $\alpha+\beta$. In dataset B, there are 498 protein domains, of which 107 are all-α domains, 126 all-β, 136 α/β, and 129 $\alpha+\beta$. The amino acid composition was used to represent the sample of a protein domain.

To demonstrate the power of 3D kernel methods, computations were performed by the Leave-One-Out Cross-Validation (LOOCV), which are widely used by more and more investigators in testing the power of various predictors. As such, the data set of n samples was divided into two disjoint subsets including a training data set ($n-1$ samples) and a test data set (only 1 sample). After developing each model based on the training set, the omitted data was predicted and the difference between experimental value and predicted value was calculated [36–38].

Based on dataset A, it was found that the projection with Gaussian (see (3), $\sigma = 0.5$) kernel function and KNN ($K = 3$) algorithm estimation was suitable for building 3D kernel PCA model with the better success rates.

Based on dataset B, it was found that the projection with polynomial (see (2), $d = 4$, $\theta = 1.5$) kernel function and class intensity model estimation was suitable for building 3D GDA model with the better success rates. Figure 4 illustrates the protein domains classes distribution of dataset B (498 samples) in 3D kernel space with GDA model. It can be seen that the data points, which belong to all-α domains, all-β domains, α/β domains, and $\alpha+\beta$ domains respectively, are located in different regions with a correct classification result.

The success rates thus obtained are given in Table 1, where, for facilitating comparison, the corresponding rates obtained by component-coupled algorithm, neural networks, support vector machines (SVMs), and AdaBoost Learner [17] are also listed.

As it can be seen from Table 1, the performance of improved 3D kernel model outperforms those of component-coupled, neural networks, SVMs models but was a little worse than that of AdaBoost model for the dataset A (277 domains) available in LOOCV test. Based on dataset B (498 domains), improved 3D kernel learner is superior to all the other predictors in identifying the structural classification.

3.3. Application of Improved 3D Kernel Approach to Classification of Membrane Proteins. The membrane proteins dataset studied here was collected from the literature [25]. The dataset contains 2059 prokaryotic proteins (type A membrane proteins: 435; type B membrane proteins: 152; type C Multi-pass transmembrane proteins: 1311; type D lipid chain-anchored membrane proteins: 51; type E GPI-anchored membrane proteins: 110). The amino acid composition was selected as the input of the classification algorithm, and the computations were performed by LOOCV to test the power of various predictors. Based on dataset of membrane proteins, the classification flow chart (Figure 5) was obtained as follows.

From Figure 5, there are two steps in building classification model. Firstly, the 3D KPCA model with projection through polynomial (see (2), $d = 2$, $\theta = 0.1$) kernel function and KNN ($K = 5$) algorithm estimation was built to classify the multipass transmembrane proteins (type C) and the other membrane proteins (type A, type B, type D, and type E). Figure 6 illustrates the data distribution of type C and other membrane proteins in 3D kernel space with KPCA model.

Secondly, the 3D GDA model with Gaussian (see (3), $\sigma = 5$) kernel function and class intensity model estimation was built to classify type A, type B, type D, and type E membrane proteins.

Table 2: LOOCV success rates by covariant discriminant, neural network, SVM, bagging, and improved 3D kernel approach.

Algorithm	Rate of correct prediction for each class					Overall rate of correct prediction
	Type A	Type B	Type C	Type D	Type E	
Covariant discriminant	74.0%	52.0%	83.7%	49%	45.4%	76.4%
Neural network	75.63%	30.92%	88.86%	50.98%	30.91%	77.76%
SVMs	77.7%	28.3%	92.5%	52.9%	35.5%	80.9%
Bagging	79.80%	48.68%	93.21%	49.02%	60.91%	84.18%
3D kernel	78.11%	31.02%	94.36%	52.63%	45.46%	84.50%

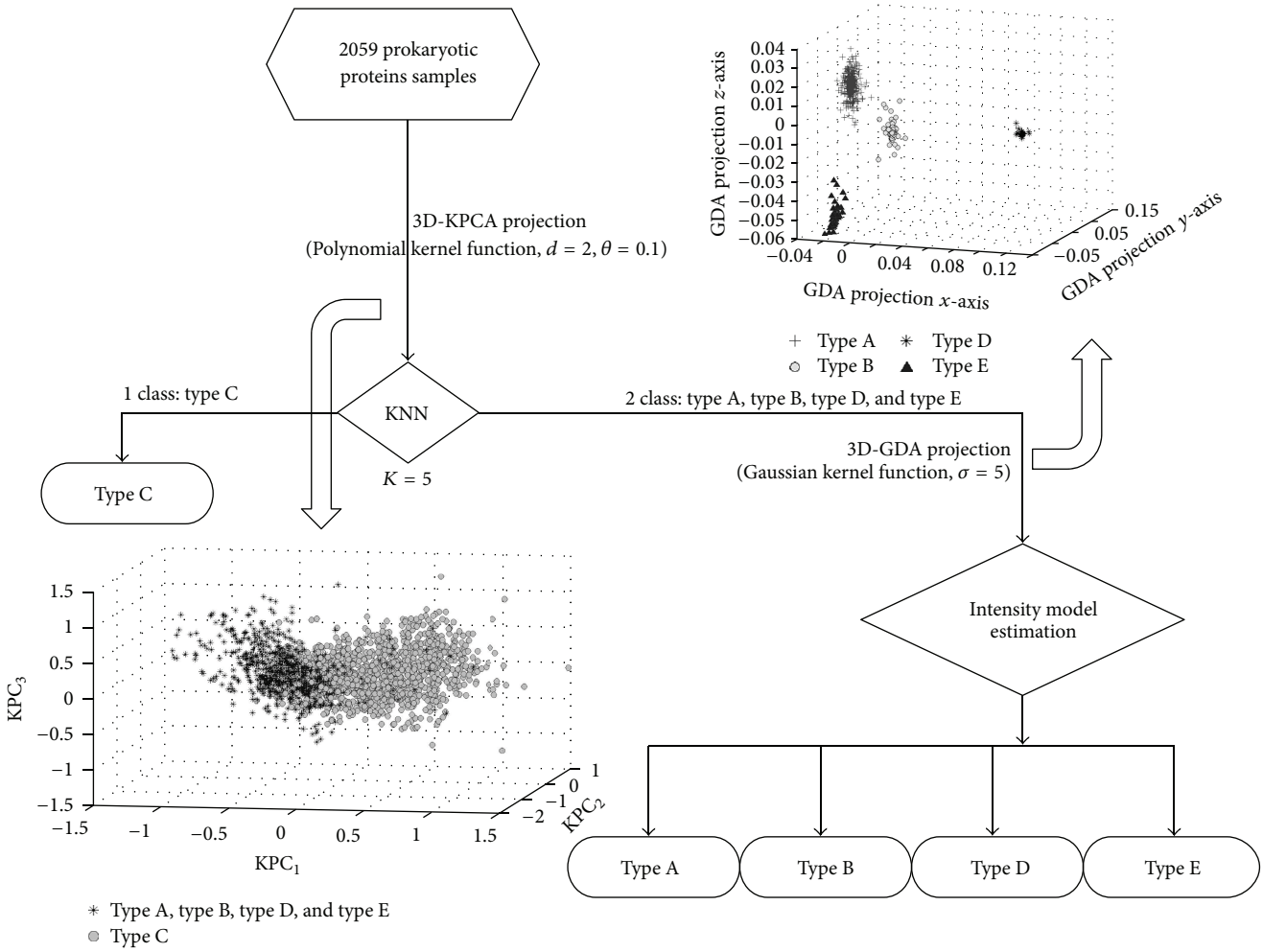


Figure 5: Classification flow chart of five type membrane proteins.

Figure 7 illustrates the data distribution of the type A, type B, type D, and type E membrane proteins in 3D kernel space with GDA model. 3D kernel method was compared with other machine learning classification methods: the covariant discriminant algorithm [23], neural networks, support vector machines, and Bagging [25], as is shown in Table 2.

As we can see from Table 2, correct classification rate of the LOOCV test applied 3D kernel algorithm outperformed other algorithms. It also means that 3D kernel method has learned very well through the membrane proteins training process.

4. Conclusions

The 3D kernel approach is very useful machine learning classifier. It has remarkably outperformed the powerful neural network, SVM classifiers, in predicting the protein domain structural classes for the two datasets constructed and membrane protein types for the same dataset constructed by previous investigators. It is thus anticipated that the 3D Kernel classifier can also be used to predict other protein attributes, such as sub-cellular localization [39–41], enzyme family and subfamily classes [42], and active sites of enzyme. The concepts of EQPC and IEFP can be easily extended to

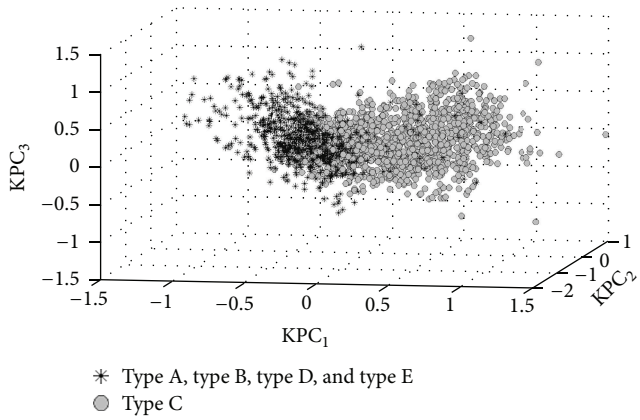


Figure 6: Distribution of the multipass transmembrane proteins (type C) and the other membrane proteins (type A, type B, type D and type E) data in 3D kernel space.

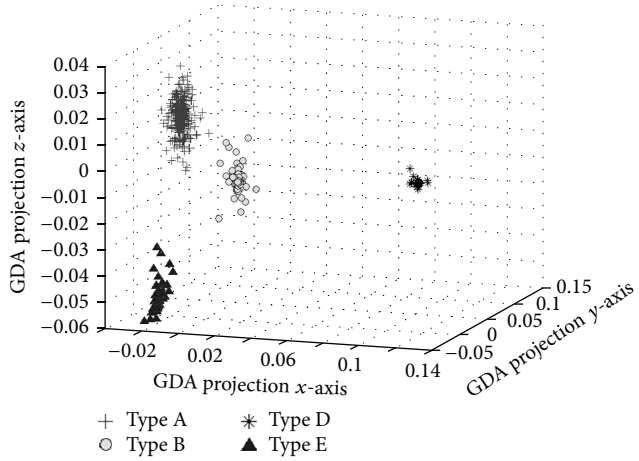


Figure 7: Distribution of the type A, type B, type D, and type E data in 3D kernel space.

many-dimensional space and could be improved to use four or more dimensions.

It could be concluded that 3D kernel approach is a robust and highly accurate classification technique that can be successfully applied to derive statistical models with statistical qualities and predictive capabilities for the protein location and function. The 3D kernel algorithm should be a complementary tool to the existing pattern recognition in chemometrics and bioinformatics.

Authors' Contribution

Xu Liu and Yuchao Zhang contributed equally to this work.

Acknowledgments

The project is financially supported by National Natural Science Foundation of China (nos. 20373040, 20973108, 20942005, and 21262005), Innovation Foundation of Guangxi University (nos. XBZ120947), and Innovation Foundation of Shanghai University (nos. A.10-0101-10-006). The work was supported by Guangxi Key Laboratory of Traditional Chinese Medicine Quality Standards (Guangxi Institute of Traditional Medical and Pharmaceutical Sciences) (guizhongzhongkai0802).

References

[1] V. Brusic, G. Rudy, M. Honeyman, J. Hammer, and L. Harrison, "Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network," *Bioinformatics*, vol. 14, no. 2, pp. 121–130, 1998.

[2] L. Xu, L. Wencong, J. Shengli, L. Yawei, and C. Nianyi, "Support vector regression applied to materials optimization of sialon ceramics," *Chemometrics and Intelligent Laboratory Systems*, vol. 82, no. 1-2, pp. 8–14, 2006.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[4] B. M. Nicolaï, K. I. Theron, and J. Lammertyn, "Kernel PLS regression on wavelet transformed NIR spectra for prediction of sugar content of apple," *Chemometrics and Intelligent Laboratory Systems*, vol. 85, no. 2, pp. 243–252, 2007.

[5] Y. Qu, B.-L. Adam, Y. Yasui et al., "Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients," *Clinical Chemistry*, vol. 48, no. 10, pp. 1835–1843, 2002.

[6] B. Niu, X.-C. Yuan, P. Roeper et al., "HIV-1 protease cleavage site prediction based on two-stage feature selection method," *Protein and Peptide Letters*, vol. 20, no. 3, pp. 290–298, 2013.

[7] B. Niu, Q. Su, X.-C. Yuan, W. Lu, and J. Ding, "QSAR study on 5-lipoxygenase inhibitors based on support vector machine," *Medicinal Chemistry*, vol. 8, no. 6, pp. 1108–1116, 2012.

[8] C.-R. Peng, W.-C. Lu, B. Niu, M.-J. Li, X.-Y. Yang, and M.-L. Wu, "Predicting the metabolic pathways of small molecules based on their physicochemical properties," *Protein & Peptide Letters*, vol. 19, pp. 1250–1256, 2012.

[9] Q. Su, W.-C. Lu, B. Niu, X. Liu, and T.-H. Gu, "Classification of the toxicity of some organic compounds to tadpoles (Rana Temporaria) through integrating multiple classifiers," *Molecular Informatics*, vol. 30, no. 8, pp. 672–675, 2011.

[10] B. Niu, W.-C. Lu, J. Ding et al., "Site of O-glycosylation prediction based on two stage feature selection," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, pp. 142–145, 2011.

[11] A. V. Finkelstein and O. B. Ptitsyn, "Why do globular proteins fit the limited set of foldin patterns?" *Progress in Biophysics and Molecular Biology*, vol. 50, no. 3, pp. 171–190, 1987.

[12] K.-C. Chou and L. Carlacci, "Energetic approach to the folding of α/β barrels," *Proteins: Structure, Function and Genetics*, vol. 9, no. 4, pp. 280–295, 1991.

[13] K.-C. Chou, "Progress in protein structural class prediction and its impact to bioinformatics and proteomics," *Current Protein & Peptide Science*, vol. 6, no. 5, pp. 423–436, 2005.

[14] K. Oxenoid and J. J. Chou, "The structure of phospholamban pentamer reveals a channel-like architecture in membranes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 31, pp. 10870–10875, 2005.

[15] J. S. Richardson, "β sheet topology and the relatedness of proteins," *Nature*, vol. 268, no. 5620, pp. 495–500, 1977.

[16] O. B. Ptitsyn and A. V. Finkelstein, "Similarities of protein topologies: evolutionary divergence, functional convergence or

principles of folding?" *Quarterly Reviews of Biophysics*, vol. 13, no. 3, pp. 339–386, 1980.

[17] B. Niu, Y.-D. Cai, W.-C. Lu, G.-Z. Li, and K.-C. Chou, "Predicting protein structural class with AdaBoost Learner," *Protein and Peptide Letters*, vol. 13, no. 5, pp. 489–492, 2006.

[18] D. A. Doyle, J. M. Cabral, R. A. Pfuetzner et al., "The structure of the potassium channel: molecular basis of K+ conduction and selectivity," *Science*, vol. 280, no. 5360, pp. 69–77, 1998.

[19] J. R. Schnell and J. J. Chou, "Structure and mechanism of the M2 proton channel of influenza A virus," *Nature*, vol. 451, no. 7178, pp. 591–595, 2008.

[20] L. Stouffer Amanda, A. Rudresh, and S. David, "Structural basis for the function and inhibition of an influenza virus proton channel," *Nature*, vol. 451, pp. 596–599, 2008.

[21] M. D. Resh, "Myristylation and palmitylation of Src family members: the fats of the matter," *Cell*, vol. 76, no. 3, pp. 411–413, 1994.

[22] K.-C. Chou and D. W. Elrod, "Protein subcellular location prediction," *Protein Engineering*, vol. 12, no. 2, pp. 107–118, 1999.

[23] K.-C. Chou and D. W. Elrod, "Prediction of membrane protein types and subcellular locations," *Proteins*, vol. 34, pp. 137–153, 1999.

[24] K.-C. Chou, "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space," *Proteins: Structure, Function and Genetics*, vol. 21, no. 4, pp. 319–344, 1995.

[25] B. Niu, Y.-H. Jin, K.-Y. Feng et al., "Predicting membrane protein types with bagging learner," *Protein & Peptide Letters*, vol. 15, no. 6, pp. 590–594, 2008.

[26] V. Vapnik, *Statistical Learning Theory*, John Wiley & Johns, New York, NY, USA, 1998.

[27] D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte, and L. Kaufman, *Chemometrics: A Textbook*, Elsevier Science Publishers B.V., Amsterdam, The Netherlands, 1988.

[28] R. Bro, "PARAFAC. Tutorial and applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 38, no. 2, pp. 149–171, 1997.

[29] W. Wu, D. L. Massart, and S. de Jong, "The kernel PCA algorithms for wide data. Part I: theory and algorithms," *Chemometrics and Intelligent Laboratory Systems*, vol. 36, no. 2, pp. 165–172, 1997.

[30] D.-S. Cao, Y.-Z. Liang, Q.-S. Xu, Q.-N. Hu, L.-X. Zhang, and G.-H. Fu, "Exploring nonlinear relationships in chemical data using kernel-based methods," *Chemometrics and Intelligent Laboratory Systems*, vol. 107, no. 1, pp. 106–115, 2011.

[31] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[32] H. Yamamoto, H. Yamaji, Y. Abe et al., "Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables," *Chemometrics and Intelligent Laboratory Systems*, vol. 98, no. 2, pp. 136–142, 2009.

[33] H. Wang, Z. Hu, and Y. Zhao, "An efficient algorithm for generalized discriminant analysis using incomplete Cholesky decomposition," *Pattern Recognition Letters*, vol. 28, no. 2, pp. 254–259, 2007.

[34] B. S. Kim and S. B. Park, "A fast k nearest neighbor finding algorithm based on the ordered partition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 761–766, 1986.

[35] S. Sonnenburg, G. Rätsch, S. Henschel et al., "The Shogun machine learning toolbox," *The Journal of Machine Learning Research*, vol. 11, pp. 1799–1802, 2010.

[36] S. R. Amendolia, G. Cossu, M. L. Ganadu, B. Golosio, G. L. Masala, and G. M. Mura, "A comparative study of K-nearest neighbour, support vector machine and multi-layer perceptron for Thalassemia screening," *Chemometrics and Intelligent Laboratory Systems*, vol. 69, no. 1-2, pp. 13–20, 2003.

[37] M. Kearns and D. Ron, "Algorithmic stability and sanity-check bounds for leave-one-out cross-validation," in *Proceedings of the 10th Annual Conference on Computational Learning Theory*, pp. 152–162, ACM Press, July 1997.

[38] S. B. Holden, "PAC-like upper bounds for the sample complexity of leave-one-out cross-validation," in *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pp. 41–50, Desenzano del Garda, Italy, July 1996.

[39] G.-P. Zhou and K. Doctor, "Subcellular location prediction of apoptosis proteins," *Proteins: Structure, Function and Genetics*, vol. 50, no. 1, pp. 44–48, 2003.

[40] Y.-X. Pan, Z.-Z. Zhang, Z.-M. Guo, G.-Y. Feng, Z.-D. Huang, and L. He, "Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach," *Journal of Protein Chemistry*, vol. 22, no. 4, pp. 395–402, 2003.

[41] K.-C. Chou and Y.-D. Cai, "Predicting protein localization in budding yeast," *Bioinformatics*, vol. 21, no. 7, pp. 944–950, 2005.

[42] K.-C. Chou and Y.-D. Cai, "Predicting enzyme family class in a hybridization space," *Protein Science*, vol. 13, no. 11, pp. 2857–2863, 2004.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 285063, 7 pages
http://dx.doi.org/10.1155/2013/285063

Research Article

MicroRNA-Mediated Regulation in Biological Systems with Oscillatory Behavior

Zhiyong Zhang,[1,2] Fengdan Xu,[1,2] Zengrong Liu,[2] Ruiqi Wang,[2] and Tieqiao Wen[2]

[1] *Department of Mathematics, Shanghai University, Shanghai 200444, China*
[2] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*

Correspondence should be addressed to Zengrong Liu; zrongliu@126.com

Received 30 April 2013; Revised 3 June 2013; Accepted 4 June 2013

Academic Editor: Tao Huang

As a class of small noncoding RNAs, microRNAs (miRNAs) regulate stability or translation of mRNA transcripts. Some reports bring new insights into possible roles of microRNAs in modulating cell cycle. In this paper, we focus on the mechanism and effectiveness of microRNA-mediated regulation in the cell cycle. We first describe two specific regulatory circuits that incorporate base-pairing microRNAs and show their fine-tuning roles in the modulation of periodic behavior. Furthermore, we analyze the effects of *miR369-3* on the modulation of the cell cycle, confirming that *miR369-3* plays a role in shortening the period of the cell cycle. These results are consistent with experimental observations.

1. Introduction

MicroRNAs are single-stranded non-coding RNA molecules containing 21~23-nucleotides. More and more works imply that microRNAs are involved in a series of important life processes, including early development, cell proliferation, differentiation, and apoptosis [1–11]. MicroRNAs act by base pairing with their target mRNAs and induce either translational repression or mRNA degradation through an RNA-induced silencing complex. Most microRNAs negatively regulate expression of their target genes. Since microRNA is a type of small molecules and needs not to be translated into proteins, it has an energy-saving advantage for the cell cycle regulation when compared to the regulation by proteins. In addition, its faster synthesis rate has more advantage in response to the changes in environment. These advantages mean that microRNA may play crucial roles in gene regulation.

A series of recent experiments show that microRNAs may play crucial roles in modulating periodic behaviors of biological systems, such as cell cycle and circadian rhythm [12–19]. Other reports indicate that microRNAs fine-tune oscillations of *p53* in the process of tumor suppression [20–29]. However, all these findings have still been confined to experimental stage. The operating mechanism and potential implication of microRNA-mediated regulation in the modulation of periodic behavior are less clear and need to be further investigated.

In this study, we aim to explore the control mechanism and kinetic characteristics of microRNA-mediated regulation in the modulation of cell cycle. First, we model two specific network motifs, which have different periodic behaviors in the absence of microRNA, that is, oscillation generated by a Hopf bifurcation and relaxation oscillation. Furthermore, microRNA is incorporated into these two motifs, respectively. The dynamical analysis confirms that microRNA can regulate these two types of oscillations by shortening their periods. Then we study the microRNA regulation of a periodic phenomenon in biological system, that is, cell cycle. The results account for the roles of microRNA in the modulation of cell cycle observed in recent experiments.

2. MicroRNA Regulation of Two Motifs and Cell Cycle

2.1. Analysis of Motif I. The first motif without and with the regulation of microRNA is shown in Figure 1. In this motif,

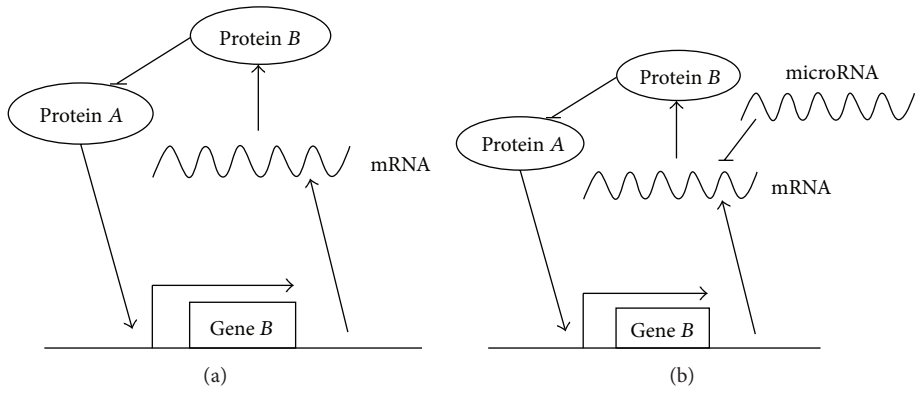


Figure 1: Schematic description of the first motif without and with microRNA regulation. (a) Without microRNA regulation; (b) with microRNA regulation.

protein A activates the transcription of gene B, which in turn inhibits the transcription of gene A, as shown in Figure 1(a).

With these assumptions, when the microRNA is not incorporated, we get a set of differential equations on the concentrations of the substrates, which describe the behaviors of the mRNA of gene B, protein B, and protein A, as follows:

$$\begin{aligned}\frac{d[m]}{dt} &= a_1 f([m]) - d_m[m],\\ \frac{d[B]}{dt} &= a_2[m] - d_B[B],\\ \frac{d[A]}{dt} &= a_3 g([B]) - d_A[A],\end{aligned} \tag{1}$$

where $[\cdot]$ means the concentration of the substrate, and $f([A]) = [A]^{n_1}/([A]^{n_1} + K_A^{n_1})$, $g([B]) = K_B^{n_2}/([B]^{n_2} + K_B^{n_2})$, where n_1, n_2 are hill coefficients.

When the microRNA regulation is incorporated into (1) (see Figure 1(b)), it becomes

$$\begin{aligned}\frac{d[m]}{dt} &= a_1 f([m]) - d_m[m] - d[m][s],\\ \frac{d[B]}{dt} &= a_2[m] - d_B[B],\\ \frac{d[A]}{dt} &= a_3 g([B]) - d_A[A],\\ \frac{d[s]}{dt} &= \lambda - d_s[s] - d[m][s],\end{aligned} \tag{2}$$

where λ is the synthesis rate of microRNA, d_s is the degradation rate of microRNA, and d is the associate rate of two RNAs.

Using the dimensionless variables scaled by $\tau = a_1 t$, $\widehat{B} = d[B]/K_B$, $\widehat{A} = d[A]/K_A$, $\widehat{a}_2 = a_2/a_1 K_B$, $\widehat{a}_3 = a_3/a_1 K_A$, $\widehat{d}_m = d_m/a_1$, $\widehat{d}_B = d_B/a1$, $\widehat{d}_A = d_A/a1$, $\widehat{f}(\widehat{A}) = \widehat{A}^{n_1}/(\widehat{A}^{n_1} + 1)$, $\widehat{g}(\widehat{B}) = 1/(\widehat{B}^{n_2} + 1)$, and $\widehat{\lambda} = \lambda/a_1$, $\widehat{d}_s = d_s/a_1$, $\widehat{d} = d/a_1$, we obtain two sets of equations as follows:

$$\begin{aligned}\frac{d[m]}{d\tau} &= \widehat{f}\left(\widehat{A}\right) - \widehat{d}_m[m],\\ \frac{d\widehat{B}}{d\tau} &= \widehat{a}_2[m] - \widehat{d}_B\widehat{B},\\ \frac{d\widehat{A}}{d\tau} &= \widehat{a}_3\widehat{g}\left(\widehat{B}\right) - \widehat{d}_A\widehat{A},\end{aligned} \tag{3}$$

$$\begin{aligned}\frac{d[m]}{d\tau} &= \widehat{f}\left(\widehat{A}\right) - \widehat{d}_m[m] - \widehat{d}\,[m]\,[s],\\ \frac{d\widehat{B}}{d\tau} &= \widehat{a}_2[m] - \widehat{d}_B\widehat{B},\\ \frac{d\widehat{A}}{d\tau} &= \widehat{a}_3\widehat{g}\left(\widehat{B}\right) - \widehat{d}_A\widehat{A},\\ \frac{d[s]}{d\tau} &= \widehat{\lambda} - \widehat{d}_s[s] - \widehat{d}\,[m]\,[s].\end{aligned} \tag{4}$$

We can calculate the equilibriums by setting the right-hand sides of the equations equal to zero. Let the equilibrium be (m^*, B^*, A^*) for (3). Linearizing (3) around the equilibrium results in the Jacobian matrix as follows:

$$\mathbf{J_1} = \begin{pmatrix} -\widehat{d}_m & 0 & \widehat{f}'(A^*) \\ \widehat{a}_2 & -\widehat{d}_B & 0 \\ 0 & \widehat{a}_3\widehat{g}'(B^*) & -\widehat{d}_A \end{pmatrix}. \tag{5}$$

By simple calculation, its characteristic polynomial is derived as follows:

$$\begin{aligned}&x^3 + \left(\widehat{d}_m + \widehat{d}_B + \widehat{d}_A\right)x^2 + \left(\widehat{d}_m\widehat{d}_B + \widehat{d}_m\widehat{d}_A + \widehat{d}_B\widehat{d}_A\right)x\\ &\quad + \widehat{d}_m\widehat{d}_B\widehat{d}_A - \widehat{a}_2\widehat{a}_3\widehat{f}'(A^*)\,\widehat{g}'(B^*) = 0.\end{aligned} \tag{6}$$

Using Routh-Hurwitz criteria, we know that when the following condition is satisfied:

$$\begin{aligned}&\left(\widehat{d}_m + \widehat{d}_B + \widehat{d}_A\right)\left(\widehat{d}_m\widehat{d}_B + \widehat{d}_m\widehat{d}_A + \widehat{d}_B\widehat{d}_A\right)\\ &\quad = \widehat{d}_m\widehat{d}_B\widehat{d}_A - \widehat{a}_2\widehat{a}_3\widehat{f}'(A^*)\,\widehat{g}'(B^*),\end{aligned} \tag{7}$$

a Hopf bifurcation occurs (and the right-hand side should be also larger than 0). In other words, when the left-hand side becomes less than the right-hand side in the above equation, oscillations will occur in (3).

Let the equilibrium be $(m^\circ, B^\circ, A^\circ, s^\circ)$ for (4). Similarly, linearizing (4) around the equilibrium results in the Jacobian matrix as follows:

$$\mathbf{J_2} = \begin{pmatrix} -\hat{d}_m - \hat{d}s^\circ & 0 & \hat{f}'(A^\circ) & -\hat{d}m^\circ \\ \hat{a}_2 & -\hat{d}_B & 0 & 0 \\ 0 & \hat{a}_3\hat{g}'(B^\circ) & -\hat{d}_A & 0 \\ -\hat{d}s^\circ & 0 & 0 & -\hat{d}_s - \hat{d}m^\circ \end{pmatrix}. \quad (8)$$

By calculation, its characteristic polynomial is derived as follows:

$$x^4 + \alpha_1 x^3 + \alpha_2 x^2 + \alpha_3 x + \alpha_4 = 0, \quad (9)$$

where $\alpha_1 = \hat{d}_m + \hat{d}_B + \hat{d}_A + \hat{d}_s + \hat{d}s^\circ + \hat{d}m^\circ$, $\alpha_2 = \hat{d}s^\circ\hat{d}_B + \hat{d}s^\circ\hat{d}_A + \hat{d}^2 s^\circ m^\circ + \hat{d}_B\hat{d}m^\circ + \hat{d}s^\circ\hat{d}_s + \hat{d}_B\hat{d}_s + \hat{d}_A\hat{d}_s + \hat{d}_m\hat{d}m^\circ - \hat{d}^2 m^{\circ 2} + \hat{d}_m\hat{d}_A + \hat{d}_B\hat{d}_A + \hat{d}_m\hat{d}_s + \hat{d}_m\hat{d}_B + \hat{d}_A\hat{d}m^\circ$, and so on. Again, according to the Routh-Hurwitz criteria, we can obtain the condition of Hopf bifurcation as follows:

$$\Delta_3 = \begin{vmatrix} \alpha_1 & \alpha_3 & 0 \\ 1 & \alpha_2 & \alpha_4 \\ 0 & \alpha_1 & \alpha_3 \end{vmatrix} = 0, \quad (10)$$

that is, when the determinant becomes less than zero (and $\alpha_1 > 0, \alpha_3 > 0, \alpha_4 > 0$), oscillations will occur in (4).

To investigate the dynamical properties of microRNA regulation in gene expression, we will next examine and compare the two models by computational analysis. To determine fundamental differences in these two models without and with microRNA regulation, it would be of interest to choose similar parameter values. We first choose values of parameters in (3), under which oscillations may occur. Then, under these fixed parameter values, other values of parameter in (4) are chosen so as to produce oscillations.

In Figure 2, the following parameter values in (3) are used: $n_1 = 2$, $n_2 = 2$, $\hat{a}_2 = 1.4$, $\hat{a}_3 = 3.4$, $\hat{d}_m = 0.1$, $\hat{d}_B = 0.1$, and $\hat{d}_A = 0.15$. Other parameter values in (4) are $\hat{\lambda} = 2.0$, $\hat{d}_s = 0.2$, and $\hat{d} = 0.8$. It can be seen that after the incorporation of the microRNA regulation, the changes of the wave form and amplitude of the oscillation are slight. In contrast, the period is shortened evidently, which means that the fine-tuning of microRNA regulation in the oscillation is period shortening.

2.2. Analysis of Motif II. The second motif is similar to the first one except the positive autoregulation of protein A, as shown in Figure 3.

When the microRNA is not incorporated, it has been discussed in [30]. It was shown that relaxation oscillation could occur due to the existence of different time scales. Using

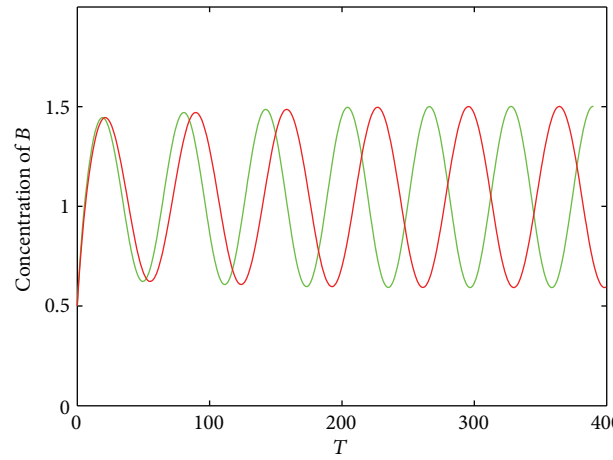


Figure 2: The modulation of oscillations by microRNA in motif I. The green and red lines show the concentration of protein B with and without microRNA regulation, respectively.

their model, the system without microRNA regulation can be expressed as follows:

$$\frac{dA}{dt} = \Delta\left(a_1 + \frac{1+\rho A^2}{1+\rho A^2+\sigma B^2} - A\right),$$
$$\frac{dB}{dt} = \Delta a_2 \frac{1+\rho A^2}{1+A^2} - B, \quad (11)$$

where Δ is the ratio of degradation rates between the two proteins A and B, $\Delta \gg 1$, which means degradation rate of A is much faster than that of B. $\alpha_2 = \epsilon\alpha_1$ with $0 < \epsilon \ll 1$ means that the synthesis rate of protein B is much slower than that of A. Similar to the first motif, after the incorporation of miRNA regulation, the system can be rewritten as follows:

$$\frac{dA}{dt} = \Delta\left(a_1 + \frac{1+\rho A^2}{1+\rho A^2+\sigma B^2} - A\right),$$
$$\frac{dB}{dt} = \Delta a_2 \frac{1+\rho A^2}{1+A^2} - B - dBs, \quad (12)$$
$$\frac{ds}{dt} = \lambda - d_s s - dBs,$$

where the first equation means that microRNA s and protein B codegrade nonlinearly at a rate d besides their respective linear degradation. Similarly, the parameter values in (11) are chosen first so as to produce relaxation oscillations. Then other parameters in (12) are chosen. More exactly, parameter values are $a_1 = 1.58$, $a_2 = 0.05a_1$, $\rho = 50$, $\sigma = 1$, $\Delta = 11$, $\lambda = 5$, $d_s = 0.1$, and $d = 0.1$.

When the microRNA is not incorporated, relaxation oscillation occurs, as shown by the red line in Figure 4. Similar to the first motif, after the incorporation of the microRNA, the changes of the wave form and amplitude of the relaxation oscillation are slight. In contrast, the period is shortened significantly too, meaning that the main modulation of microRNA regulation in relaxation oscillation is also period shortening.

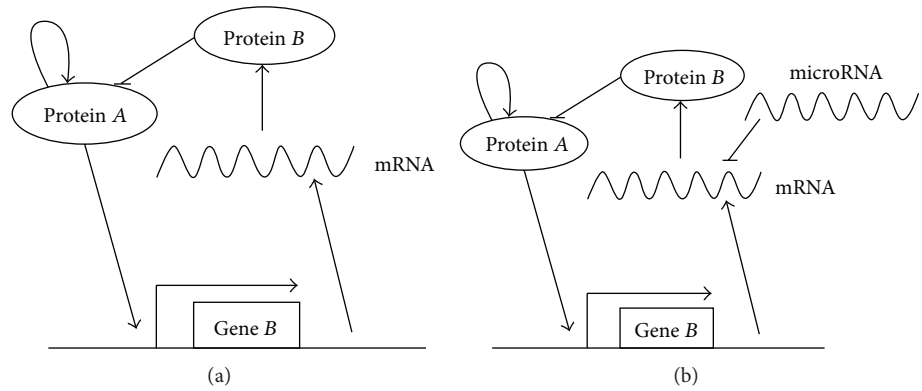


Figure 3: Schematic description of the second motif without and with microRNA regulation. (a) Without microRNA regulation; (b) with microRNA regulation.

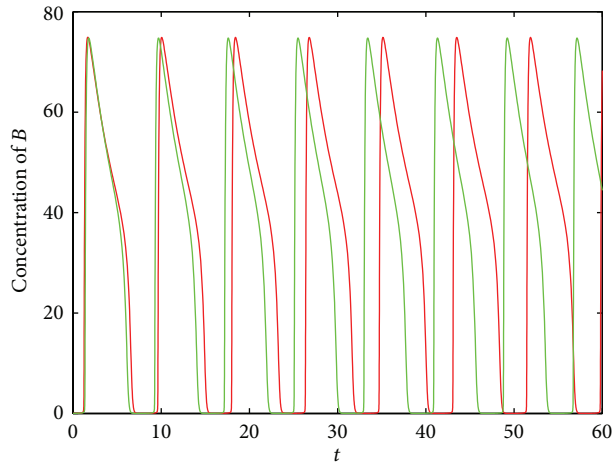


Figure 4: The modulation of relaxation oscillation by microRNA in motif II. The green and red lines show the concentration of protein B with and without microRNA regulation, respectively.

For both of the motifs, we get similar results; that is, the main modulation of microRNA regulation in periodic behavior is period shortening. In contrast, the types of oscillations, the wave forms, and amplitude of oscillations do not change significantly, thereby fine-tuning periodic behavior of biological systems.

2.3. Cell Cycle Regulation. Next, we study the roles of microRNA in modulation of cell cycle by incorporating a microRNA, that is, *miR369-3*, into a detailed model.

Recently, experimental report by Vasudevan et al. revealed that microRNA can up-regulate gene expression in G0/G1 arrest during mammalian cell cycle [31]. In fact, they carried out a series of experiments by connecting a reporter mRNA to study *miR369-3* expression in different conditions. These results show that *miR369-3* promotes translation in G0/G1 arrest and suppresses translation in proliferating cells. It is equivalent to shortening the cell cycle process from the perspective of the entire cell cycle.

Many excellent theoretical models have been proposed to study the cell cycle [32–45]. The cell cycle consists of four phases: G1 phase, S phase, G2 phase, and M phase. Activation of the next stage is dependent on the normal completion of previous one. In [32], Novák and Tyson model the cell cycle as a two-steady-state process in which the two steady states are corresponding to G1 stage and S/G2/M stage, respectively. And the periodic switching between these two stable states is a result of the antagonism between CycB/Cdk1 and Cdh1/APC, which creates a G1 state with active Cdh1 and low CycB activity and an S/G2/M state with high CycB level and Cdh1 turned off.

Here we develop a computational model based on Tyson's cell cycle model [32] to investigate the posttranscriptional role of microRNA in the modulation of the mammalian cell cycle. By using computational target prediction algorithm miRBase [46], we screen out all the potential target genes of *miR369-3*; all of them are not directly in the pathway given in Tyson's model. However, among these potential target genes, *Lox* is found to inactivate CycD, an important gene in cell

cycle during the period G0/G1 [47]. Therefore, by introducing the new pathway *miR369-3-Lox-CycD* into the Tyson's model, miR369-3 is incorporated into the cell cycle model.

Now, let us study the regulatory effects of microRNA in the modified model with kinetic analysis. The dynamical equations are different from Tyson's model in two points: two new equations for Lox and *miR369-3* and one additional term in equation for CycD to show the inactivation of CycD by Lox. Refer to following equations:

$$\begin{aligned}
\frac{d\left[CycD\right]}{dt} &= \epsilon k_9\left[DRG\right] + V_6\left[CycD : Kip1\right] \\
&\quad + k_{24r}\left[CycD : Kip1\right] - k_{24}\left[CycD\right]\left[Kip1\right] \\
&\quad - k_{10}\left[CycD\right] - k_{40}\left[CycD\right]\left[Lox\right], \\
\frac{d\left[Lox\right]}{dt} &= k_{41} - k_{40}\left[CycD\right]\left[Lox\right] \\
&\quad - k_{42}\left[Lox\right]\left[miR369\text{-}3\right] - k_{43}\left[Lox\right], \\
\frac{d\left[miR369\text{-}3\right]}{dt} &= k_{44} - k_{42}\left[Lox\right]\left[miR369\text{-}3\right] \\
&\quad - k_{45}\left[miR369\text{-}3\right].
\end{aligned} \tag{13}$$

The new parameters k_{44}, k_{45}, k_{42}, k_{41}, k_{43}, and k_{40} are the production rate of *miR369-3*, linear degradation rate of *miR369-3*-association rate of *miR369-3* with *Lox*, production rate of *Lox*, linear degradation rate of *Lox*, and inactivation rate of *CycD* by *Lox*, respectively.

Next, numerical simulations are conducted. In the simulations, all parameter values are the same as those used in Tyson's model except the newly introduced ones. We run the simulation with a wide range of the new parameters, and similar results are derived. Here we show the results with the following set of parameters. Before the incorporation of the microRNA, the parameter values are taken as $k_{40} = 50$, $k_{41} = 200$, $k_{42} = 0$, $k_{43} = 5$, $k_{44} = 0$, and $k_{45} = 0$. After incorporating it, they are $k_{40} = 50$, $k_{41} = 200$, $k_{42} = 80$, $k_{43} = 5$, $k_{44} = 200$, and $k_{45} = 5$.

A simulation with and without microRNA regulation is presented in Figure 5, where concentrations of CycB and Cdh1 are shown. It can be seen that the incorporation of miR369-3 shortens the period of CycB with low expression representing the G1 phase and cells enter M phase earlier, which is in agreement with the experimental observations. We infer from the above results that miR369-3 exerts its regulation on cell cycle through the cell cycle period modulation. This founding is also similar to the results for the two motifs studied above.

The linear nature of the dependence of period on progressively varied k_{44} is shown in Figure 6, which further reflects the roles played by the microRNA-mediated regulation. The period of oscillations decreases slightly but almost linearly with k_{44}, meaning the fine-tuning of cell cycle by the miR369-3.

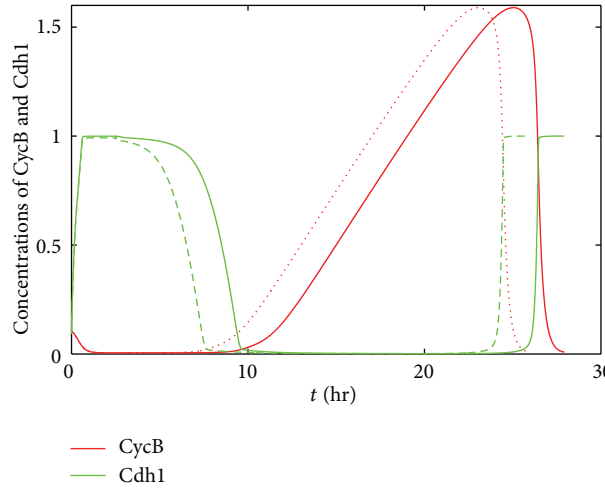


Figure 5: The modulation of cell cycle by *miR369-3*. The dashed and solid lines show the concentration of CycB and Cdh1 with and without microRNA regulation, respectively.

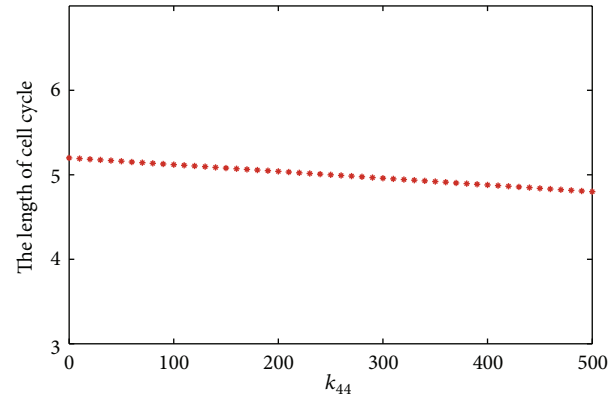


Figure 6: The effect of production rate of miR369-3 on the length of the cell cycle.

3. Discussion

MicroRNA-mediated regulation has gained recent attention. With the accumulation of our knowledge about microRNA, more and more regulatory mechanisms will be revealed.

Recent experimental works have implicated that microRNAs may play a fine-tuning role in cell cycle regulation. However, most of them are based on experimental works, and the regulatory mechanisms are less clear theoretically and need to be further investigated.

Beginning with two motifs, we find that microRNAs do not significantly change the type of oscillations, including their wave forms and amplitudes, while the period changes significantly. It seems that such a fine-tuning is general for both motifs. With those results for simplified models, we further incorporate a microRNA, *miR369-3*, into a classical cell cycle model and study the microRNA regulatory effects and mechanisms for cell cycle. We found that the *miR369-3* regulates cell cycle by slightly shortening the cell cycle period, thus accelerating the cell cycle, which is in agreement

with experimental observations. With the accumulatation of experimental results related to microRNA regulation in cell cycle, we will further incorporate more microRNAs into cell cycle and investigate the overall regulatory effects of microRNAs regulation in the future, which may give deep insights into microRNA regulatory mechanisms in cell cycle.

Finally, it is worth noting that microRNA-mediated regulation has gained recent attention, and computational studies have revealed various regulatory properties unique to microRNAs. These findings will be helpful for our understating the operating mechanisms and biological implications of microRNA-mediated regulation. They also have great potential for biotechnological, therapeutic applications, and synthetic biology.

Acknowledgment

This work is supported by NSFC no. 10832006 and no. 11172158.

References

[1] A. Chakrabarty, S. Tranguch, T. Daikoku, K. Jensen, H. Furneaux, and S. K. Dey, "MicroRNA regulation of cyclooxygenase-2 during embryo implantation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 38, pp. 15144–15149, 2007.

[2] E. Wienholds, W. P. Kloosterman, E. Miska et al., "Cell biology: microRNA expression in zebrafish embryonic development," *Science*, vol. 309, no. 5732, pp. 310–311, 2005.

[3] L. He, J. M. Thomson, M. T. Hemann et al., "A microRNA polycistron as a potential human oncogene," *Nature*, vol. 435, no. 7043, pp. 828–833, 2005.

[4] A. M. Krichevsky, K.-C. Sonntag, O. Isacson, and K. S. Kosik, "Specific MicroRNAs modulate embryonic stem cell-derived neurogenesis," *Stem Cells*, vol. 24, no. 4, pp. 857–864, 2006.

[5] J. Takamizawa, H. Konishi, K. Yanagisawa et al., "Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival," *Cancer Research*, vol. 64, no. 11, pp. 3753–3756, 2004.

[6] M. Z. Michael, S. M. O'Connor, N. G. Van Holst Pellekaan, G. P. Young, and R. J. James, "Reduced accumulation of specific microRNAs in colorectal neoplasia," *Molecular Cancer Research*, vol. 1, no. 12, pp. 882–891, 2003.

[7] H. Tagawa and M. Seto, "A microRNA cluster as a target of genomic amplification in malignant lymphoma [11]," *Leukemia*, vol. 19, no. 11, pp. 2013–2016, 2005.

[8] R. Schickel, B. Boyerinas, S.-M. Park, and M. E. Peter, "MicroRNAs: key players in the immune system, differentiation, tumorigenesis and cell death," *Oncogene*, vol. 27, no. 45, pp. 5959–5974, 2008.

[9] H.-W. Hwang and J. T. Mendell, "MicroRNAs in cell proliferation, cell death, and tumorigenesis," *British Journal of Cancer*, vol. 94, no. 6, pp. 776–780, 2006.

[10] H. R. Shcherbata, S. Hatfield, E. J. Ward, S. Reynolds, K. A. Fischer, and H. Ruohola-Baker, "The MicroRNA pathway plays a regulatory role in stem cell division," *Cell Cycle*, vol. 5, no. 2, pp. 172–175, 2006.

[11] P. P. Medina and F. J. Slack, "microRNAs and cancer: an overview," *Cell Cycle*, vol. 7, no. 16, pp. 2485–2492, 2008.

[12] M. Carleton, M. A. Cleary, and P. S. Linsley, "MicroRNAs and cell cycle regulation," *Cell Cycle*, vol. 6, no. 17, pp. 2127–2132, 2007.

[13] S. A. Georges, M. C. Biery, S.-Y. Kim et al., "Coordinated regulation of cell cycle transcripts by p53-inducible microRNAs, miR-192 and miR-215," *Cancer Research*, vol. 68, no. 24, pp. 10105–10112, 2008.

[14] S. D. Hatfield, H. R. Shcherbata, K. A. Fischer, K. Nakahara, R. W. Carthew, and H. Ruohola-Baker, "Stem cell division is regulated by the microRNA pathway," *Nature*, vol. 435, no. 7044, pp. 974–978, 2005.

[15] J. A. Pulikkan, V. Dengler, P. S. Peramangalam et al., "Cell-cycle regulator E2F1 and microRNA-223 comprise an autoregulatory negative feedback loop in acute myeloid leukemia," *Blood*, vol. 115, no. 9, pp. 1768–1778, 2010.

[16] Q. Liu, H. Fu, F. Sun et al., "miR-16 family induces cell cycle arrest by regulating multiple cell cycle genes," *Nucleic Acids Research*, vol. 36, no. 16, pp. 5391–5404, 2008.

[17] M. J. Bueno, I. P. De Castro, and M. Malumbres, "Control of cell proliferation pathways by microRNAs," *Cell Cycle*, vol. 7, no. 20, pp. 3143–3148, 2008.

[18] R. R. Chivukula and J. T. Mendell, "Circular reasoning: microRNAs and cell-cycle control," *Trends in Biochemical Sciences*, vol. 33, no. 10, pp. 474–481, 2008.

[19] Z. Yu, R. Baserga, L. Chen, C. Wang, M. P. Lisanti, and R. G. Pestell, "MicroRNA, cell cycle, and human breast cancer," *American Journal of Pathology*, vol. 176, no. 3, pp. 1058–1064, 2010.

[20] G. T. Bommer, I. Gerin, Y. Feng et al., "p53-mediated activation of miRNA34 candidate tumor-suppressor genes," *Current Biology*, vol. 17, no. 15, pp. 1298–1307, 2007.

[21] T.-C. Chang, D. Yu, Y.-S. Lee et al., "Widespread microRNA repression by Myc contributes to tumorigenesis," *Nature Genetics*, vol. 40, no. 1, pp. 43–50, 2008.

[22] D. C. Corney, A. Flesken-Nikitin, A. K. Godwin, W. Wang, and A. Y. Nikitin, "MicroRNA-34b and MicroRNA-34c are targets of p53 and cooperate in control of cell proliferation and adhesion-independent growth," *Cancer Research*, vol. 67, no. 18, pp. 8433–8438, 2007.

[23] L. He, X. He, L. P. Lim et al., "A microRNA component of the p53 tumour suppressor network," *Nature*, vol. 447, no. 7148, pp. 1130–1134, 2007.

[24] H. Hermeking, "The miR-34 family in cancer and apoptosis," *Cell Death and Differentiation*, vol. 17, no. 2, pp. 193–199, 2010.

[25] V. Tarasov, P. Jung, B. Verdoodt et al., "Differential regulation of microRNAs by p53 revealed by massively parallel sequencing: miR-34a is a p53 target that induces apoptosis and G 1-arrest," *Cell Cycle*, vol. 6, no. 13, pp. 1586–1593, 2007.

[26] H. Tazawa, N. Tsuchiya, M. Izumiya, and H. Nakagama, "Tumor-suppressive miR-34a induces senescence-like growth arrest through modulation of the E2F pathway in human colon cancer cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 39, pp. 15472–15477, 2007.

[27] H. I. Suzuki and K. Miyazono, "Emerging complexity of microRNA generation cascades," *Journal of Biochemistry*, vol. 149, no. 1, pp. 15–25, 2011.

[28] L. Santarpia, M. Nicoloso, and G. A. Calin, "MicroRNAs: a complex regulatory network drives the acquisition of malignant cell phenotype," *Endocrine-Related Cancer*, vol. 17, no. 1, pp. F51–F75, 2010.

[29] H. Hermeking, "p53 enters the microRNA world," *Cancer Cell*, vol. 12, no. 5, pp. 414–418, 2007.

[30] R. Guantes and J. F. Poyatos, "Dynamical principles of two-component genetic oscillators," *PLoS Computational Biology*, vol. 2, no. 3, pp. 188–197, 2006.

[31] S. Vasudevan, Y. Tong, and J. A. Steitz, "Switching from repression to activation: microRNAs can up-regulate translation," *Science*, vol. 318, no. 5858, pp. 1931–1934, 2007.

[32] B. Novák and J. J. Tyson, "A model for restriction point control of the mammalian cell cycle," *Journal of Theoretical Biology*, vol. 230, no. 4, pp. 563–579, 2004.

[33] Z. Qu, W. R. MacLellan, and J. N. Weiss, "Dynamics of the cell cycle: checkpoints, sizers, and timers," *Biophysical Journal*, vol. 85, no. 6, pp. 3600–3611, 2003.

[34] Z. Qu, J. N. Weiss, and W. R. MacLellan, "Regulation of the mammalian cell cycle: a model of the G1-to-S transition," *American Journal of Physiology*, vol. 284, no. 2, pp. C349–C364, 2003.

[35] Z. Qu, J. N. Weiss, and W. R. MacLellan, "Coordination of cell growth and cell division: a mathematical modeling study," *Journal of Cell Science*, vol. 117, no. 18, pp. 4199–4207, 2004.

[36] L. Yang, Z. Han, W. Robb MacLellan, J. N. Weiss, and Z. Qu, "Linking cell division to cell growth in a spatiotemporal model of the cell cycle," *Journal of Theoretical Biology*, vol. 241, no. 1, pp. 120–133, 2006.

[37] Z. Han, L. Yang, W. R. MacLellan, J. N. Weiss, and Z. Qu, "Hysteresis and cell cycle transitions: how crucial is it?" *Biophysical Journal*, vol. 88, no. 3, pp. 1626–1634, 2005.

[38] B. Pfeuty, T. David-Pfeuty, and K. Kaneko, "Underlying principles of cell fate determination during G1 phase of the mammalian cell cycle," *Cell Cycle*, vol. 7, no. 20, pp. 3246–3257, 2008.

[39] B. Pfeuty, "Strategic cell-cycle regulatory features that provide mammalian cells with tunable G1 length and reversible G1 arrest," *PLoS ONE*, vol. 7, no. 4, Article ID e35291, 2012.

[40] J. J. Tyson, A. Csikasz-Nagy, and B. Novak, "The dynamics of cell cycle regulation," *BioEssays*, vol. 24, no. 12, pp. 1095–1109, 2002.

[41] A. Csikász-Nagy, D. Battogtokh, K. C. Chen, B. Novák, and J. J. Tyson, "Analysis of a generic model of eukaryotic cell-cycle regulation," *Biophysical Journal*, vol. 90, no. 12, pp. 4361–4379, 2006.

[42] C. Gérard and A. Goldbeter, "Temporal self-organization of the cyclin/Cdk network driving the mammalian cell cycle," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 51, pp. 21643–21648, 2009.

[43] C. Gérard and A. Goldbeter, "From simple to complex patterns of oscillatory behavior in a model for the mammalian cell cycle containing multiple oscillatory circuits," *Chaos*, vol. 20, no. 4, Article ID 045109, 2010.

[44] C. Grard and A. Goldbeter, "A skeleton model for the network of cyclin-dependent kinases driving the mammalian cell cycle," *Interface Focus*, vol. 1, no. 1, pp. 24–35, 2011.

[45] A. Altinok, D. Gonze, and F. Levi, "An automaton model for the cell cycle," *Interface Focus*, vol. 1, no. 1, pp. 36–47, 2011.

[46] http://www.mirbase.org/.

[47] G. P. Pidgeon, M. Kandouz, A. Meram, and K. V. Honn, "Mechanisms controlling cell cycle arrest and induction of apoptosis after 12-lipoxygenase inhibition in prostate cancer cells," *Cancer Research*, vol. 62, no. 9, pp. 2721–2727, 2002.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 798743, 7 pages
http://dx.doi.org/10.1155/2013/798743

Research Article

Predicting the DPP-IV Inhibitory Activity pIC_{50} Based on Their Physicochemical Properties

Tianhong Gu,[1] Xiaoyan Yang,[2] Minjie Li,[2] Milin Wu,[2] Qiang Su,[1] Wencong Lu,[2] and Yuhui Zhang[3]

[1] *School of Materials Science and Engineering, Shanghai University, 149 Yan-Chang Road, Shanghai 200072, China*
[2] *Department of Chemistry, College of Sciences, Shanghai University, 99 Shang-Da Road, Shanghai 200444, China*
[3] *Department of Neurosurgery, Changhai Hospital, Second Military Medical University, 168 Chang-Hai Road, Shanghai 200433, China*

Correspondence should be addressed to Wencong Lu; wclu@shu.edu.cn and Yuhui Zhang; gong_chang2008@126.com

Received 29 March 2013; Revised 10 May 2013; Accepted 28 May 2013

Academic Editor: Yudong Cai

The second development program developed in this work was introduced to obtain physicochemical properties of DPP-IV inhibitors. Based on the computation of molecular descriptors, a two-stage feature selection method called mRMR-BFS (minimum redundancy maximum relevance-backward feature selection) was adopted. Then, the support vector regression (SVR) was used in the establishment of the model to map DPP-IV inhibitors to their corresponding inhibitory activity possible. The squared correlation coefficient for the training set of LOOCV and the test set are 0.815 and 0.884, respectively. An online server for predicting inhibitory activity pIC_{50} of the DPP-IV inhibitors as described in this paper has been given in the introduction.

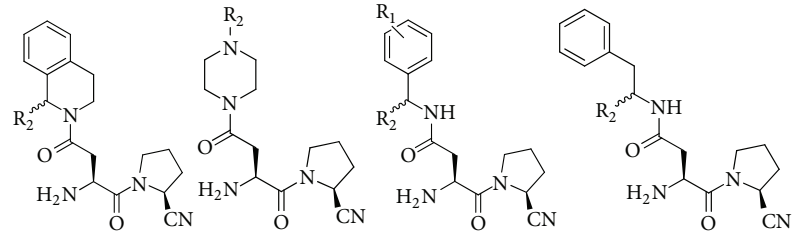
1. Introduction

The incretin hormones glucagon-like peptide-1 (GLP-1) and glucose-dependent insulinotropic polypeptide (GIP) are the endogenous peptides that stimulate glucose-dependent insulin secretion [1]. One of the important roles of dipeptidyl peptidase IV (DPP-IV) [2] is a rapid inactivation of the GLP-1 and GIP. Inhibition of DPP-4 increases the levels of endogenous intact circulating GLP-1 and GIP. Consequently, inhibitors of DPP-4 or gliptins have been recently regarded as a prospective approach for the treatment of type-2 diabetes mellitus.

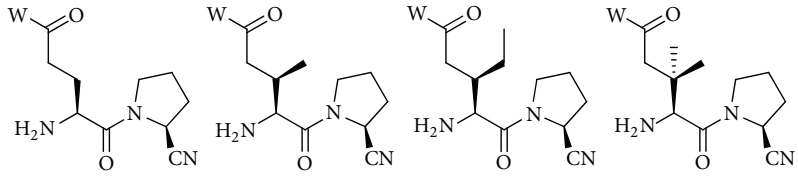
In recent years, multiple small-molecule DPP-4 inhibitors have been reported [3, 4]. The development of a structurally diverse collection of DPP-4 inhibitors is a hot research [5–8]. Computational and various mathematical approaches have been widely employed in the quantitative structure-activity relationship (QSAR) analysis [9–13]. Using statistical methods, QSAR analyses were carried out on a dataset of 47 pyrrolidine analogs acting as DPP-IV inhibitors by Paliwal et al. [14]. Murugesan et al. used the comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) to analyze the structural requirements of a DPP-IV active site [15]. Gao et al. developed a novel 3D-QSAR model to assist rational design of novel, potent, and selective pyrrolopyrimidine DPP-4 inhibitors [16]. Moreover, several efforts by using computational and mathematical approaches have been made in investigating small molecules of DPP-4 inhibitors. In our previous studies [17], we have attempted to use the quantum chemistry method [18] to optimize a series of DPP-IV inhibitors, and a 2D-QSAR model has been built, which can predict the inhibitory activity of small molecule with satisfying results. However, it is time consuming to calculate the molecular descriptors adopted in 2D-QSAR model.

In view of this, here we will try to devise an effective method to correctly recognize the possible activity prediction of small molecules based on physical and chemical properties of the compounds.

According to the general development trend [19, 20] and the recent research progress [21–31], the following procedures should be considered to establish a powerful statistical predictor for a biological system: (i) a valid benchmark dataset is constructed or selected to train and test the predictor;



(a) Glutamate cyanopyrrolidine analogues



(b) (2S)-cyanopyrrolidines with glutamic acid derivatives

Figure 1: Molecular structure of cyanopyrrolidine amides as DPP-IV inhibitors.

(ii) the samples are formulated with potent mathematical functions that are contributed to the prediction; (iii) a powerful algorithm is introduced or developed to operate the prediction; (iv) cross-validation tests are used to estimate the performance of the predictor; (v) a user-friendly online-server is established for the predictor that is accessible to the public. In this study, we attempt to describe how to deal with these steps for predicting the DPP-IV inhibitory activity pIC_{50} based on their physicochemical properties available via our program.

2. Materials and Methods

2.1. Data Preparation. The dataset used in the present work contains 48 pyrrolidine amides derivatives. In the current study, a diverse series of DPP-IV inhibitors with known IC_{50} values were collected from the papers [32, 33]. The detailed structures are documented in Supplementary Materials.(See Supplementary Material available at http://dx.doi.org/10.1155/2013/798743.) Figure 1 demonstrates the common structure of all of these analogues. All of the structures of compounds under investigation are based on the structure of Figure 1.

How to describe the molecules is an important problem in the establishment of the statistical model. In this study, the molecular descriptors for the 48 molecules were calculated by the second development software based on the calculator plugins, which is a product of ChemAxon [34]. ChemAxon is a company that provides chemical software development platforms and desktop applications for the biotechnology and pharmaceutical industries [35].

2.2. The Introduction of Procedure. Due to the use of Marvin Sketch graphic interface and JChem for Excel program, the calculations of small molecular descriptors are not very convenient. ChemAxon provides the calculation plugins of invoking function API, so our lab members have made a careful study and repeated experiments. The calculation

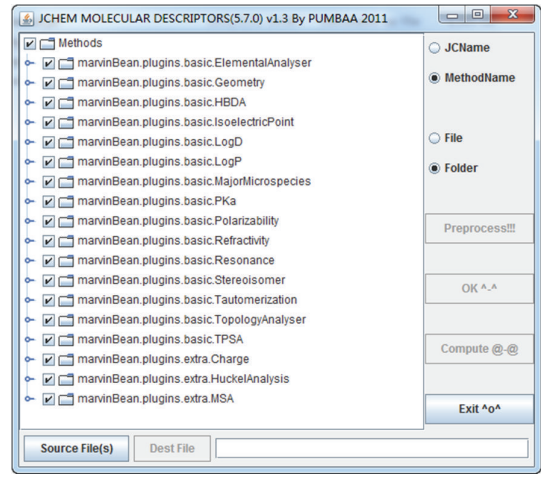


Figure 2: The program interface for the computation of molecular descriptors.

results are compared with the ones of Gaussian 09 [18], JChem for Excel [34], HyperChem 7.5 [20, 36], and Dragon [37] programs calculation. By invoking the Calculator Plugins and using the Java language, we successfully developed a convenient and available customized batch calculation program (second development software) for the small molecular descriptors.

This program contains a selection of tree box; the user can choose the visual way to the calculation of molecular descriptors (as shown in Figure 2, command-line version does not provide molecular descriptor selection). The molecule structures are constructed from Gauss View 5.0 package [38, 39] as MOL-format file. Command-line version of the program is operated commonly in Linux server, through the similar execution command as follows:

java-jar JChemCmd.jar Molecules Pathway Result.csv Method.xml

2.3. Model Validation

2.3.1. Dataset. The full dataset included training set (36 compounds) and test set (12 compounds). The whole samples were ranked by activity and were extracted every fourth sample for the generation of the test set.

2.3.2. Leave-One-Out Cross-Validation (LOOCV) and Predictive Validation. In this study, Leave-one-out cross-validation (LOOCV) [40, 41] was used to investigate the prediction quality of training set. In the cross-validation, each sample is used to test the model that is established by all of the other samples at the same time.

2.3.3. Fitting and Predictive Performances of Models. The fitting and predictive performances of model were measured by the squared correlation coefficient (q^2) and root mean square error (RMSE) for both the training set and the external test set. Here the performances of models can be estimated by q^2 and RMSE defined as follows, respectively:

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{\text{mean}})^2}, \qquad \text{RMSE} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N}}, \tag{1}$$

where y_i and $\hat{y}_i$ are the actual and predicted pIC_{50} values of i sample, respectively, and y_{mean} is the average pIC_{50} value of the entire samples. N is the numbers of the training set.

2.4. Methods. For the sake of the redundancy of some features, the selection of descriptors before establishing a suitable model is necessary. The selection of descriptors plays an important role in construction for the actual model. In this work, mRMR-BFS method (minimum redundancy maximum relevance-backward feature selection) [42, 43] was used for the selection of molecular descriptors. The support vector regression (SVR) model was established based on the feature selection results.

2.4.1. mRMR-BFS Algorithm. The mRMR (minimum-redundancy maximum-relevance) algorithm was introduced by Ding and Ping [44], which was used usually for feature selection. It sorts a feature based on score function which is maximum relevance to target and minimum redundancy to the already selected features. The score function is defined as follows:

$$\text{score}_j = I\left(f_j, c\right) - \frac{1}{m}\sum_{i=1}^{m} I\left(f_i, f_j\right), \tag{2}$$

where $f_j \subset S_n$, $f_i \subset S_m$, $S_m = S - S_n$, and S_m, S_n, and S are the feature sets. m and n are the feature numbers. The mutual information $I(x, y)$ is as follows:

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} dx\, dy, \tag{3}$$

where $p(x, y)$, $p(x)$, and $p(y)$ are the probabilistic density functions.

More details about mRMR algorithm can be found in [44, 45].

To gain an even better performance of predictor and feature selection, backward feature selection (BFS) based on the result of mRMR is also used in this study. The most important 50 variables were obtained from the mRMR procedure. We initialize the BFS-selected feature set S_s with all features in S:

$$S_s = \left\{f_1', f_2', \ldots, f_k'\right\} \quad (1 \le k \le 50). \tag{4}$$

With the mRMR-selected feature subset S_s, the next BFS-selected feature set can be gained by the following steps.

(1) Suppose that the candidate feature set is $S_C = S_S - f_k$. Then an SVR model based on each S_C is established and evaluated by LOOCV method.

(2) The feature f which gets the lowest RMSE is selected when removed from S_S.

(3) The feature f is removed from S_s forming the next BFS-selected feature set.

2.4.2. SVM (Support Vector Machine). Vapnik and his coworkers developed the SVM algorithm, which is a supervised machine-learning method that is used for classification and regression analysis. Owing to embodying the structural risk minimization principle, the SVM exhibits a better whole performance. The SVM is suitable for the problems which are involved in the small sample set. In this work, SVM was applied to regression. The details of the algorithm can be found in reference [46]. The algorithm was performed by using the software package Weka 3.6.7 [47, 48].

3. Results and Discussion

3.1. Selection of Features. Firstly, mRMR method was applied to rank the total 75 features according to their mRMR scores. Secondly, we used the backward feature selection (BFS) algorithm based on SVR to search for the feature combinations. As different machine learning methods will lead to different results, several robust machine learning methods like the nearest-neighbor algorithm (NNA), support vector machine (SVM based on RBF kernel function), and Adaboost were employed to find an optimal feature subset with leave-one-out cross-validation, respectively. As a result, we adopted the SVM as the prediction engine based on the LOOCV in this study.

Table 1 lists an optimal subset attained by employing the above two-stage feature selection method, mRMR-BFS. The six features in optimal subset can be clustered into three categories (based on the category of Calculator Plugins [49]): elemental analysis, geometry, topology, and others. The geometry and topology factor are more important in this work. The geometry and topology factor are related to the size of the molecule as it indicates that the size of cyanopyrrolidine amides derivatives plays a main role in the inhibitory activity.

Table 1: Symbols for molecular descriptors involved in the model.

Molecular descriptor	Type	Description
OComposition	Elemental analysis functions	O Composition
MaximalProjectionArea	Geometry	Calculates the maximal projection area
MinimalProjectionArea	Geometry	Calculates the minimal projection area
BasicpKa	pKa	Constant denoting basic pKa
RingBondCount	Topology	Ring bond count
AliphaticRingCount	Topology	Aliphatic ring count

3.2. Results of Computation. In this work, q^2_{train}, $q^2_{\text{train-CV}}$, and q^2_{test} were used to present the squared correlation coefficients for the training set, cross-validation set, and external test set, respectively. Also $\text{RMSE}_{\text{train}}$, $\text{RMSE}_{\text{train-CV}}$, and $\text{RMSE}_{\text{test}}$ were adopted to present the root mean square errors for the training set, cross-validation set, and external test set, respectively.

The final model was built by the SVR based on the Gaussian kernel function (RBF) with the parameters C, ε, and γ that are 2.0, 0.05, and 1.0, respectively. The Gaussian kernel function (RBF) is given as follows:

$$K(x, x_i) = \exp\left(-\gamma \|x - x_i\|^2\right). \quad (5)$$

The model based on the above parameters with original data is given as follows:

$$\text{pIC}_{50} = 2.10 * \left[\sum_{i\text{CSV}} \beta_i \exp\left(-\|x - x_i\|^2\right) + 0.207\right] + 6.60, \quad (6)$$

where β_i is the Lagrange coefficient of support vectors.

The experimental versus predicted pIC_{50} values based on the SVR model for the training set and test set are shown in Figure 3. As a result, the values of q^2_{train}, $q^2_{\text{train-CV}}$, and q^2_{test} were 0.953, 0.815, and 0.884, respectively. And the values of $\text{RMSE}_{\text{train}}$, $\text{RMSE}_{\text{train-CV}}$, and $\text{RMSE}_{\text{test}}$ were 0.123, 0.247, and 0.193, respectively. Figure 3 illustrates that the regression straight line is appropriate not only for the fitting pIC_{50} values of the training set but also for the predicted pIC_{50} values of the external test set. Table 2 shows the experimental and the calculated values over the training set and the test set. From Figure 3 and Table 2, it can be concluded that the predicted values are in good agreement with the experimental ones. Figure 4 illustrates the dispersion plot of the residuals for the training and test sets. The predicted values are randomly dispersed around the zero-value line in Figure 4. It means that the model is appropriate for the data.

3.3. Analysis of the New Method. The secondary development program developed in this work was used to establish a robust model with $q^2_{\text{train}} = 0.953$, $q^2_{\text{train-CV}} = 0.815$, and $q^2_{\text{test}} = 0.884$, respectively. In order to validate the generalization and reliability of the descriptors obtained by using our secondary development program, the same training and test sets were also constructed and optimized at the HF/6 – 31G* level of theory with the Gaussian program; 1262 descriptors were computed by HyperChem 7.5 program [20], JChem for Excel package [34], and the Dragon program [37]. And a robust and reliable model was obtained with $q^2_{\text{train}} = 0.969$, $q^2_{\text{train-CV}} = 0.868$, and $q^2_{\text{test}} = 0.891$, respectively. The statistical comparisons were summarized in Table 3.

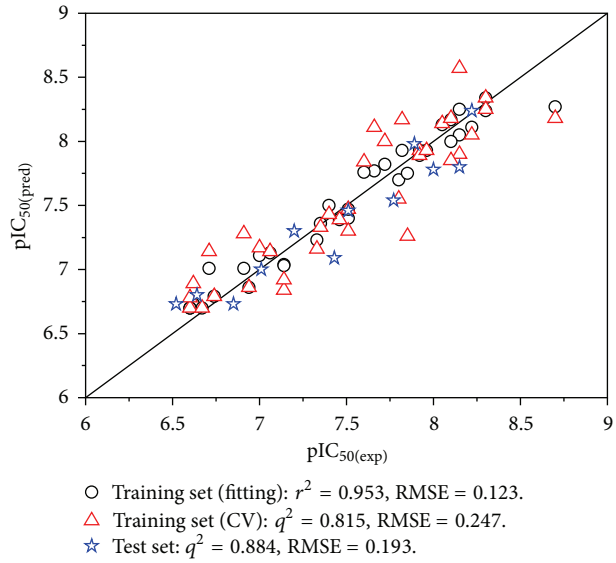


Figure 3: Predicted versus experimental pIC_{50} for the training (circles for fitting and triangle for CV, respectively) and test (stars) sets.

It is indicated that it takes less than 30 minutes for a molecule from the structure optimization to the computation of descriptors by using the second development program. In contrast, more than 36 hours were taken based on the Gaussian program. These results show that the computing speeds are greatly improved by using the secondary development program, while the statistical parameters of models are as good as those obtained with the Gaussian method. Therefore, the second development program is very helpful not only for saving the time of descriptor computation but also for providing the effective QSPR models online available in the future.

In a benchmark test, the support vector regression (SVR) was contrasted with the multiple linear regression (MLR) and the back propagation-artificial neural network (BP-ANN) on the $q^2_{\text{train-CV}}$. The statistical comparisons were shown in Table 4. From Table 4, SVR has a better generalization ability in our work.

Table 2: Experimental and predicted pIC_{50} for the training and test sets.

No.	$pIC_{50(exp)}$	$pIC_{50(Pred)}$	$pIC_{50(LOOCV)}$
1	7.00	7.11	7.17
2[T]	7.20	7.30	—
3	7.35	7.36	7.33
4	7.33	7.23	7.16
5[T]	7.01	7.00	—
6	7.14	7.04	6.92
7	7.14	7.03	6.84
8	6.71	7.01	7.14
9[T]	6.64	6.80	—
10	7.06	7.13	7.14
11	6.91	7.01	7.28
12	6.62	6.73	6.89
13	6.60	6.70	6.78
14[T]	6.85	6.73	—
15	6.67	6.70	6.70
16	6.60	6.70	6.70
17	6.94	6.86	6.86
18	6.74	6.79	6.79
19[T]	6.52	6.73	—
20	8.70	8.27	8.18
21	8.30	8.34	8.34
22	7.46	7.39	7.39
23	7.40	7.50	7.43
24[T]	8.22	8.24	—
25	8.15	8.25	8.57
26	8.30	8.24	8.25
27	8.05	8.13	8.14
28	8.22	8.11	8.05
29	8.15	8.05	7.90
30[T]	8.00	7.78	—
31	7.66	7.77	8.11
32[T]	8.15	7.80	—
33	7.82	7.93	8.17
34[T]	7.77	7.54	—
35[T]	7.51	7.46	—
36	8.10	8.00	7.85
37	7.72	7.82	8.00
38[T]	7.43	7.09	—
39	7.96	7.93	7.93
40	8.10	8.17	8.18
41	7.51	7.40	7.30
42	7.92	7.89	7.89
43	7.51	7.47	7.47
44	7.92	7.93	7.93
45	7.80	7.70	7.55
46	7.60	7.76	7.84
47	7.85	7.75	7.26
48[T]	7.89	7.98	—

[T]indicates the test samples.

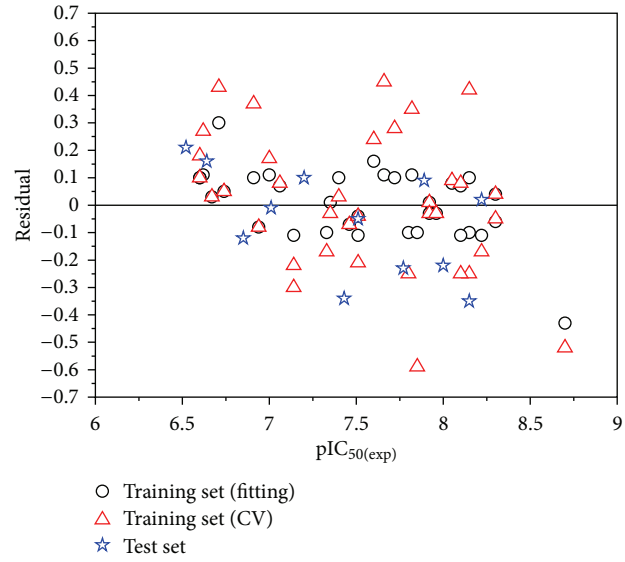


Figure 4: Dispersion plot of the residuals for the training and test sets.

Table 3: Comparative statistical parameters obtained by the secondary development program and the Gaussian program concerning the same compounds.

Program	q^2_{train}	$q^2_{train\text{-}CV}$	q^2_{test}
The secondary development program developed in this work	0.953	0.815	0.884
Gaussian, HyperChem 7.5, JChem for Excel package, Dragon	0.969	0.868	0.891

Table 4: $q^2_{train\text{-}CV}$ of different methods.

Method	SVR	BP-ANN	MLR
$q^2_{train\text{-}CV}$	0.815	0.761	0.721

3.4. The Online Web Server. Since user-friendly and publicly accessible online servers represent the trend for developing more useful models or predictors, we established a web server for predicting the DPP-IV inhibitory activity pIC_{50} at http://chemdata.shu.edu.cn:8080/QSARPrediction/index.jsp.

The web server allows users to upload the MOL-format file of a molecule, and the server will return the result of prediction according to the model of our mRMR-BFS-SVR method. In this course, the Calculator Plugins [49] of ChemAxon was invoked in the background program. The server developed has the most outstanding characteristic that users need to do nothing except for uploading the file of the unknown small molecule. Then they can get the predicted result after waiting for some time. It is a remarkable advance compared to our previous work [17, 20, 36].

4. Conclusions

In this paper, the secondary development program was proposed to bring an efficient and fast calculation means for molecular descriptors. The mRMR-BFS was adopted in the procedure of feature selection. The SVR was used to construct the model to map DPP-IV inhibitors to their corresponding inhibitory activity. The q^2_{train}, $q^2_{\text{train-CV}}$, and q^2_{test} of the model are 0.953, 0.815, and 0.884, respectively. These results are as good as those obtained with the Gaussian method. The web server, which provides a quick approach to predict the DPP-IV inhibitory activities pIC_{50} of unknown small molecules based on their MOL-format files, was established by using our secondary development program at http://chemdata.shu.edu.cn:8080/QSARPrediction/index.jsp. A user-friendly and rapid approach whose accuracy is approximate with the Gaussian method is proposed in this work.

Acknowledgments

This study was supported by the National Science Foundation of China (20973108, 20902056), the Shanghai Education Committee Project (11ZZ83), and the Leading Academic Discipline Project of Shanghai Municipal Education Commission, China (J50101). The authors also acknowledge ChemAxon for their excellent products.

References

[1] M. H. Kim and M. K. Lee, "The incretins and pancreatic beta-cells: use of glucagon-like peptide-1 and glucose-dependent insulinotropic polypeptide to cure type 2 diabetes mellitus," *Korean Diabetes Journal*, vol. 34, no. 1, pp. 2–9, 2010.

[2] A. Sarashina, S. Sesoko, M. Nakashima et al., "Linagliptin, a dipeptidyl peptidase-4 inhibitor in development for the treatment of type 2 diabetes mellitus: a phase I, randomized, double-blind, placebo-controlled trial of single and multiple escalating doses in healthy adult male japanese subjects," *Clinical Therapeutics*, vol. 32, no. 6, pp. 1188–1204, 2010.

[3] K. Augustyns, P. Van der Veken, and A. Haemers, "Inhibitors of proline-specific dipeptidyl peptidases: DPP IV inhibitors as a novel approach for the treatment of type 2 diabetes," *Expert Opinion on Therapeutic Patents*, vol. 15, no. 10, pp. 1387–1407, 2005.

[4] A. E. Weber, "Dipeptidyl peptidase IV inhibitors for the treatment of diabetes," *Journal of Medicinal Chemistry*, vol. 47, no. 17, pp. 4135–4141, 2004.

[5] S. D. Edmondson, A. Mastracchio, R. J. Mathvink et al., "(2S, 3S)-3-amino-4-(3,3-difluoropyrrolidin-1-yl)-N,N-dimethyl-4-oxo-2-(4-[1,2,4]triazolo[1,5-a]-pyridin-6-ylphenyl)butanamide: a selective α-amino amide dipeptidyl peptidase IV inhibitor for the treatment of type 2 diabetes," *Journal of Medicinal Chemistry*, vol. 49, no. 12, pp. 3614–3627, 2006.

[6] J. L. Duffy, B. A. Kirk, L. Wang et al., "4-Aminophenylalanine and 4-aminocyclohexylalanine derivatives as potent, selective, and orally bioavailable inhibitors of dipeptidyl peptidase IV," *Bioorganic and Medicinal Chemistry Letters*, vol. 17, no. 10, pp. 2879–2885, 2007.

[7] J. Xu, L. Wei, R. J. Mathvink et al., "Discovery of potent, selective, and orally bioavailable oxadiazole-based dipeptidyl peptidase IV inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 16, no. 20, pp. 5373–5377, 2006.

[8] J. Xu, L. Wei, R. Mathvink et al., "Discovery of potent, selective, and orally bioavailable pyridone-based dipeptidyl peptidase-4 inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 16, no. 5, pp. 1346–1349, 2006.

[9] T. S. Garcia and K. M. Honório, "Two-dimensional quantitative structure-activity relationship studies on bioactive ligands of peroxisome proliferator-activated receptor δ," *Journal of the Brazilian Chemical Society*, vol. 22, no. 1, pp. 65–72, 2011.

[10] G. C. García, I. Luque Ruiz, and M. Á. Gómez-Nieto, "Analysis and study of molecule data sets using snowflake diagrams of weighted maximum common subgraph trees," *Journal of Chemical Information and Modeling*, vol. 51, no. 6, pp. 1216–1232, 2011.

[11] D. Jana, A. K. Halder, N. Adhikari, M. K. Maiti, C. Mondal, and T. Jha, "Chemometric modeling and pharmacophore mapping in coronary heart disease: 2-arylbenzoxazoles as cholesteryl ester transfer protein inhibitors," *MedChemComm*, vol. 2, no. 9, pp. 840–852, 2011.

[12] V. Kovalishyn, V. Tanchuk, L. Charochkina, I. Semenuta, and V. Prokopenko, "Predictive QSAR modeling of phosphodiesterase 4 inhibitors," *Journal of Molecular Graphics and Modelling*, vol. 32, pp. 32–38, 2012.

[13] B. Niu, Q. Su, X. C. Yuan, W. Lu, and J. Ding, "QSAR study on 5-lipoxygenase inhibitors based on support vector machine," *Medicinal Chemistry*, vol. 8, no. 6, pp. 1108–1116, 2012.

[14] S. Paliwal, D. Seth, D. Yadav, R. Yadav, and S. Paliwal, "Development of a robust QSAR model to predict the affinity of pyrrolidine analogs for dipeptidyl peptidase IV (DPP-IV)," *Journal of Enzyme Inhibition and Medicinal Chemistry*, vol. 26, no. 1, pp. 129–140, 2011.

[15] V. Murugesan, N. Sethi, Y. S. Prabhakar, and S. B. Katti, "CoMFA and CoMSIA of diverse pyrrolidine analogues as dipeptidyl peptidase IV inhibitors: active site requirements," *Molecular Diversity*, vol. 15, no. 2, pp. 457–466, 2011.

[16] Y. D. Gao, D. Feng, R. P. Sheridan et al., "Modeling assisted rational design of novel, potent, and selective pyrrolopyrimidine DPP-4 inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 17, no. 14, pp. 3877–3879, 2007.

[17] X. Y. Yang, M. J. Li, Q. Su, M. Wu, T. Gu, and W. Lu, "QSAR studies on pyrrolidine amides derivatives as DPP-IV inhibitors for type 2 diabetes," *Medicinal Chemistry Research*, 2013.

[18] S. Peng, Z. Jian-Wei, Z. Peng, and X. Lin, "QSPR modeling of bioconcentration factor of nonionic compounds using Gaussian processes and theoretical descriptors derived from electrostatic potentials on molecular surface," *Chemosphere*, vol. 83, no. 8, pp. 1045–1052, 2011.

[19] T. Gu, W. Lu, X. Bao, and N. Chen, "Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors," *Solid State Sciences*, vol. 8, no. 2, pp. 129–136, 2006.

[20] J. Zhu, W. Lu, L. Liu, T. Gu, and B. Niu, "Classification of Src kinase inhibitors based on support vector machine," *QSAR and Combinatorial Science*, vol. 28, no. 6-7, pp. 719–727, 2009.

[21] V. Kovalishyn, J. Aires-de-Sousa, C. Ventura, R. Elvas Leitão, and F. Martins, "QSAR modeling of antitubercular activity of diverse organic compounds," *Chemometrics and Intelligent Laboratory Systems*, vol. 107, no. 1, pp. 69–74, 2011.

[22] L. Xing, R. Goulet, and K. Johnson, "Statistical analysis and compound selection of combinatorial libraries for soluble epoxide hydrolase," *Journal of Chemical Information and Modeling*, vol. 51, no. 7, pp. 1582–1592, 2011.

[23] S. Kar, O. Deeb, and K. Roy, "Development of classification and regression based QSAR models to predict rodent carcinogenic potency using oral slope factor," *Ecotoxicology and Environmental Safety*, vol. 82, pp. 85–95, 2012.

[24] B. Niu, X. C. Yuan, P. Roeper et al., "HIV-1 protease cleavage site prediction based on two-stage feature selection method," *Protein and Peptide Letters*, vol. 20, no. 3, pp. 290–298, 2013.

[25] B. Niu, Y. D. Cai, W. C. Lu, G. Z. Li, and K. C. Chou, "Predicting protein structural class with AdaBoost Learner," *Protein and Peptide Letters*, vol. 13, no. 5, pp. 489–492, 2006.

[26] B. Niu, Y. H. Jin, K. Y. Feng et al., "Predicting membrane protein types with bagging learner," *Protein and Peptide Letters*, vol. 15, no. 6, pp. 590–594, 2008.

[27] B. Niu, Y. H. Jin, K. Y. Feng, W. C. Lu, Y. D. Cai, and G. Z. Li, "Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins," *Molecular Diversity*, vol. 12, no. 1, pp. 41–45, 2008.

[28] B. Niu, Y. Jin, L. Lu et al., "Prediction of interaction between small molecule and enzyme using AdaBoost," *Molecular Diversity*, vol. 13, no. 3, pp. 313–320, 2009.

[29] B. Niu, Y. Jin, W. Lu, and G. Li, "Predicting toxic action mechanisms of phenols using AdaBoost Learner," *Chemometrics and Intelligent Laboratory Systems*, vol. 96, no. 1, pp. 43–48, 2009.

[30] B. Niu, L. Lu, L. Liu et al., "HIV-1 protease cleavage site prediction based on amino acid property," *Journal of Computational Chemistry*, vol. 30, no. 1, pp. 33–39, 2009.

[31] Q. Su, W. C. Lu, B. Niu, X. Liu, and T. H. Gu, "Classification of the toxicity of some organic compounds to tadpoles (Rana Temporaria) through integrating multiple classifiers," *Molecular Informatics*, vol. 30, no. 8, pp. 672–675, 2011.

[32] I. L. Lu, S. J. Lee, H. Tsu et al., "Glutamic acid analogues as potent dipeptidyl peptidase IV and 8 inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 15, no. 13, pp. 3271–3275, 2005.

[33] T. Y. Tsai, T. Hsu, C. T. Chen et al., "Rational design and synthesis of potent and long-lasting glutamic acid-based dipeptidyl peptidase IV inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 19, no. 7, pp. 1908–1912, 2009.

[34] L. Weber, "JChem base—chemAxon," *Chemistry World*, vol. 5, no. 10, pp. 65–66, 2008.

[35] 2013, http://www.chemaxon.com/.

[36] S. S. Yang, W. C. Lu, T. H. Gu, L. M. Yan, and G. Z. Li, "QSPR study of n-octanol/water partition coefficient of some aromatic compounds using support vector regression," *QSAR and Combinatorial Science*, vol. 28, no. 2, pp. 175–182, 2009.

[37] T. Todeschini, "Dragon 5.0: software for molecular descriptors," in *Milano Chemometrics and QSAR Research Group*, University of Milano-Bicocca, Milan, Italy, 2004.

[38] V. Mukherjee, K. Singh, N. P. Singh, and R. A. Yadav, "Quantum chemical determination of molecular geometries and interpretation of FTIR and Raman spectra for 2,4,5- and 3,4,5-trifluoro-benzonitriles," *Spectrochimica Acta A*, vol. 71, no. 4, pp. 1571–1580, 2008.

[39] Y. Chen, Z. Yi, S. J. Chen, J. S. Luo, Y. G. Yi, and Y. J. Tang, "Study of density functional theory for surface-enhanced raman spectra of p-aminothiophenol," *Spectroscopy and Spectral Analysis*, vol. 31, no. 11, pp. 2952–2955, 2011.

[40] T. Zhang, "A leave-one-out cross validation bound for kernel methods with applications in learning," *Computational Learning Theory Proceedings*, vol. 2111, pp. 427–443, 2001.

[41] J. Yuan, Y. M. Li, C. L. Liu, and X. F. Zha, "Leave-one-out cross-validation based model selection for manifold regularization," in *Advances in Neural Networks*, vol. 6063 of *Lecture Notes in Computer Science*, pp. 457–464, 2010.

[42] M. Kompany-Zareh, "An improved QSPR study of the toxicity of aliphatic carboxylic acids using genetic algorithm," *Medicinal Chemistry Research*, vol. 18, no. 2, pp. 143–157, 2009.

[43] M. Goodarzi, B. Dejaegher, and Y. Vander Heyden, "Feature selection methods in QSAR studies," *Journal of Aoac International*, vol. 95, no. 3, pp. 636–651, 2012.

[44] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," in *Proceedings of the IEEE Bioinformatics Conference*, pp. 185–205, August 2003.

[45] Z. He, J. Zhang, X. H. Shi et al., "Predicting drug-target interaction networks based on functional groups and biological features," *PLoS ONE*, vol. 5, no. 3, Article ID e9603, 2010.

[46] B. Üstün, W. J. Melssen, and L. M. C. Buydens, "Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel," *Chemometrics and Intelligent Laboratory Systems*, vol. 81, no. 1, pp. 29–40, 2006.

[47] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using Weka," *Bioinformatics*, vol. 20, no. 15, pp. 2479–2481, 2004.

[48] L. Chen, L. Lu, K. Feng et al., "Multiple classifier integration for the prediction of protein structural classes," *Journal of Computational Chemistry*, vol. 30, no. 14, pp. 2248–2254, 2009.

[49] 2013, http://www.chemaxon.com/products/calculator-plugins/.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 630803, 11 pages
http://dx.doi.org/10.1155/2013/630803

Research Article

Dynamic Actin Gene Family Evolution in Primates

Liucun Zhu,[1] Ying Zhang,[2] Yijun Hu,[3] Tieqiao Wen,[1,3] and Qiang Wang[4]

[1] *Institute of System Biology, Shanghai University, Shanghai 200444, China*
[2] *Yangzhou Breeding Biological Agriculture Technology Co. Ltd., Yangzhou 225200, China*
[3] *School of Life Sciences, Shanghai University, Shanghai 200444, China*
[4] *State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing 210093, China*

Correspondence should be addressed to Qiang Wang; wangq@nju.edu.cn

Received 10 April 2013; Revised 17 May 2013; Accepted 18 May 2013

Academic Editor: Tao Huang

Actin is one of the most highly conserved proteins and plays crucial roles in many vital cellular functions. In most eukaryotes, it is encoded by a multigene family. Although the actin gene family has been studied a lot, few investigators focus on the comparison of actin gene family in relative species. Here, the purpose of our study is to systematically investigate characteristics and evolutionary pattern of actin gene family in primates. We identified 233 actin genes in human, chimpanzee, gorilla, orangutan, gibbon, rhesus monkey, and marmoset genomes. Phylogenetic analysis showed that actin genes in the seven species could be divided into two major types of clades: orthologous group versus complex group. Codon usages and gene expression patterns of actin gene copies were highly consistent among the groups because of basic functions needed by the organisms, but much diverged within species due to functional diversification. Besides, many great potential pseudogenes were found with incomplete open reading frames due to frameshifts or early stop codons. These results implied that actin gene family in primates went through "birth and death" model of evolution process. Under this model, actin genes experienced strong negative selection and increased the functional complexity by reproducing themselves.

1. Introduction

Actin is an abundant and highly conserved protein that is found in all eukaryotic cells [1]. It is also a major component of total amount of proteins in various kinds of cells [2, 3] and plays an essential role in a variety of important cellular processes including vesicle and organelle movements [4, 5], cell motility [6], cell division [7] and cytokinesis [8], muscle contraction [9], and the establishment and maintenance of cell junctions and cell shape [10]. Except for conventional actin, eukaryotic cells also contain actin-like (ALPs) and actin-related proteins (ARPs), which have well-characterized roles in cytoskeletal functions [11, 12]. Actins, ALPs, and ARPs, comprising a large family of homologous proteins, share the same structural architecture, known as the "actin fold" [13]. These three kinds of proteins are encoded by a multigene family in all animals, plants, and many protozoans examined to date, making up actin superfamily [14–16], which is called actin gene family in this work.

Compared to its functional studies, the organization and evolution of actin gene family are not discussed extensively. Comparisons of nucleotide sequences from the protein coding regions and exon-intron arrangements of related genes provide a means of tracing their evolution pathways [17, 18]. Before the advent of the era of large-scale sequencing, actin gene family has been investigated in many organisms [19–24]. Those results indicate that actin gene family is highly conserved, and the number of actin genes among these organisms is variable. With the development of sequencing technology, recent studies of dynamic actin gene evolution in lower organisms like algae reveal distinct phylogenetic structures and evolution histories [25, 26]. In most of algae, actin genes morphologically cluster with each other on the phylogenetic tree among different algal lineages [25]. In each algal clade of actin tree, at least two subclades are found, in which one contains highly conserved sequences, whereas the other one has very diverged actin isoforms. On the other hand, phylogenetic analysis in dinoflagellates exhibits at least

three types of clusters [26]. The first type contains recently duplicated copies within each species, and the other two types form clades including sequences from different species, in which one type contains very similar copies and the other one has divergent copies across species.

Although there are many studies for this family, no systematic research has been made in primates. Consequently, the purpose of this study is to investigate characteristics and evolutionary pattern of all actin genes in primates. We first identified 233 actin genes including actin-like and actin-related gene plus 337 pseudogenes residing in human, chimpanzee, gorilla, orangutan, gibbon, rhesus monkey, and marmoset genomes. And then, we analyzed and compared their phylogenetic distribution, codon usage, and expression pattern between orthologs and paralogs. Our results indicated that actin genes in primates extraordinarily diverged among paralogs, but were highly conserved across orthologs. In this case, we suggested that actin gene family experienced a duplication followed by mutation process, according with birth and death model of evolution.

2. Material and Method

2.1. Identification of Actin Genes. The genome and protein sequences of human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus abelii*), gibbon (*Nomascus leucogenys*), rhesus monkey (*Macaca mulatta*), and marmoset (*Callithrix jacchus*) were downloaded from Ensembl ftp site (ftp://ftp.ensembl.org/pub/release-69/fasta/). We identified actin genes as follows: first of all, we downloaded protein sequences which were limited to genes with actin domain (Pfam: PF00022) from Biomart [27] (website: http://asia.ensembl.org/biomart/martview). Then, the amino acid sequences of all known actin genes were adopted as queries in local BLASTP (Basic Local Alignment Search Tool) searches for potential homologs in seven genomes with 1e-10 as the threshold expectation value. Based on the BLASTP results, all genes were verified with the conserved actin domain by searching in corresponding Conserved Domain Database (CDD) online [28] (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Thus, the entire actin genes were identified in the seven genomes. The actin gene, in which the amino acid length of actin domain was less than 160aa, was excluded for further analysis. The associated gene name or ensembl protein id, in which this copy was not given associated gene names was used for each actin gene. The suffixal letters "Hsa," "Ptr," "Ggo," "Ppy," "Nle," "Mmu," and "Cja" of the gene names represented the actin copies from human, chimpanzee, gorilla, orangutan, gibbon, rhesus monkey, and marmoset genome, respectively.

2.2. Sequence Alignment and Phylogenetic Analysis. The amino acid sequence of actin domain was aligned by MEGA4 [29] in ClustalW with default options [30]. The resulting amino acid sequence alignments were then used to guide the alignments of nucleotide coding sequences (CDSs). Phylogenetic trees were constructed based on the bootstrap neighbor-joining method with a Jukes-Cantor model by MEGA4. The stability of internal nodes was assessed by bootstrap analysis with 1000 replicates.

Based on the nucleotide diversity/divergence between homologs within major clades (<30%) and bootstrap values, the phylogenetic tree of all actin genes from the seven genomes can be divided into two major types (Figure 1(a)) and several single genes. The first type, which contained a single copy of actin genes from each of the seven species, was designated as orthologous group, shown in Figure 1(d). The bootstrap value of every clade should be more than 80, which was considered as a credible clade. On the other hand, complex group exhibited multi-copy number or none of actin genes from one of the seven species mixed in the clade, as illustrated in Figures 1(b) and 1(c).

Nucleotide divergence among homologs was estimated by divergence (d) with the Jukes and Cantor correction [31]. The number of nonsynonymous substitutions per nonsynonymous site and the number of synonymous substitution per synonymous site were denoted by K_a and K_s, respectively. The K_a and K_s were calculated based on Nei and Gojobori [32]. A K_a/K_s ratio greater than 1 suggested positive selection, and the ratio less than 1 suggested negative selection generally.

2.3. Identification of Pseudogenes. To identify actin pseudogenes, all of the nucleotide sequences of actin domains from seven species were employed to search in all the genomes used in this work (BLASTN). After excluding the hit sequences which were identified as actin genes above, a PERL script was written to remove the length of hit sequences which was shorter than 450 bp. The rest of hit positions on the chromosomes were considered as locations of actin pseudogenes.

2.4. Codon Usage Estimates Using Relative Synonymous Codon Usage (RSCU). The codon usage analysis for every actin gene was estimated by relative synonymous codon usage (RSCU) value. The RSCU value of a codon [33] is calculated by dividing the observed codon usage by that expected when all codons for the same amino acid are used equally. Due to an amino acid coded by a single codon (such as ATG: methionine and TGG: tryptophan), these two codons and stop codons were not included in an RSCU analysis. RSCU values are not affected by sequence length and amino acid frequency since these factors are eliminated during the computation. The RSCU values <1, 1, and >1 indicated that the codons used less than average, at average level (no bias), and more than average [34–36]. For actin domain nucleotide sequences of each actin gene in this study, RSCU values were calculated for the 59 relevant codons by a PERL script. The variation of RSCU value for each codon from actin genes within every genome or complex groups/orthologous groups (see Section 2.3) was calculated to estimate codon usage pattern.

2.5. Actin Gene Expression Analysis. The array datasets of transcription profiling of human and chimpanzee were

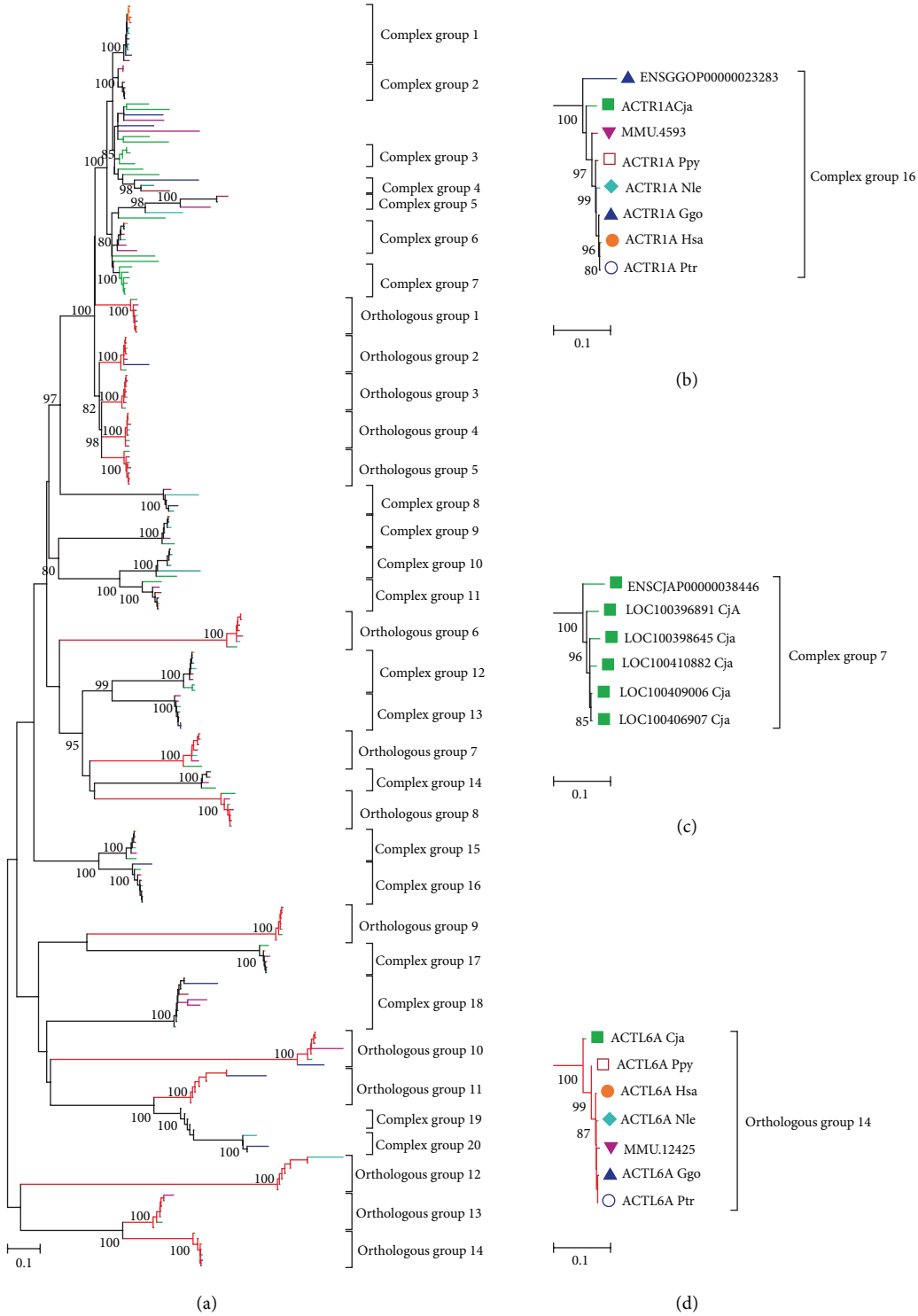


Figure 1: Schematic for the whole phylogenetic tree of actin genes in the seven species using nucleotide alignment of actin domain (a). *Orange* (human), *Dark Cyan* (chimpanzee), *Blue* (gorilla), *Wine* (orangutan), *Cyan* (gibbon), *Magenta* (rhesus monkey), and *Green* (marmoset). The *Red* clades represent the orthologous groups. The representative actin domain phylogenies for (b) clades in the complex group displayed multicopies from the same species, (c) clades in the complex group which lost copies from some species, and (d) clades in the orthologous group. The tree was built based on the neighboring-joining method with a Jukes-Cantor model.

downloaded from ARRAYEXPRESS database at the European Bioinformatics Institute (EBI, website: http://www.ebi.ac.uk/arrayexpress/). The accession number of the experiment was E-AFMX-11 [37], processing on the platform of "Affymetrix GeneChip Human Genome U133 plus 2.0 [HG-U133_Plus_2]." An R script was developed to extract the information of array probe, values of expressed level, and P values from the array data. The coefficient of variation (CV; SD/mean) of the expression values for actin genes was calculated to estimate expression pattern.

3. Result

3.1. Phylogeny and Classification of Actin Genes. According to the characteristic domain of actin gene (PF00022) reported previously, we identified 233 actin genes in seven genomes using BLASTP search and CDD analysis (see Supplementary Material Table S1, see Section 2.1). Based on the alignment results for actin domain sequences of all the actin genes found in the seven species, we constructed a phylogenetic tree using the Bootstrap neighbor-joining (NJ) method with a Jukes-Cantor model by MEGA v4.0 [29] (Figure 1). According to the nucleotide diversity/divergence between homologs (<30%, see Table S2) and bootstrap values (>80), we split the tree into 34 groups to investigate evolution of actin genes in detail. Under these criteria, 14 genes could not be included in any group.

The whole phylogenetic tree and representative major clades were shown in Figure S1 and Figure 1, exhibiting two dominant types of phylogenetic structures. The first type of major clades that consist of seven copies of actin genes from all species was designated as orthologous group, shown in Figure 1(d). On the other side, each clade of complex group contained more or less than one copy of actin genes from one of the seven species, as illustrated in Figures 1(b) and 1(c). For example, complex group 16 contained two copies of actin genes from gorilla genome (Figure 1(b)), while complex group 7 just included six copies from marmoset, none from other six species (Figure 1(c)).

Following the definition of two kinds of clades, there were 14 major clades (41.2% of the total clades) found in the orthologous groups, comprising 98 (42.0%) of the total actin genes. The average nucleotide divergence (d) of the actin domain sequences within the orthologous groups was 2.75% (Table 1). Twenty clades of complex groups were identified on the phylogenetic tree, which had 121 actin genes in the clades (Table 2). The average d value of the actin domain sequences in the complex groups was 5.66%, which was significantly greater than that in the orthologous groups ($P = 0.028$, using two-tailed t-test). In addition, the total copy number for each species in all the complex groups was from 11 to 21 (Table 2). The large d value and variable copy number of complex groups implied that these actin genes diverged across species.

3.2. Nonsynonymous to Synonymous Substitution. According to the multiple alignments of all actin genes from seven species, we calculated the average nonsynonymous substitutions (K_a) and synonymous substitutions (K_s) for actin

Table 1: Average nucleotide diversity, nonsynonymous, and synonymous substitutions of actin domain from actin genes in each group and species.

Name	d	K_a	K_s	K_a/K_s
Orthologous group 1	0.0234	0.0058	0.0819	0.0711
Orthologous group 2	0.0374	0.0188	0.1002	0.1874
Orthologous group 3	0.0149	0.0000	0.0642	0.0000
Orthologous group 4	0.0163	0.0010	0.0674	0.0153
Orthologous group 5	0.0204	0.0000	0.0902	0.0000
Orthologous group 6	0.0362	0.0156	0.1050	0.1488
Orthologous group 7	0.0510	0.0262	0.1268	0.2064
Orthologous group 8	0.0366	0.0184	0.0899	0.2047
Orthologous group 9	0.0107	0.0005	0.0483	0.0107
Orthologous group 10	0.0777	0.0606	0.1349	0.4487
Orthologous group 11	0.0094	0.0000	0.0416	0.0000
Orthologous group 12	0.0186	0.0006	0.0831	0.0069
Orthologous group 13	0.0207	0.0007	0.0893	0.0077
Orthologous group 14	0.0125	0.0005	0.0563	0.0089
Average for orthologous groups	**0.0276**	**0.0106**	**0.0842**	**0.0940**
Complex group 1	0.0208	0.0136	0.0445	0.3067
Complex group 2	0.0183	0.0011	0.0755	0.0146
Complex group 3	0.0350	0.0364	0.0305	1.1935
Complex group 4	0.2291	0.2027	0.3317	0.6112
Complex group 5	0.1455	0.1362	0.1770	0.7697
Complex group 6	0.0494	0.0183	0.1570	0.1164
Complex group 7	0.0364	0.0236	0.0793	0.2973
Complex group 8	0.0760	0.0674	0.1053	0.6399
Complex group 9	0.0324	0.0199	0.0754	0.2647
Complex group 11	0.0555	0.0312	0.1382	0.2255
Complex group 10	0.0945	0.0735	0.1574	0.4668
Complex group 12	0.0392	0.0123	0.1343	0.0919
Complex group 13	0.0217	0.0081	0.0657	0.1229
Complex group 14	0.0516	0.0376	0.0928	0.4054
Complex group 15	0.0330	0.0022	0.1388	0.0155
Complex group 16	0.0380	0.0158	0.1133	0.1397
Complex group 17	0.0265	0.0068	0.0924	0.0736
Complex group 18	0.0630	0.0540	0.0931	0.5797
Complex group 19	0.0107	0.0000	0.0466	0.0000
Complex group 20	0.0554	0.0482	0.0816	0.5909
Average for complex groups	**0.0566**	**0.0404**	**0.1115**	**0.3463**
Homo sapiens (human)	0.8350	0.6650	1.3310	0.4996
Callithrix jacchus (marmoset)	0.7360	0.5890	1.2230	0.4816
Gorilla gorilla (gorilla)	0.8510	0.6890	1.3410	0.5138
Macaca mulatta (rhesus monkey)	0.9040	0.7270	1.4500	0.5014
Nomascus leucogenys (Gibbon)	1.0050	0.8630	1.5680	0.5504

Table 1: Continued.

Name	d	K_a	K_s	K_a/K_s
Pan troglodytes (chimpanzee)	0.9980	0.8300	1.4780	0.5616
Pongo pygmaeus abelii (orangutan)	0.8830	0.7050	1.5140	0.4657
Average for all species	**0.8874**	**0.7240**	**1.4150**	**0.5106**

domain among each pair of homologs within clades from every complex group and orthologous group (Table 1 and detail data see Table S2). Whether in orthologous groups or complex groups, the average K_a/K_s ratios of most groups (82.4%) are much smaller than 1 (only six groups of average K_a/K_s ratios are greater than 0.5, all of them belong to complex groups), indicating that the actin genes code highly conserved proteins because of important functions and were under strong negative selection.

However, the average K_a/K_s ratio in all the complex groups was significantly greater than that in the orthologous groups (0.346 versus 0.0941, $P = 0.003$, using two-tailed t-test). Furthermore, the average K_a in complex groups was significantly greater than that in orthologous groups (0.0404 versus 0.0106, $P = 0.021$, using two-tailed t-test), while the average K_s for both two types of clades were not significantly different from each other (0.1115 for complex groups versus 0.0842 for orthologous groups; $P = 0.108$, using two-tailed t-test). This suggested that actin genes included in the two types of clades experienced similar evolutionary time, but undergone uneven selections. The results confirmed that the actin genes included in the orthologous groups were higher conserved, and the actin genes from complex groups may experience a relatively relaxed negative selection during a certain period.

At the same time, we also separately aligned the actin genes from each species and calculated the average nucleotide divergence, nonsynonymous, and synonymous substitutions in the genome (Table 1). Our results showed that the average K_a and K_s values for all pairs of paralogs in seven species were from 0.5890 to 0.8630 and from 1.2230 to 1.5680, respectively. Nevertheless, the maximum averages of K_a and K_s in both complex groups and orthologous groups were 0.2027 and 0.3317. The average K_a and K_s values for the paralogs within species were significantly greater than those for the homologous actin genes from different species within the same group ($P < 0.001$ for both K_a and K_s), implying that different actin genes with distinct functions may undergo diverse selective pressures.

3.3. Pseudogenes Contained Actin Domain. Discriminating pseudogene from live actin gene could help us to understand the evolutionary history of actin gene family. In our work, three-hundred and thirty-seven actin pseudogenes were identified in the seven genomes (see Section 2.3, Table S3). The number of pseudogenes was much greater than that of live actin copies. Marmoset genome has the largest number (63 copies) of pseudogenes, and the number of pseudogenes in other genomes in descending order was 59 in human, 51 in chimpanzee, 48 in orangutan, 41 in gorilla, 40 in gibbon, and 35 in rhesus monkey. All species except rhesus monkey own more dead actin genes than live ones. Actually, the number of pseudogene in rhesus monkey was equal to that of live ones, (see Table S1 and Table S3). The frameshift insertions or deletions and premature stop codons were observed in all the pseudogene sequences. The dead actin genes abundantly existing in all the seven genomes provided evidence that actin genes went through a duplication first and then varied in the evolutionary process.

3.4. Codon Usage. The synonymous codons, which code for the same amino acid, were reported to be used unequally in almost all species [38–43] and present the evolutionary pattern of genes. For this reason, the study of codon usage pattern could be helpful to understand actin gene family. To study the codon usage of actin genes within species and within groups (including complex groups and orthologous groups split from the tree), all the actin domain sequences of copies were examined by RSCU values of the 59 relevant codons (see Section 2.4). The variation of RSCU values for each codon from actin genes within each group and every species was calculated to examine the extent of difference in codon usage pattern. The larger the variation value was, the more various codon usage patterns among the groups or species there were. The differences of average variations were revealed between 34 groups and 7 species for every codon (Figure 2). The average variations in all the species were significantly greater than those in 34 groups for the total 59 codons ($P < 0.001$ for all codons, using two-tailed t-test, see Table S4). These results demonstrated that actin genes within groups come from different species that had relative coincident codon usage pattern, while the codon usage of actin genes within species diverged a lot.

3.5. Actin Gene Expression Pattern. The transcription profiling data of humans and chimpanzees in brain, heart, liver, kidney, and testis were employed to detect whether there were any differences of the expression patterns for actin genes between and within species [37]. The expression level values of 23 actin genes from human and 21 from chimpanzee in the five tissues were extracted from array data. Because the gene expression data were measured in multiple samples, with the addition of some actin genes represented by more than one corresponding probes, the average value of actin gene expression using every probe in all samples was adopted as the expression level for this actin copy in each tissue. The extent of difference in gene expression between paralogs and orthologs was measured by coefficient of variation (CV) of expression values (Table S5). In our results, a large proportion of the CV's (65/67) for actin genes within groups in the five tissues was smaller than 0.4; however, the minimum CV of actin genes within species was 0.875, and 80% of them were greater than 1.50 as well. The significant differences of average CV results between paralogs and orthologs in all the tissues were shown as in Figure 3 ($P < 0.001$, using two-tailed t-test). These results demonstrated that actin genes within species, which possessed distinct categories of functions, had differential expression levels among each other, whereas actin

Table 2: Distribution of actin copies from seven species in each complex group.

Group name	Human	Chimpanzee	Gorilla	Orangutan	Gibbon	Rhesus monkey	Marmoset	Total
Complex group 1	3	4	2	1	0	0	0	10
Complex group 2	2	1	1	1	1	2	0	8
Complex group 3	0	0	0	0	0	0	4	4
Complex group 4	0	1	1	1	0	0	0	3
Complex group 5	0	1	1	0	0	1	0	3
Complex group 6	1	0	1	1	1	2	0	6
Complex group 7	0	0	0	0	0	0	6	6
Complex group 8	0	1	1	1	1	1	0	5
Complex group 9	1	1	1	0	1	1	1	6
Complex group 11	1	1	1	1	0	1	1	6
Complex group 10	1	1	1	2	0	0	1	6
Complex group 12	1	1	1	1	1	1	2	8
Complex group 13	1	1	2	1	1	1	0	7
Complex group 14	0	0	1	1	0	1	1	4
Complex group 15	1	1	0	1	1	1	1	6
Complex group 16	1	1	2	1	1	1	1	8
Complex group 17	1	1	0	1	1	1	1	6
Complex group 18	1	1	1	2	1	3	1	10
Complex group 19	1	1	0	1	0	1	1	5
Complex group 20	0	2	0	0	1	1	0	4
Total	16	20	17	17	11	19	21	121

genes within groups but from different species, which might be involved in the identical function, expressed in the same level.

4. Discussion

4.1. Phylogenetic Analysis. Actin is reported as an abundant cytoskeletal protein that plays a central role in many cellular processes. The phylogenetic analysis of actin genes in multicellular animals showed that phylogeny corresponded well with distinct functional categories into, for example, cytosolic, smooth, and cardiac muscle actins [44] and more divergent actin-related proteins [13, 14]. However, the phylogenetic structures in dinoflagellates exhibited at least three types of clusters [26].

Based on our results, the apparent feature of orthologous groups was one actin gene copy from each species clustered together on the phylogenetic tree, possessing distinct functions, which were coincident with Oota's and Muller's results [13, 44]. Nevertheless, more than 50% of the total actin genes incompletely interspecifically or monophyletically clustered on the clades formed complex groups. In fact, actin genes within the complex groups could be divided into three types in detail, based on the branch length and organization of the groups. The first type consisted of complex groups 1, 3, 7, 12, and 13, which had more than one copy from a species in the clades, indicating recent duplication that occurred after speciation. The complex groups 4 and 5, which contained much more divergent actin gene sequences than the other groups did (see Table 1), were designated as type 2. These actin genes would possibly subject to faster relative mutation rate or longer divergence time than other genes. And the other complex groups belong to the third type, in which one or two orthologous copies were lost in some primates. Furthermore, 85% of the lost copies in the third type were found to become pseudogenes in the corresponding genomes or have truncated actin domain which were excluded in the work (the nucleotide length of actin domain was smaller than 160 bp). Thus, actin gene sequences within the three types of complex groups plus the orthologous groups in the primates, which had similar phylogenetic structures in dinoflagellates at some extent [26], appeared to have diverged from one another at different time points during and after speciation. The copy number variation on the phylogenetic tree reflected complicated evolutionary patterns of actin gene family. The results also implied that the actin gene family might obtain new function or alter original function by changing copy number in the genome during the evolutionary process.

4.2. Distinct Selection. Actin genes within orthologous groups and complex groups showed significantly different levels of nucleotide diversity, K_a and K_a/K_s ratios, suggesting they had undergone nonuniformly selections. The K_a/K_s ratios in the orthologous groups were significantly smaller than those in the complex groups (Table 1), implying that actin genes within orthologous groups were highly conserved under strong background selection owning to their basic functions in the cells. On the other hand, relaxation of negative selection or positive selection was associated with

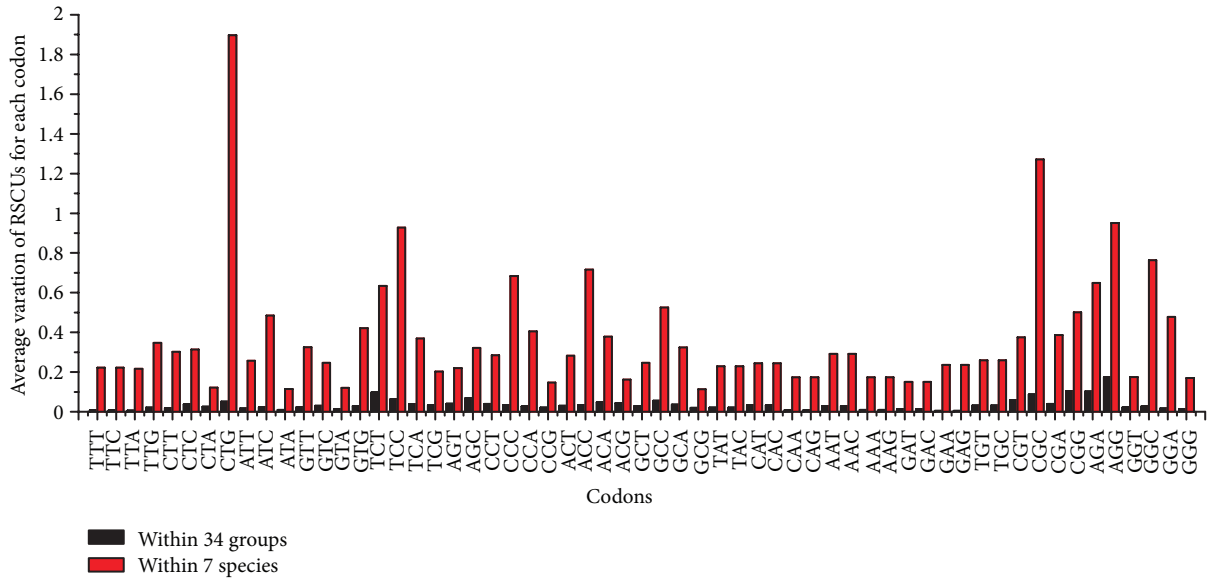


Figure 2: Variations of relative synonymous codon usage (RSCU) values for actin genes within groups (black columns) and species (red columns) in each codon. Each column represents average variations of all groups (or species) for one codon. For all the 59 codons, the red columns are significantly higher than black ones ($P < 0.001$ for all codons, detailed corresponding P value for each codon see Table S4 in Supplementary Material available online at http://dx.doi.org/10.1155/2013/630803), representing that the codon usage patterns of actin genes within species were very distinct among each other, while those of actin genes within groups were much similar.

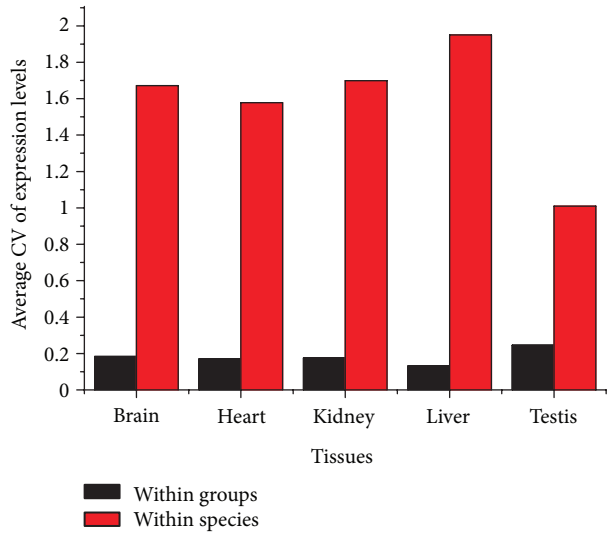


Figure 3: Coefficient of variations (CVs) of gene expression values for actin genes within groups (black columns) and species (red columns) in each tissue. Each column represents average CVs of all groups (or species) for one tissue. For all the five tissues, the black columns are significantly lower than red ones, indicating that the actin genes within groups exhibited similar expression patterns, while actin genes within species did not ($P < 0.001$, using two-tailed t-test).

copy number variation, which was detected in complex groups, leading to relatively rapid diversification of actin genes within complex groups.

At the same time, we found that actin genes were tremendously divergent from each other within species, which the average nucleotide diversity, K_a and K_s of the actin domain sequences in each genome was greater than 0.70, 0.58 and 1.20, respectively (Table 1). This result was much unexpected, for actin was one of the most highly conserved proteins [24]. Actins in plant genomes, such as *Populus* and *Arabidopsis thaliana*, were reported to share high sequence homology, larger than 90% identity [45, 46]. Similarly, K_a and K_s for actin genes in dinoflagellate species were less than 0.05 and 0.55 [26], much smaller than our results. In consideration of the fundamental importance of actins, we inferred that molecular diversification of actin gene family could result in

functional diversification in the complex higher organisms like primates. Besides, 82.8% of actins were conserved across species instead of within species, suggesting that parallel selection played a major role in the evolution of actins.

4.3. The Different Characteristics of Actin Genes across and within Species. On the basis of the transcription profiling data of humans and chimpanzee in brain, heart, liver, kidney, and testis, 72.7% of actin genes appear to be differentially expressed in different tissues. The actin genes, twelve from human and ten from chimpanzee, have available array data in all the five tissues (see Table S6), for which the average CVs of expression level values were 0.584 and 0.527 of human and chimpanzee, respectively. The CVs for 16 out of 22 actin genes were greater than 0.4, significantly greater than those for almost all the orthologs of actin genes between species, suggesting overlapping and unique expression patterns of actin gene family members due to distinct functions. The findings were in agreement with previous studies, in which the isoforms of nonplant actin appear to be differentially expressed in striated muscle, smooth muscle, and nonmuscle tissues [47], and individual actins from plants, such as *Arabidopsis thaliana* and *Populus*, show specific expression patterns, congruent with their evolutionary relationships [45, 46, 48–50].

In addition, the average CVs of expressed values for actin copies within species in every tissue were significantly greater than those of each actin copy among the tissues, implying that actin genes with distinct functions had different expression levels. The CVs of gene expression values for actin copes within species were much greater than those for actin genes within groups across species, suggesting a big difference between paralogs and orthologs of actin copies. Besides, similar results in the codon usage pattern as well as the findings for divergence and ratio of nonsynonymous to synonymous substitutions were also revealed between paralogs and orthologs of actin copies, implying that actin copies were highly homologous within groups. All these results might provide a clue for paralogs and orthologs of actin copies through different evolutionary histories.

4.4. Dynamic Actin Gene Family Evolution in Primates. In agreement with previous studies, actin was highly conserved across primates due to its important functions, proved by our results that about 40% of actin genes belong to the orthologous groups with well-interspecific distribution and little divergence. On the other hand, actin was needed to obtain new function constantly in order to adapt more and more complicated system in complex higher organisms. How did actin evolve to meet the pair of conflict demands in primates?

Generally speaking, gene acquired new function resulting from increase of self-complexity or copy number variation. Increasing gene length or fusing with other domains could increase its complex, while duplication offered a chance to gain new functions without losing the original ones.

Interestingly, several actins were found to fuse with other domains to generate new functions. For example, actins within complex group 1 also contained POTE ankyrin domain [51, 52]. And a length of DUF1542 domain sequences was examined to insert into actin genes within complex group 17 that encode *ACTR5* protein. Similarly, actin domain of *ACTR8* genes comprising orthologous group 12 was encompassed in COG5277 domain. However, the rest of the actin genes comprising complex groups chose the other way. The tremendous pseudogenes, presenting for the copies that failed to gain new function, also gave an evidence for the numerous duplications of actin genes. The organization of actins and characteristics of actin gene family indicated that actins acquired new function in various evolutionary pathways. Both of increasing self-complexity and copy number, especially the second way, played important roles in promoting the evolution of actin.

Taken together, several evolutionary characteristics of actin gene family in primates were observed in our results. First, the phylogenetic tree structure for all the actin domain sequences exhibited that 89.7% of actin genes clustered with other orthologous copies from distinct species, implying incomplete lineage sorting [53] during the divergence of the seven primates and inconsistent divergence time or rate of variance between gene copies. Second, the differences of divergences, codon usage, and expression patterns between orthologs and paralogs of actin copies within groups and within species indicated that actin genes within groups were highly homologous, but actin genes within species were very divergent. Therefore, we deduced that multiple rounds of gene duplication events have occurred and that the most multiple actin gene homologs likely existed in the recent common ancestor. Finally, the presence of a great deal of pseudogenes provided convinced evidence for actin gene experiencing duplicated, mutated, and dead process. We conclude that gene family expansion and contraction have continued during and after speciation of these primates. These features of actin gene family in primates provided evidence for us to explain how actin gene family evolved leading to the contradictory characteristics of conserved across species but divergent within species in the evolutionary history in primates.

Generally, "concerted evolution" and "birth and death" models were often invoked to explain the divergence and evolution of multigene families [54]. Under concerted evolution model, new gene copies were duplicated, homogenized, and deleted by interlocus recombination or intergenic gene conversion, resulting in a high degree of sequence similarity among multigene family members [55–57]. In contrast, under the model of "birth and death," new gene copy was created by various forms of gene duplications in order to diverge functionally, in which some duplicated copies with new function or original function maintained in the genomes, and others became nonfunctional or deleted due to mutation and degeneration. Thus, the predicted end result of this model was a mixture of divergent groups of genes and highly homologous genes within groups plus many great pseudogenes present in the multigene family [54].

Obviously, although actin gene copies from the same species shared highly similar sequences clustered on the first type of complex groups partly corresponded with convert evolution model, the major characteristics of actin genes,

such as the variation in copy number, the structure of the phylogenetic tree with a mixture of divergent groups of gene copies, the differences of divergences, codon usage, expression patterns between orthologs and paralogs of actin genes across and within species, and the presence of many pseudogenes, fit well with "birth and death" model of multigene family evolution [54].

Since actin family plays such a crucial role in all aspects of cell activities, their related functions cannot be easily altered or removed. However, the way of the copy number of actin genes changed following "birth and death model" maybe affording an alternative evolutionary pathway to meet the conflicting demands that actin was conserved to maintain vital functions and evolved new functions in the body in order to help adapting to environmental pressure. Under this scenario, organisms may not only keep bodies working regularly, but make species evolving from simple to complex, from rough to fine. We infer that birth and death evolution model might be a common evolutionary mechanism in other highly conserved multigene families.

5. Conclusions

In summary, 233 actin genes and 337 pseudogenes were identified in the seven primates. Phylogenetic analysis for actin genes exhibited two major types of clades. Actin genes interspecifically clustered that belong to the orthologous groups were highly conserved because of fundamental importance. On the contrary, complex groups contained actin gene members that displayed copy number variation with significantly higher levels of average nucleotide divergence and K_a/K_s ratios compared to the orthologous groups. Analysis of codon bias and gene expression level revealed that actin genes in primates were extraordinarily divergent from each other within species, but were highly conserved within groups across species. These results may be explained by a birth and death evolutionary process of actin gene families, which would be the general evolutionary mechanism for other highly conserved multigene families.

Abbreviations

CDS: Coding sequences
d: Divergence/diversity
K_a: Nonsynonymous substitutions
K_s: Synonymous substitutions
RSCU: Relative synonymous codon usage
CV: Coefficient of variation
CDD: Conserved domain database
BLAST: Basic Local Alignment Search Tool.

Conflict of Interests

The authors declare that they have no conflict of interests.

Authors' Contribution

L. Zhu and Y. Zhang contributed equally to this work.

Acknowledgments

The authors thank the editor and the reviewers for their suggestions. This work was supported by grants from the Shanghai Science Foundation of China (12ZR1444200), National Natural Science Foundation of China (61103075), and the First-Class Discipline of Universities in Shanghai.

References

[1] I. M. Sehring, J. Mansfeld, C. Reiner, E. Wagner, H. Plattner, and R. Kissmehl, "The actin multigene family of Paramecium tetraurelia," *BMC Genomics*, vol. 8, article 82, 2007.

[2] J. F. Morrow, R. S. Stearman, C. G. Peltzman, and D. A. Potter, "Induction of hepatic synthesis of serum amyloid A protein and actin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 78, no. 8, pp. 4718–4722, 1981.

[3] J. L. Degen, M. G. Neubauer, S. J. Degen, C. E. Seyfried, and D. R. Morris, "Regulation of protein synthesis in mitogen-activated bovine lymphocytes. Analysis of actin-specific and total mRNA accumulation and utilization," *The Journal of Biological Chemistry*, vol. 258, no. 20, pp. 12153–12162, 1983.

[4] C. J. Staiger, Y. Ming, R. Valenta, P. J. Shaw, R. M. Warn, and C. W. Lloyd, "Microinjected profilin affects cytoplasmic streaming in plant cells by rapidly depolymerizing actin microfilaments," *Current Biology*, vol. 4, no. 3, pp. 215–219, 1994.

[5] J. Snider, F. Lin, N. Zahedi, V. Rodionovt, C. C. Yu, and S. P. Gross, "Intracellular actin-based transport: how far you go depends on how often you switch," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 36, pp. 13204–13209, 2004.

[6] M. Ghosh, X. Song, G. Mouneimne, M. Sidani, D. S. Lawrence, and J. S. Condeelis, "Cofilin promotes actin polymerization and defines the direction of cell motility," *Science*, vol. 304, no. 5671, pp. 743–746, 2004.

[7] K. L. Hill, N. L. Catlett, and L. S. Weisman, "Actin and myosin function in directed vacuole movement during cell division in Saccharomyces cerevisiae," *The Journal of Cell Biology*, vol. 135, no. 6, pp. 1535–1549, 1996.

[8] M. S. Otegui, K. J. Verbrugghe, and A. R. Skop, "Midbodies and phragmoplasts: analogous structures involved in cytokinesis," *Trends in Cell Biology*, vol. 15, no. 8, pp. 404–413, 2005.

[9] E. Reisler, "Actin molecular structure and function," *Current Opinion in Cell Biology*, vol. 5, no. 1, pp. 41–47, 1993.

[10] P. K. Howard, B. M. Sefton, and R. A. Firtel, "Tyrosine phosphorylation of actin in Dictyostelium associated with cell-shape changes," *Science*, vol. 259, no. 5092, pp. 241–244, 1993.

[11] D. A. Schafer and T. A. Schroer, "Actin-related proteins," *Annual Review of Cell and Developmental Biology*, vol. 15, pp. 341–363, 1999.

[12] L. M. Machesky and R. C. May, "Arps: actin-related proteins," *Results and Problems in Cell Differentiation*, vol. 32, pp. 213–229, 2001.

[13] J. Muller, Y. Oma, L. Vallar, E. Friederich, O. Poch, and B. Winsor, "Sequence and comparative genomic analysis of actin-related proteins," *Molecular Biology of the Cell*, vol. 16, no. 12, pp. 5736–5748, 2005.

[14] H. V. Goodson and W. F. Hawse, "Molecular evolution of the actin family," *Journal of Cell Science*, vol. 115, no. 13, pp. 2619–2622, 2002.

[15] M. K. Kandasamy, R. B. Deal, E. C. McKinney, and R. B. Meagher, "Plant actin-related proteins," *Trends in Plant Science*, vol. 9, no. 4, pp. 196–202, 2004.

[16] J. L. Hodgkinson, C. Peters, S. A. Kuznetsov, and W. Steffen, "Three-dimensional reconstruction of the dynactin complex by single-particle image analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 10, pp. 3667–3672, 2005.

[17] A. Efstratiadis, J. W. Posakony, T. Maniatis et al., "The structure and evolution of the human β-globin gene family," *Cell*, vol. 21, no. 3, pp. 653–668, 1980.

[18] G. G. Brown, J. S. Lee, N. Brisson, and D. P. S. Verma, "The evolution of a plant globin gene family," *Journal of Molecular Evolution*, vol. 21, no. 1, pp. 19–32, 1984.

[19] J. N. Engel, P. W. Gunning, and L. Kedes, "Isolation and characterization of human actin genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 78, no. 8, pp. 4674–4678, 1981.

[20] E. A. Fyrberg, B. J. Bond, N. D. Hershey, K. S. Mixter, and N. Davidson, "The actin genes of drosophila: protein coding regions are highly conserved but intron positions are not," *Cell*, vol. 24, no. 1, pp. 107–116, 1981.

[21] R. Zakut, M. Shani, D. Givol, S. Neuman, D. Yaffe, and U. Nudel, "Nucleotide sequence of the rat skeletal muscle actin gene," *Nature*, vol. 298, no. 5877, pp. 857–859, 1982.

[22] J. G. Files, S. Carr, and D. Hirsh, "Actin gene family of Caenorhabditis elegans," *Journal of Molecular Biology*, vol. 164, no. 3, pp. 355–375, 1983.

[23] J. A. Fornwald, G. Kuncio, I. Peng, and C. P. Ordahl, "The complete nucleotide sequence of the chick a-actin gene and its evolutionary relationship to the actin gene family," *Nucleic Acids Research*, vol. 10, no. 13, pp. 3861–3876, 1982.

[24] R. C. Hightower and R. B. Meagher, "The molecular evolution of actin," *Genetics*, vol. 114, no. 1, pp. 315–332, 1986.

[25] M. Wu, J. M. Comeron, H. S. Yoon, and D. Bhattacharya, "Unexpected dynamic gene family evolution in algal actins," *Molecular Biology and Evolution*, vol. 26, no. 2, pp. 249–253, 2009.

[26] S. Kim, T. R. Bachvaroff, S. M. Handy, and C. F. Delwiche, "Dynamics of actin evolution in dinoflagellates," *Molecular Biology and Evolution*, vol. 28, no. 4, pp. 1469–1480, 2011.

[27] R. J. Kinsella, A. Kahari, S. Haider et al., "Ensembl BioMarts: a hub for data retrieval across taxonomic space," *Database*, vol. 2011, Article ID bar030, 2011.

[28] A. Marchler-Bauer, J. B. Anderson, P. F. Cherukuri et al., "CDD: a conserved domain database for protein classification," *Nucleic Acids Research*, vol. 33, pp. D192–D196, 2005.

[29] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1596–1599, 2007.

[30] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[31] M. Lynch and T. J. Crease, "The analysis of population survey data on DNA sequence variation," *Molecular Biology and Evolution*, vol. 7, no. 4, pp. 377–394, 1990.

[32] M. Nei and T. Gojobori, "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions," *Molecular Biology and Evolution*, vol. 3, no. 5, pp. 418–426, 1986.

[33] P. M. Sharp, T. M. F. Tuohy, and K. R. Mosurski, "Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes," *Nucleic Acids Research*, vol. 14, no. 13, pp. 5125–5143, 1986.

[34] T. Zhou, W. Gu, J. Ma, X. Sun, and Z. Lu, "Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses," *BioSystems*, vol. 81, no. 1, pp. 77–86, 2005.

[35] I. Ahn, B. J. Jeong, S. E. Bae, J. Jung, and H. S. Son, "Genomic analysis of influenza A viruses, including avian flu (H5N1) strains," *European Journal of Epidemiology*, vol. 21, no. 7, pp. 511–519, 2006.

[36] G. Perrière and J. Thioulouse, "Use and misuse of correspondence analysis in codon usage studies," *Nucleic Acids Research*, vol. 30, no. 20, pp. 4548–4555, 2002.

[37] P. Khaitovich, I. Hellmann, W. Enard et al., "Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees," *Science*, vol. 309, no. 5742, pp. 1850–1854, 2005.

[38] R. Percudani and S. Ottonello, "Selection at the wobble position of codons read by the same tRNA in Saccharomyces cerevisiae," *Molecular Biology and Evolution*, vol. 16, no. 12, pp. 1752–1762, 1999.

[39] S. Kanaya, Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura, "Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis," *Journal of Molecular Evolution*, vol. 53, no. 4-5, pp. 290–298, 2001.

[40] G. Sánchez, A. Bosch, and R. M. Pintó, "Genome variability and capsid structural constraints of hepatitis A virus," *Journal of Virology*, vol. 77, no. 1, pp. 452–459, 2003.

[41] L. A. Shackelton, C. R. Parrish, and E. C. Holmes, "Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses," *Journal of Molecular Evolution*, vol. 62, no. 5, pp. 551–563, 2006.

[42] C. T. Tsai, C. H. Lin, and C. Y. Chang, "Analysis of codon usage bias and base compositional constraints in iridovirus genomes," *Virus Research*, vol. 126, no. 1-2, pp. 196–206, 2007.

[43] K. M. Pepin, J. Domsic, and R. McKenna, "Genomic evolution in a virus under specific selection for host recognition," *Infection, Genetics and Evolution*, vol. 8, no. 6, pp. 825–834, 2008.

[44] S. Oota and N. Saitou, "Phylogenetic relationship of muscle tissues deduced from superimposition of gene trees," *Molecular Biology and Evolution*, vol. 16, no. 6, pp. 856–867, 1999.

[45] D. Zhang, Q. Du, B. Xu, Z. Zhang, and B. Li, "The actin multigene family in Populus: organization, expression and phylogenetic analysis," *Molecular Genetics and Genomics*, vol. 284, no. 2, pp. 105–119, 2010.

[46] K. Slajcherova, J. Fiserova, L. Fischer, and K. Schwarzerova, "Multiple actin isotypes in plants: diverse genes for diverse roles?" *Frontiers in Plant Science*, vol. 3, article 226, 2012.

[47] P. Gunning, P. Ponte, L. Kedes, R. J. Hickey, and A. I. Skoultchi, "Expression of human cardiac actin in mouse L cells: a sarcomeric actin associates with a nonmuscle cytoskeleton," *Cell*, vol. 36, no. 3, pp. 709–715, 1984.

[48] J. Vandekerckhove and K. Weber, "At least six different actins are expressed in a higher mammal: an analysis based on the amino acid sequence of the amino-terminal tryptic peptide," *Journal of Molecular Biology*, vol. 126, no. 4, pp. 783–802, 1978.

[49] A. S. Tsang, H. Mahbubani, and J. G. Williams, "Cell-type specific actin mRNA populations in dictyostelium discoideum," *Cell*, vol. 31, no. 2, pp. 375–382, 1982.

[50] E. A. Fyrberg, J. W. Mahaffey, B. J. Bond, and N. Davidson, "Transcripts of the six Drosophila actin genes accumulate in a stage- and tissue-specific manner," *Cell*, vol. 33, no. 1, pp. 115–123, 1983.

[51] Y. Lee, T. Ise, D. Ha et al., "Evolution and expression of chimeric POTE-actin genes in the human genome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 47, pp. 17885–17890, 2006.

[52] T. K. Bera, D. B. Zimonjic, N. C. Popescu et al., "POTE, a highly homologous gene family located on numerous chromosomes and expressed in prostate, ovary, testis, placenta, and prostate cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 26, pp. 16975–16980, 2002.

[53] A. P. Rooney and T. J. Ward, "Birth-and-death evolution of the internalin multigene family in Listeria," *Gene*, vol. 427, no. 1-2, pp. 124–128, 2008.

[54] M. Nei, X. Gu, and T. Sitnikova, "Evolution by the birth-and-death process in multigene families of the vertebrate immune system," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 15, pp. 7799–7806, 1997.

[55] G. P. Smith, "Mouse immunoglobulin kappa chain MPC 11: extra amino terminal residues," *Science*, vol. 181, no. 4103, pp. 941–943, 1973.

[56] E. A. Zimmer, S. L. Martin, S. M. Beverley, Y. W. Kan, and A. C. Wilson, "Rapid duplication and loss of genes coding for the α chains of hemoglobin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, no. 4, pp. 2158–2162, 1980.

[57] D. M. Irwin and A. C. Wilson, "Concerted evolution of ruminant stomach lysozymes. Characterization of lysozyme cDNA clones from sheep and deer," *The Journal of Biological Chemistry*, vol. 265, no. 9, pp. 4944–4952, 1990.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 267375, 8 pages
http://dx.doi.org/10.1155/2013/267375

Research Article

Identification of Lung-Cancer-Related Genes with the Shortest Path Approach in a Protein-Protein Interaction Network

Bi-Qing Li,[1,2] Jin You,[1,3] Lei Chen,[4] Jian Zhang,[5] Ning Zhang,[6] Hai-Peng Li,[7] Tao Huang,[8] Xiang-Yin Kong,[1,3] and Yu-Dong Cai[9]

[1] *The Key Laboratory of Stem Cell Biology, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[2] *Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[3] *Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China*
[4] *College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*
[5] *Department of Ophthalmology, Shanghai First People's Hospital, Shanghai Jiaotong University, Shanghai 200080, China*
[6] *Department of Biomedical Engineering Tianjin University, Tianjin Key Lab of BME Measurement, Tianjin 300072, China*
[7] *CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[8] *Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA*
[9] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*

Correspondence should be addressed to Xiang-Yin Kong; xykong@sibs.ac.cn and Yu-Dong Cai; cai_yud@yahoo.com.cn

Received 12 March 2013; Revised 19 April 2013; Accepted 29 April 2013

Academic Editor: Bing Niu

Lung cancer is one of the leading causes of cancer mortality worldwide. The main types of lung cancer are small cell lung cancer (SCLC) and nonsmall cell lung cancer (NSCLC). In this work, a computational method was proposed for identifying lung-cancer-related genes with a shortest path approach in a protein-protein interaction (PPI) network. Based on the PPI data from STRING, a weighted PPI network was constructed. 54 NSCLC- and 84 SCLC-related genes were retrieved from associated KEGG pathways. Then the shortest paths between each pair of these 54 NSCLC genes and 84 SCLC genes were obtained with Dijkstra's algorithm. Finally, all the genes on the shortest paths were extracted, and 25 and 38 shortest genes with a permutation P value less than 0.05 for NSCLC and SCLC were selected for further analysis. Some of the shortest path genes have been reported to be related to lung cancer. Intriguingly, the candidate genes we identified from the PPI network contained more cancer genes than those identified from the gene expression profiles. Furthermore, these genes possessed more functional similarity with the known cancer genes than those identified from the gene expression profiles. This study proved the efficiency of the proposed method and showed promising results.

1. Introduction

Lung cancer is one of the leading causes of cancer mortality worldwide [1]. Two main types of lung cancer are non-small cell lung cancer (NSCLC), which accounts for 80%–85%, and small cell lung cancer (SCLC), which accounts for around 20% of all cases. However, the SCLC has an extraordinarily high degree of metastasis and a strong association with smoking [2]. Diagnosis and treatment at the early stage of the disease process could reduce fatalities and increase the probability of disease-free survival. Therefore, it is meaningful to screen lung-cancer-related genes that could be used as prognostic factors or to help elucidate the mechanism of the disease.

Recently, as high-throughput biotechnologies develop rapidly, numerous biological data have been generated from processes such as protein complex, yeast two-hybrid systems, and gene expression profiles. These data are useful resources for understanding and deducing gene function. So far, protein-protein interaction (PPI) data has been widely utilized to annotate and predict the gene function assuming that interaction proteins possess the similar or identical

functions and thus may participate in the same pathways. This so-called "guilt by association" rule was initially proposed by Nabieva et al. [3]. This rule could also be utilized to identify novel cancer-related genes.

Search Tool for the Retrieval of Interacting Genes (STRING) is an online database resource [4] that provides both predicted and experimental interaction information with a confidence score. It has been shown that proteins with short distances between each other in the PPI network tend to have the same biological functions [5–8], and interactive neighbors are prone to have the same biological functions as noninteractive ones [9, 10]. The possible reason is that the query protein and its interactive proteins might form a protein complex to exert a particular function or might participate in the same pathways.

Though great successes have been achieved for gene function prediction and identification of novel cancers-related genes with the application of the high-throughput data, yet high-throughput data is not error free. In this work, we proposed a computational method for identifying lung-cancer-related genes based on PPI network constructed from STRING. 54 NSCLC and 84 SCLC related genes were retrieved from associated KEGG pathways. Then, Dijkstra's algorithm [11] was employed to obtain the shortest paths between each pair of the 54 NSCLC and 84 SCLC genes. All the genes present on the shortest paths were extracted and analyzed. Several of these genes have been reported to be related to lung cancer. However, some of them were not previously reported. Therefore, there are probably novel lung-cancer-related genes and have the potential to be biomarkers for diagnosis of lung cancer.

2. Materials and Methods

2.1. Lung-Cancer-Related Gene List. We compiled all 54 genes existing in the human nonsmall cell lung cancer (NSCLC) pathway and 84 genes in the small cell lung cancer (SCLC) pathway from KEGG [12]. These two gene sets and corresponding Ensembl protein IDs are listed in Additional file S1 in Supplementary Matrial available online at doi: http://dx.doi.org/10.1155/2013/267375.

2.2. Lung Cancer Gene Expression Data. The gene expression profiling in Kastner et al.'s work was used in our study [13], which includes 8 SCLC, 16 NSCLC, and 14 normal lung tissue samples. It was retrieved from NCBI Gene Expression Omnibus (GEO) (Accession number: GSE40275). The gene expression profile was obtained by the Human Exon 1.0 ST Array with 56283 probes corresponding to 26410 genes. Signal intensity was first log2 transformed and then quantile normalized with "preprocessCore" package of R [14].

2.3. Identifying Differentially Expressed Genes. The "samr" package of R [15] was utilized to identify the differentially expressed genes between NSCLC, SCLC, and normal tissues separately with the criterion that false-discovery-rate- (FDR-) adjusted P value [16] was less than 0.01 and fold change was greater than 3 or less than 0.33.

2.4. Cancer-Related Gene List. A list of 742 cancer-related genes was compiled from three different sources [17]. First, a list of 457 cancer-related genes was collected from the Cancer Gene Census. Second, a list of cancer-related genes from the Atlas of Genetics and Cytogenetic in Oncology was retrieved [18]. We compiled the third list from the Human Protein Reference Database (HPRD) [19].

2.5. STRING PPI Data and Shortest Path Identification. The initial weighted PPI network was constructed based on data from STRING (version 9.0) [4] (http://string.embl.de/). Each interaction in STRING is evaluated by an interaction confidence score in range from 1 to 999 to quantify the likelihood that an interaction may occur. We used Dijkstra's algorithm which has also been used in our previous works [20, 21] to identify the shortest path between each protein pair corresponding to the 54 NSCLC and 84 SCLC genes in the PPI network, respectively. Finally, all the proteins present on the shortest paths were ranked according to their betweenness. The Dijkstra's algorithm was implemented with R package "igraph" [22].

2.6. KEGG Pathway Enrichment Analysis. KEGG pathway enrichment analysis was performed with the functional annotation tool DAVID [23]. The enrichment P value was corrected with the Benjamin multiple testing correction method to control family-wide false discovery rate less than 0.05 [24]. All the protein-coding genes in human genome were taken as background during the enrichment analysis.

3. Results

3.1. Differentially Expressed Genes of the Gene Expression Profile. With the SAMR method, 1918 significantly upexpressed probes and 2243 downexpressed probes corresponding to 1825 genes were identified for NSCLC when compared with 14 normal lung tissues (for probes see additional file S2, and for gene symbols see additional file S3). For SCLC, 819 significantly up-expressed probes and 820 down-expressed probes corresponding to 1063 genes were identified (for probes see additional file S2, and for gene symbols see additional file S3).

3.2. Shortest Path Genes and Enrichment Analysis. An undirected graph was constructed with the PPI data from STRING. Subsequently, we repeatedly chose a pair of proteins corresponding to 54 NSCLC genes and the 84 SCLC genes respectively, and the shortest path between these two proteins was determined with Dijkstra's algorithm. A total of 1711 and 3916 shortest paths were obtained (see additional file S4) with lowest cost for NSCLC and SCLC containing 114 and 161 path genes, respectively. Shown in Figure 1 are the 1711 shortest paths between the 54 NSCLC genes. The weight was labeled on the edge between each of the interaction gene pairs. Shown in Figure 2 are the 3916 shortest paths between the 84 SCLC genes. To determine whether our 114 and 161 shortest path genes were also hubs in the background network, we performed a permutation to count the number

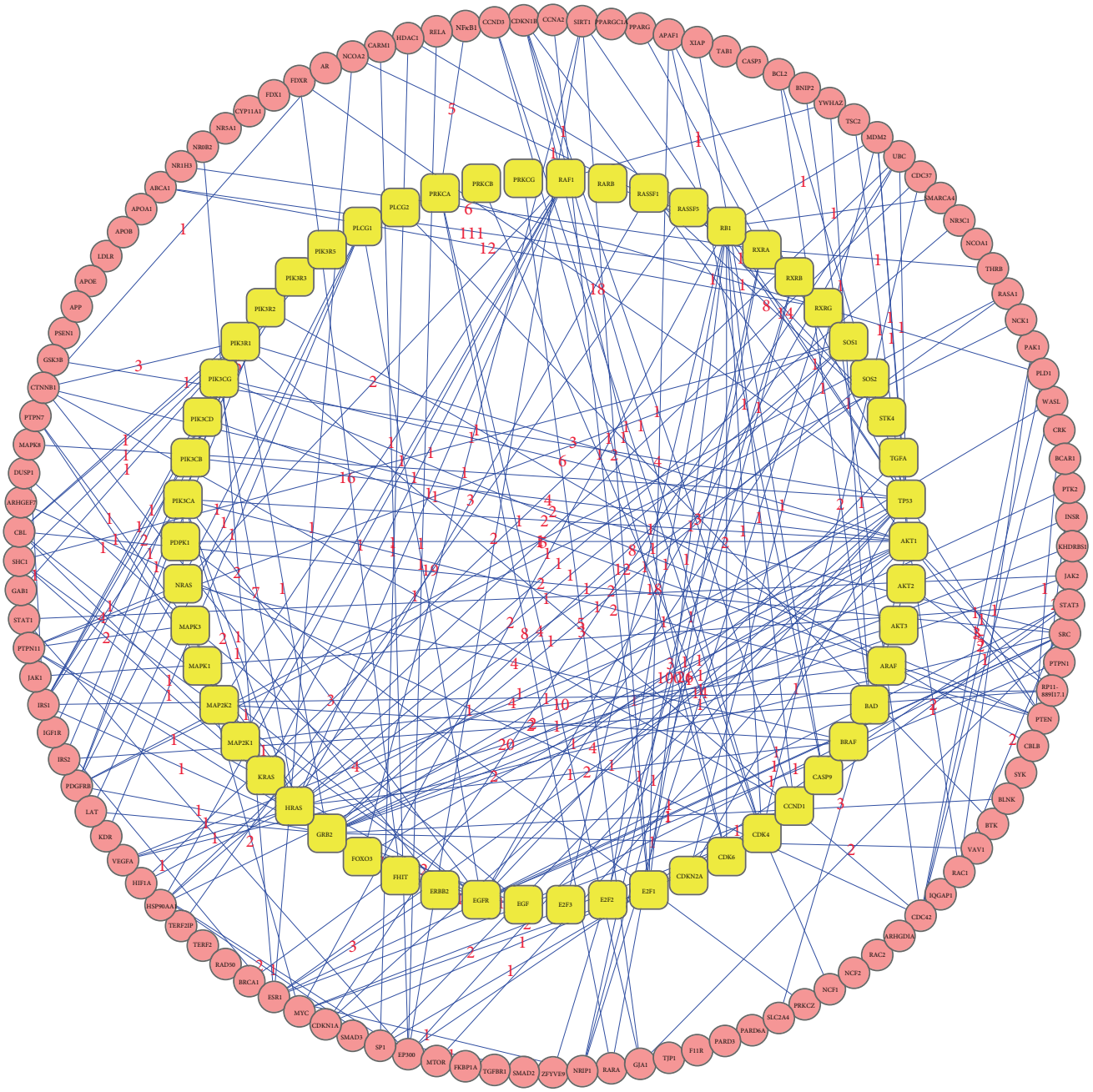


FIGURE 1: 1711 shortest paths between 54 NSCLC genes. The 1171 shortest paths between 54 NSCLC genes were identified with Dijkstra's algorithm based on PPI data from STRING. Yellow round represents 54 NSCLC genes. Red round represents 114 genes existing on shortest paths. Numbers on edges represent the edge weight to quantify the interaction confidence. The smaller the number, the stronger the interaction between two nodes.

of their occurrences on the shortest paths between 54 and 84 randomly selected genes only if they had a greater betweenness than that in our study. This process was repeated 2000 times, and the proportion of occurrences for the 114 and 161 shortest path genes was regarded as the P value. The detailed results thus obtained are given in additional file S5. Then we chose the 25 NSCLC and 38 SCLC shortest path genes with a P value less than 0.05 for further analysis (see additional file 5).

The GO enrichment analysis of 25 NSCLC shortest path genes indicated that they were significantly enriched in the regulation of intracellular signaling cascades and regulation of macromolecule metabolic processes (see additional file S6). These terms had been demonstrated to make great contributions to the survival and reproduction of cancer cells, while they also appeared in the enriched GO terms of 38 SCLC shortest path genes (see additional file S6). Besides these terms, the analysis result of SCLC shortest path

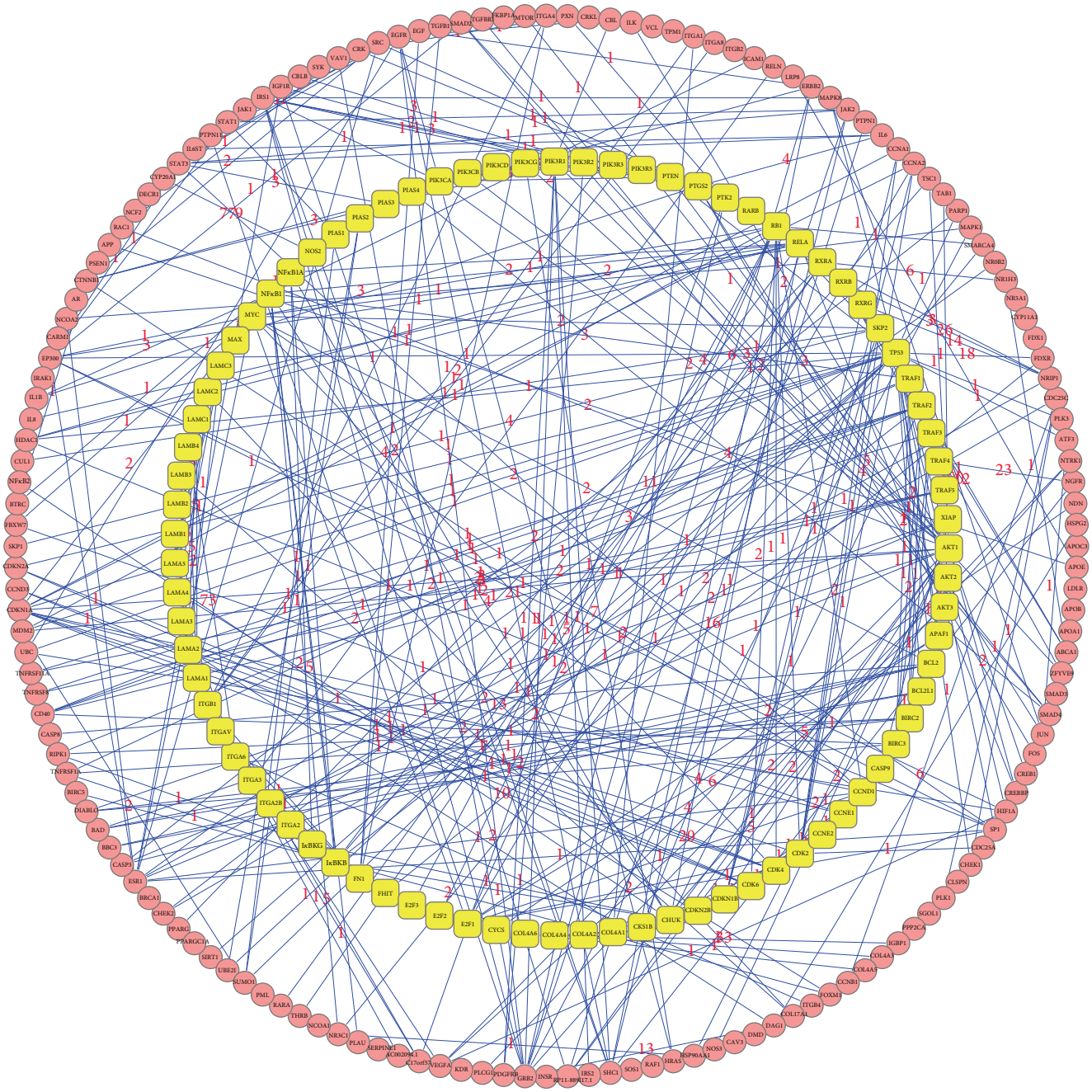


Figure 2: 3916 shortest paths between 84 SCLC genes. The 3916 shortest paths between 84 SCLC genes were identified with Dijkstra's algorithm based on PPI data from STRING. Yellow round represents 84 SCLC genes. Red round represents 161 genes existing on shortest paths. Numbers on edges represent the edge weight to quantify the interaction confidence. The smaller the number, the stronger the interaction between two nodes.

genes showed that they were significantly enriched in cell adhesion processes, suggesting that genes in this term might play an important role in differentiating SCLC from NSCLC (see additional file S6). The KEGG pathway enrichment of these 38 SCLC shortest path genes indicated that they were enriched in canonical-cancer-related pathways such as the cell cycle and p53 signaling pathway (Table 1).

3.3. Comparing the Overlap between Candidate Genes with 742 Cancer-Related Genes. The 25 and 38 shortest path genes were regarded as candidate genes for NSCLC and SCLC, respectively. We checked the overlap between 742 cancer genes and differentially expressed genes from the gene expression array as well as the overlap between the candidate genes identified in our study (Table 2). The entire 5-gene

Table 1: KEGG enrichment analysis of 38 SCLC shortest path genes.

Term	Count[a]	Percentage[b]	P value	Benjamini adjusted P value
Focal adhesion	8	21.1	$1.40E-05$	$6.70E-04$
Regulation of actin cytoskeleton	7	18.4	$2.20E-04$	$5.40E-03$
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	5	13.2	$2.70E-04$	$4.40E-03$
ECM-receptor interaction	5	13.2	$4.00E-04$	$4.80E-03$
Hypertrophic cardiomyopathy (HCM)	5	13.2	$4.20E-04$	$4.10E-03$
Dilated cardiomyopathy	5	13.2	$5.70E-04$	$4.60E-03$
Cell cycle	5	13.2	$1.80E-03$	$1.20E-02$
p53-signaling pathway	4	10.5	$2.90E-03$	$1.70E-02$

[a]The number of genes belonging to a certain pathway.
[b]The percentage of genes belonging to a certain pathway accounts for all the genes undergoing KEGG pathway analysis.

Table 2: Overlap between candidate genes and cancer-related genes.

Gene set	Number of candidate genes	Overlap with 742 cancer genes	P value
NSCLC from array	1825	93	$6.698e-04$
SCLC from array	1063	69	$2.218e-06$
NSCLC in our study	25	6	$2.518e-05$
SCLC in our study	38	5	$2.559e-03$

P value was calculated with the hypergeometric test assuming the total number of protein-coding genes was 20000.

Table 3: Comparing the overlap between candidate genes with cancer-related genes.

Gene set	Number of candidate genes	Overlap with 742 cancer genes	P value
NSCLC from array	1825	93	
NSCLC in our study	25	6	$3.858e-03$
SCLC from array	1063	69	
SCLC in our study	38	5	0.186

P value was calculated with Fisher's exact test.

set can be found in additional file S3. From Table 2, we can see that both the lung cancer candidate genes identified from the gene expression array and those identified by our method had a significant overlap with the 742 cancer genes. However, the 25 NSCLC candidate genes identified with our method contained more cancer genes than those from the gene expression array (P value = $3.858e-03$) (Table 3). The 38 SCLC candidate genes had a higher percentage of cancer-related genes (0.1316) than those from expression array (0.0649) though the P value of Fisher's exact test was not significant (P value = 0.186). At least, the 38 SCLC candidate genes contained comparable cancer-related genes as those from gene expression array.

4. Discussion

4.1. Shortest Path Genes in Nonsmall Cell Lung Cancer (NSCLC). We identified 25 shortest path genes in NSCLC and 38 shortest path genes in SCLC with a permutation P value less than 0.05. Intriguingly the top five shortest path genes in NSCLC are also among the most significant genes in SCLC, while SCLC has several unique genes with large betweenness values. These may help to reveal the relationship between the two major types of lung cancer.

As in NSCLC, HSP90AA1 [25–27] has been well documented to be relevant to lung cancer. We focus on candidate genes with large betweenness values and discuss the potential relationship between them and lung cancer.

Estrogen receptor 1 (ESR1) belongs to the nuclear steroid hormone receptor superfamily which acts as ligand-dependent, sequence-specific transcription factors and regulates the expression of genes involved in signal transduction, cell-cycle control, and cell survival [28]. Previous evidence showed that the proportion of never smokers among women with lung cancer is higher compared with men. Hypermethylation of ESR1 was reported to be detected only in lung tumors, but not in normal lung tissues, with a higher frequency being found in male patients than in female patients [29]. These all indicated ESR1 as a prognostic factor in lung cancer and as a potential target of hormone therapy.

ATP-binding cassette sub-family A member 1 (ABCA1) is a sulfonylurea-sensitive and cAMP-dependent anion transporter with critical impact on intracellular cholesterol transport. Cholesterol level increase has been found in cancers compared with normal tissue in many kinds of cancers [30], such as oral cancer [31]. Smith and Land demonstrated in colon cancer cells that ABCA1 had an anticancer activity in which deficiency allowed for increased mitochondrial cholesterol, inhibited release of mitochondrial cell death-promoting molecules, and facilitated cancer cell survival [32]. As abnormal metabolism is generally found in cancer, ABCA1 deserves further investigation with regard to its role in lung cancer.

Insulin receptor substrate 1 (IRS1) is an adaptor protein for insulin-like growth factor (IGF) signaling and is associated with IGF-stimulated proliferation [33]. It has been reported to be downregulated in NSCLC [34], and

its degradation accelerates lung tumor growth by upgrading interaction between the potent mitogen platelet-derived growth factor receptor (PDGFR) and phosphatidylinositol 3 kinase (PI3 K) [35]. Correspondingly, our study shows that the shortest path of IRS1 is designated more than 100 and is significant in both NSCLC genes and SCLC genes, indicating that it may play a crucial part in lung cancer development.

FDXR (NADPH: adrenodoxin oxidoreductase) serves as the first electron transfer protein in the mitochondrial P450 systems. FDXR is identified to be target of the p53 family. It could be induced in a p53-dependent way by DNA damage in cells and participated in p53-mediated apoptosis via generating oxidative stress in mitochondria [36, 37]. Owing to the significance of p53 in apoptosis during tumorigenesis, the contribution of FDXR to lung cancer is worthy of further elucidation.

4.2. Shortest Path Genes in Small Cell Lung Cancer (SCLC). The KEGG pathway enrichment analysis shows that there is a distinct group of shortest path genes in SCLC compared with NSCLC. These are the extracellular-matrix- (ECM-) related genes (Table 1). This coincides with the KEGG pathway enrichment analysis result of known SCLC pathway genes. ECM surrounds SCLC cells and includes collagen IV, tenascin, fibronectin, and laminin. Cell surface receptor integrins interact with ECM components and numerous signal transduction pathways which play important roles in cell cycle regulation, apoptosis, and so on and thus promote cancer cell proliferation [38]. Hodkinson et al. found that ECM can inhibit the caspase-3 activation and subsequent cell apoptosis induced by etoposide via stimulating phosphatidyl inositol 3-kinase- (PI3K-) signaling pathway in SCLC cells in a ITGB1/PI3K-dependent way [39]. Choi et al. demonstrated that downregulation of the phosphorylation activity of ILK (integrin-linked kinase) by single deletion of ILK protein itself or deletion of ITGB4/ILK complex could suppress the invasion of ovarian cancer [40]. Other studies also demonstrate that the intracellular signals activated by ECM components account for the high metastasis potential and drug resistance of SCLC [41]. In this work, we found that collagen IV members COL4A5 and COL4A3, integrin members ITGA1, ITGB4 and ITGA4, and linked kinase ILK all have a betweenness of more than 80 and a P value < 0.05, all of which may indicate their crucial roles in SCLC.

Forkhead box protein M1 (FOXM1) is a transcription factor regulating cell proliferation and DNA damage repair [42, 43]. Research shows that it could be phosphorylated by MAPK (ERK) kinase [44] and then activate the expression of a number of cell-cycle-related genes which are crucial for DNA replication and mitotic division in the Ras-mitogen-activated protein-kinase- (MAPK-) signaling pathway, such as cyclin A2, cyclin B1, Aurora B kinase, Cdc25B phosphatase- and Polo-like kinase1 [45]. Additionally, the protein level of FXOM1 has been found increased in prostate adenocarcinomas [46], infiltrating ductal breast carcinomas [47], basal cell carcinomas [48], intrahepatic cholangiocarcinomas [49], and in many other solid tumors [50]. A study by Kim et al. [51] showed that in human NSCLC Foxm1 protein is overexpressed and promotes tumor cells proliferation during the development of NSCLC. These all indicate that FXOM1 may play an import role in SCLC as well.

Table 4: The functional similarity between identified lung cancer genes and 742 cancer genes.

	742 cancer genes
1825 NSCLC genes from array	0.4314*
1063 SCLC genes from array	0.4845*
25 NSCLC genes from our study	0.5554*
38 SCLC genes from our study	0.6919*

*Pearson correlation coefficient of functional profiles.

Immunoglobulin-binding protein 1 (IGBP1) was formerly identified as a signal transduction molecule with a surface IgM receptor. More recently, it has been shown to regulate the phosphatase catalytic activity of protein phosphatase 2A (PP2A) [52]. PP2A is composed of a majority of cellular serine/threonine phosphatases [53] and regulates a number of important cellular processes, such as cell cycle transition, apoptosis, transcription, translation, autophagy [54], and cell transformation [55]. IGBP1 directly interacts with the catalytic subunit of PP2A [56], and this interaction leads to an antiapoptosis function. Recent studies show that, in carcinogen-transformed human cells and primary human cancers such as primary lung cancers, primary hepatocellular carcinomas and primary breast cancers, the expression level of IGBP1 is upregulated, [57]. Sakashita et al. found its overexpression in small cell adenocarcinomas [58], and Li et al. found that in a lung adenocarcinoma cell line the interaction of IGBP1 and Lactoferrin could induce cell apoptosis [59], implying IGBP1 to be a candidate target for SCLC therapy.

4.3. Functional Similarities between Candidate Genes and Known Cancer Genes. In order to compare the functional similarities between our candidate genes and the 742 known cancer genes, their functional profiles were constructed using the −log10 of the hypergeometric test P value on Gene Ontology (GO) terms [20, 21]. Then the Pearson correlation coefficient of their functional profiles was calculated [20, 21]. The functional similarities of five gene sets are shown in Table 4. All five gene sets can be found in additional file S3. Our 25 NSCLC (0.5554) and 38 SCLC (0.6919) candidate genes both had greater functional similarity with the cancer genes than the NSCLC (0.43139) and SCLC (0.48451) genes identified from gene expression profiles. It is suggested that our way is more efficient in identifying cancer-related genes.

5. Conclusion

In this study, we propose a computational method based on a protein-protein interaction network to identify cancer-related genes. We applied this method to lung cancer to find the shortest paths between 54 NSCLC and 84 SCLC genes in the protein-protein interaction network constructed based on STRING data and selected the 25 and 38 genes with a significant P value for NSCLC and SCLC, respectively.

Analysis of these shortest path genes indicates that some of these genes, such as ESR1, FDXR, ABCA1, IRS1, HSP90AA1, FOXM1, and IGBP1 are related to lung cancer. In addition, the candidate genes of lung cancer identified in our study contain more cancer genes than those identified from gene expression profiles. Moreover, it is revealed that our candidate genes have greater functional similarity with the cancer genes than those identified from gene expression profiles. These candidate genes may be worth experiment validation and further research. It is expected that this method is useful in predicting novel cancer-related genes and has widespread use in cancer research.

Authors' Contribution

B.-Q. Li and J. You contributed equally to this work.

Acknowledgments

This work was supported by grants from National Basic Research Program of China (2011CB510102 and 2011CB510101), Innovation Program of Shanghai Municipal Education Commission (12ZZ087), and the Grant of "The First-class Discipline of Universities in Shanghai."

References

[1] R. Siegel, D. Naishadham, and A. Jemal, "Cancer statistics," *A Cancer Journal for Clinicians*, vol. 62, pp. 10–29, 2012.

[2] J. P. van Meerbeeck, D. A. Fennell, and De Ruysscher, "DKM Small-cell lung cancer," *The Lancet*, vol. 378, pp. 1741–1755.

[3] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh, "Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps," *Bioinformatics*, vol. 21, supplement 1, pp. i302–i310, 2005.

[4] D. Szklarczyk, A. Franceschini, M. Kuhn et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, no. 1, pp. D561–D568, 2011.

[5] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular Systems Biology*, vol. 3, article 88, 2007.

[6] P. Bogdanov and A. K. Singh, "Molecular function prediction using neighborhood features," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 2, pp. 208–217, 2010.

[7] Y. A. Kourmpetis, A. D. van Dijk, M. C. Bink, R. C. van Ham, and C. J. ter Braak, "Bayesian Markov Random Field analysis for protein function prediction based on network data," *PLoS One*, vol. 5, article e9293, 2010.

[8] K. L. Ng, J. S. Ciou, and C. H. Huang, "Prediction of protein functions based on function-function correlation relations," *Computers in Biology and Medicine*, vol. 40, no. 3, pp. 300–305, 2010.

[9] U. Karaoz, T. M. Murali, S. Letovsky et al., "Whole-genome annotation by using evidence integration in functional-linkage networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2888–2893, 2004.

[10] S. Letovsky and S. Kasif, "Predicting protein function from protein/protein interaction data: a probabilistic approach," *Bioinformatics*, vol. 19, supplement 1, pp. i197–i204, 2003.

[11] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[12] M. Kanehisa and Subramaniam, "The KEGG database," *Novartis Foundation Symposium*, vol. 247, pp. 91–103, 2002.

[13] S. Kastner, T. Voss, S. Keuerleber, C. Glickel, M. Freissmuth, and W. Sommergruber, "Expression of G protein-coupled receptor 19 in human lung cancer cells is triggered by entry into S-phase and supports G(2)-m cell-cycle progression," *Molecular Cancer Research*, vol. 10, pp. 1343–1358, 2012.

[14] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.

[15] S. Zhang, "A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance," *BMC Bioinformatics*, vol. 8, article 230, 2007.

[16] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society Series B*, vol. 57, pp. 289–300, 1995.

[17] S. H. Nagaraj and A. Reverter, "A Boolean-based systems biology approach to predict novel genes associated with cancer: application to colorectal cancer," *BMC Systems Biology*, vol. 5, article 35, 2011.

[18] J. L. Huret, P. Dessen, and A. Bernheim, "Atlas of genetics and cytogenetics in oncology and haematology, year 2003," *Nucleic Acids Research*, vol. 31, no. 1, pp. 272–274, 2003.

[19] T. S. K. Prasad, R. Goel, K. Kandasamy et al., "Human protein reference database—2009 update," *Nucleic Acids Research*, vol. 37, no. 1, pp. D767–D772, 2009.

[20] B. Q. Li, J. Zhang, T. Huang, L. Zhang, and Y. D. Cai, "Identification of retinoblastoma related genes with shortest path in a protein-protein interaction network," *Biochimie*, vol. 94, pp. 1910–1917, 2012.

[21] B. Q. Li, T. Huang, L. Liu, Y. D. Cai, and K. C. Chou, "Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network," *PLoS One*, vol. 7, Article ID e33393, 2012.

[22] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal Complex Systems*, 2006.

[23] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[24] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

[25] L. Whitesell and S. L. Lindquist, "HSP90 and the chaperoning of cancer," *Nature Reviews Cancer*, vol. 5, no. 10, pp. 761–772, 2005.

[26] J. Trepel, M. Mollapour, G. Giaccone, and L. Neckers, "Targeting the dynamic HSP90 complex in cancer," *Nature Reviews Cancer*, vol. 10, no. 8, pp. 537–549, 2010.

[27] M. Senju, N. Sueoka, A. Sato et al., "Hsp90 inhibitors cause G2/M arrest associated with the reduction of Cdc25C and Cdc2 in lung cancer cell lines," *Journal of Cancer Research and Clinical Oncology*, vol. 132, no. 3, pp. 150–158, 2006.

[28] J. Frasor, J. M. Danes, B. Komm, K. C. N. Chang, C. Richard Lyttle, and B. S. Katzenellenbogen, "Profiling of estrogen up- and

down-regulated gene expression in human breast cancer cells: insights into gene networks and pathways underlying estrogenic control of proliferation and cell phenotype," *Endocrinology*, vol. 144, no. 10, pp. 4562–4574, 2003.

[29] J. C. Lai, Y. W. Cheng, H. L. Chiou, M. F. Wu, C. Y. Chen, and H. Lee, "Gender difference in estrogen receptor alpha promoter hypermethylation and its prognostic value in non-small cell lung cancer," *International Journal of Cancer*, vol. 117, no. 6, pp. 974–980, 2005.

[30] Y. Yoshioka, J. Sasaki, M. Yamamoto et al., "Quantitation by 1H-NMR of dolichol, cholesterol and choline-containing lipids in extracts of normal and phathological thyroid tissue," *NMR in Biomedicine*, vol. 13, pp. 377–383, 2000.

[31] K. Kolanjiappan, C. R. Ramachandran, and S. Manoharan, "Biochemical changes in tumor tissues of oral cancer patients," *Clinical Biochemistry*, vol. 36, no. 1, pp. 61–65, 2003.

[32] B. Smith and H. Land, "Anticancer activity of the cholesterol exporter ABCA1 gene," *Cell Reports*, vol. 2, pp. 580–590.

[33] D. Yee, "Targeting insulin-like growth factor pathways," *British Journal of Cancer*, vol. 94, no. 4, pp. 465–468, 2006.

[34] C. H. Han, J. Y. Cho, J. T. Moon et al., "Clinical significance of insulin receptor substrate-I down-regulation in non-small cell lung cancer," *Oncology Reports*, vol. 16, no. 6, pp. 1205–1210, 2006.

[35] A. M. Houghton, D. M. Rzymkiewicz, H. Ji et al., "Neutrophil elastase-mediated degradation of IRS-1 accelerates lung tumor growth," *Nature Medicine*, vol. 16, no. 2, pp. 219–223, 2010.

[36] G. Liu and X. Chen, "The ferredoxin reductase gene is regulated by the p53 family and sensitizes cells to oxidative stress-induced apoptosis," *Oncogene*, vol. 21, no. 47, pp. 7195–7204, 2002.

[37] P. M. Hwang, F. Bunz, J. Yu et al., "Ferredoxin reductase affects p53-dependent, 5-fluorouracil-induced apoptosis in colorectal cancer cells," *Nature Medicine*, vol. 7, no. 11, pp. 1111–1117, 2001.

[38] A. J. Ridley, M. A. Schwartz, K. Burridge et al., "Cell migration: integrating signals from front to back," *Science*, vol. 302, no. 5651, pp. 1704–1709, 2003.

[39] P. S. Hodkinson, T. Elliott, W. S. Wong et al., "ECM overrides DNA damage-induced cell cycle arrest and apoptosis in small-cell lung cancer cells through β1 integrin-dependent activation of PI3-kinase," *Cell Death and Differentiation*, vol. 13, no. 10, pp. 1776–1788, 2006.

[40] Y. P. Choi, B. G. Kim, M. Q. Gao, S. Kang, and N. H. Cho, "Targeting ILK and β4 integrin abrogates the invasive potential of ovarian cancer," *Biochemical and Biophysical Research Communications*, vol. 427, no. 3, pp. 642–648, 2012.

[41] T. Sethi, R. C. Rintoul, S. M. Moore et al., "Extracellular matrix proteins protect small cell lung cancer cells against apoptosis: a mechanism for small cell lung cancer growth and drug resistance in vivo," *Nature Medicine*, vol. 5, no. 6, pp. 662–668, 1999.

[42] Y. Tan, P. Raychaudhuri, and R. H. Costa, "Chk2 mediates stabilization of the FoxM1 transcription factor to stimulate expression of DNA repair genes," *Molecular and Cellular Biology*, vol. 27, no. 3, pp. 1007–1016, 2007.

[43] Z. Fu, L. Malureanu, J. Huang et al., "Plk1-dependent phosphorylation of FoxM1 regulates a transcriptional programme required for mitotic progression," *Nature Cell Biology*, vol. 10, no. 9, pp. 1076–1082, 2008.

[44] R. Y. M. Ma, T. H. K. Tong, A. M. S. Cheung, A. C. C. Tsang, W. Y. Leung, and K. M. Yao, "Raf/MEK/MAPK signaling stimulates the nuclear translocation and transactivating activity of FOXM1c," *Journal of Cell Science*, vol. 118, no. 4, pp. 795–806, 2005.

[45] R. H. Costa, V. V. Kalinichenko, M. L. Major, and P. Raychaudhuri, "New and unexpected: forkhead meets ARF," *Current Opinion in Genetics and Development*, vol. 15, no. 1, pp. 42–48, 2005.

[46] T. V. Kalin, I. C. Wang, T. J. Ackerson et al., "Increased levels of the FoxM1 transcription factor accelerate development and progression of prostate carcinomas in both TRAMP and LADY transgenic mice," *Cancer Research*, vol. 66, no. 3, pp. 1712–1720, 2006.

[47] D. R. Wonsey and M. T. Follettie, "Loss of the forkhead transcription factor FoxM1 causes centrosome amplification and mitotic catastrophe," *Cancer Research*, vol. 65, no. 12, pp. 5181–5189, 2005.

[48] M. T. Teh, S. T. Wong, G. W. Neill, L. R. Ghali, M. P. Philpott, and A. G. Quinn, "FOXM1 is a downstream target of Gli1 in basal cell carcinomas," *Cancer Research*, vol. 62, no. 16, pp. 4773–4780, 2002.

[49] K. Obama, K. Ura, M. Li et al., "Genome-wide analysis of gene expression in human intrahepatic cholangiocarcinoma," *Hepatology*, vol. 41, no. 6, pp. 1339–1348, 2005.

[50] C. Pilarsky, M. Wenzig, T. Specht, H. D. Saeger, and R. Grützmann, "Identification and validation of commonly overexpressed genes in solid tumors by comparison of microarray data," *Neoplasia*, vol. 6, no. 6, pp. 744–750, 2004.

[51] I. M. Kim, T. Ackerson, S. Ramakrishna et al., "The Forkhead Box m1 transcription factor stimulates the proliferation of tumor cells during development of lung cancer," *Cancer Research*, vol. 66, no. 4, pp. 2153–2161, 2006.

[52] M. Kong, D. Ditsworth, T. Lindsten, and C. B. Thompson, "α4 is an essential regulator of PP2A phosphatase activity," *Molecular Cell*, vol. 36, no. 1, pp. 51–60, 2009.

[53] V. Janssens and J. Goris, "Protein phosphatase 2A: a highly regulated family of serine/threonine phosphatases implicated in cell growth and signalling," *Biochemical Journal*, vol. 353, no. 3, pp. 417–439, 2001.

[54] T. Yorimitsu, C. He, K. Wang, and D. J. Klionsky, "Tap42-associated protein phosphatase type 2A negatively regulates induction of autophagy," *Autophagy*, vol. 5, no. 5, pp. 616–624, 2009.

[55] J. D. Arroyo and W. C. Hahn, "Involvement of PP2A in viral and cellular transformation," *Oncogene*, vol. 24, no. 52, pp. 7746–7755, 2005.

[56] K. Murata, J. Wu, and D. L. Brautigan, "B cell receptor-associated protein α4 displays rapamycin-sensitive binding directly to the catalytic subunit of protein phosphatase 2A," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 20, pp. 10624–10629, 1997.

[57] L. P. Chen, Y. D. Lai, D. C. Li et al., "α4 is highly expressed in carcinogen-transformed human cells and primary human cancers," *Oncogene*, vol. 30, no. 26, pp. 2943–2953, 2011.

[58] S. Sakashita, D. Li, N. Nashima et al., "Overexpression of immunoglobulin (CD79a) binding protein1 (IGBP-1) in small lung adenocarcinomas and its clinicopathological significance," *Pathology International*, vol. 61, no. 3, pp. 130–137, 2011.

[59] D. Li, S. Sakashita, Y. Morishita et al., "Binding of lactoferrin to IGBP1 triggers apoptosis in a lung adenocarcinoma cell line," *Anticancer Research*, vol. 31, no. 2, pp. 529–534, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 846387, 9 pages
http://dx.doi.org/10.1155/2013/846387

Research Article

Molecular Profiling Predicts the Existence of Two Functionally Distinct Classes of Ovarian Cancer Stroma

Loukia N. Lili,[1] Lilya V. Matyunina,[1] L. DeEtte Walker,[1] Benedict B. Benigno,[2] and John F. McDonald[1,2]

[1] *Integrated Cancer Research Center, School of Biology and Parker H. Petit Institute of Bioengineering and Bioscience, Georgia Institute of Technology, 315 Ferst Dr., Atlanta, GA 30332, USA*
[2] *Ovarian Cancer Institute, 960 Johnson Ferry Road, Suite 130, Atlanta, GA 30342, USA*

Correspondence should be addressed to John F. McDonald; mcgene@gatech.edu

Received 6 March 2013; Revised 16 April 2013; Accepted 18 April 2013

Academic Editor: Tao Huang

Although stromal cell signaling has been shown to play a significant role in the progression of many cancers, relatively little is known about its importance in modulating ovarian cancer development. The purpose of this study was to investigate the process of stroma activation in human ovarian cancer by molecular analysis of matched sets of cancer and surrounding stroma tissues. RNA microarray profiling of 45 tissue samples was carried out using the Affymetrix (U133 Plus 2.0) gene expression platform. Laser capture microdissection (LCM) was employed to isolate cancer cells from the tumors of ovarian cancer patients (Cepi) and matched sets of surrounding cancer stroma (CS). For controls, ovarian surface epithelial cells (OSE) were isolated from the normal (noncancerous) ovaries and normal stroma (NS). Hierarchical clustering of the microarray data resulted in clear separations between the OSE, Cepi, NS, and CS samples. Expression patterns of genes encoding signaling molecules and compatible receptors in the CS and Cepi samples indicate the existence of two subgroups of cancer stroma (CS) with different propensities to support tumor growth. Our results indicate that functionally significant variability exists among ovarian cancer patients in the ability of the microenvironment to modulate cancer development.

1. Introduction

The epithelial cells of the ovary interact with the cells of the surrounding microenvironment in order to regulate tissue homeostasis. Morphologically, the normal ovarian epithelial cells form a flat-to-cuboidal monolayer supported by a basement membrane. Cells located below this basement membrane are composed of various cell types collectively referred to as stromal cells. The most common types of stromal cells are fibroblasts, pericytes, endothelial cells, and various immune and inflammatory cells. Stromal and epithelial cells communicate through the secretion and binding of growth factors and other signaling molecules that promote reciprocal cellular responses appropriate for coordinated cell functions, for example, those required for the replication of ovarian surface epithelial cells following ovulation [1–3].

During cancer progression, genetic and epigenetic alterations lead to changes in the morphology and behavior of both epithelial and stromal cells by disrupting the tissue architecture and by interfering with signaling mechanisms. For example, signaling changes in a wide variety of developing cancer cells have been shown to result in the disruption of tissue homeostasis by inducing extracellular matrix (ECM) turnover, basement membrane disassociation, and increased stromal cell proliferation [1, 4].

Despite the well-documented role of stromal cell signaling in cancer progression, relatively few studies have been focused specifically on epithelial ovarian cancer-stromal interactions (EOC-SI). Previously reported studies on EOC-SI have focused on particular stromal components [5, 6], on specific cell lines [7], and/or employed in-house fabricated microarrays of limited scope [8]. We report here the results of a study of EOC-SI using high-throughput gene expression (microarray) analysis of normal ovarian surface epithelial cells and cells captured from normal stroma, cancer epithelia, and cancer stroma using laser capture microdissection

Table 1: Patient samples used in this study.

Patient ID	Age at time of surgery	Tissue for microarray	Histopathology	Stage	Grade
460	65	OSE	WNL (within normal limits)	N/A	N/A
552	41	OSE	WNL	N/A	N/A
563	66	OSE	WNL	N/A	N/A
567	78	OSE	WNL	N/A	N/A
434	41	OSE/NS	WNL	N/A	N/A
437	54	OSE/NS	WNL	N/A	N/A
440	50	OSE/NS	WNL	N/A	N/A
448	63	OSE/NS	WNL	N/A	N/A
452	51	OSE/NS	WNL	N/A	N/A
463	48	OSE/NS	WNL	N/A	N/A
470	44	OSE/NS	WNL	N/A	N/A
475	63	OSE/NS	WNL	N/A	N/A
317	59	Cepi	Serous adenocarcinoma	Ic	3
489	48	Cepi	Serous adenocarcinoma	IV	3
528	66	Cepi	Serous adenocarcinoma	IIIc	3
537	64	Cepi	Serous adenocarcinoma	IIIa	2
542	61	Cepi	Serous adenocarcinoma	IV	3
551	59	Cepi	Serous adenocarcinoma	IIIc/IV	3
588	71	Cepi	Serous adenocarcinoma	IIIc	2
606	54	Cepi	Serous adenocarcinoma	IIIa	3
617	64	Cepi	Serous adenocarcinoma	IIIc	2
620	62	Cepi	Serous adenocarcinoma	III/IV	3
651	46	Cepi	Serous adenocarcinoma	IIIb/IIIc	3
183	66	Cepi/CS	Serous adenocarcinoma	III	2
369	52	Cepi/CS	Serous adenocarcinoma	IIIc	2
229	58	Cepi/CS	Serous adenocarcinoma	IIIc	3
242	63	Cepi/CS	Serous adenocarcinoma	IIIb	3
336	63	Cepi/CS	Serous adenocarcinoma	Ic	3
367	56	Cepi/CS	Serous adenocarcinoma	II	3
413	49	Cepi/CS	Serous adenocarcinoma	IIb	3

(LCM). Our results reveal the existence of two categories of ovarian cancer stroma. Analysis of ligand-receptor patterns of gene expression indicates that one of these classes of cancer stroma may be more permissive and one more resistant to associated cancer cell growth.

2. Materials and Methods

2.1. Tissue Collection. Tissues were collected at Northside Hospital (Atlanta, GA, USA) under appropriate Institutional Review Board protocols. Following resection, the tumor tissues were placed in cryotubes and immediately (<1 minute) frozen in liquid nitrogen. Samples were transported on dry ice to Georgia Institute of Technology (Atlanta, GA, USA) and stored at −80°C. All tissues were examined, and diagnoses were made by a pathologist. The histopathology for each sample is listed in Table 1.

For each of the cancer tissue samples, 7 mm frozen sections were cut from samples embedded in cryomatrix (Shandon) and attached to uncharged microscope slides. Immediately following dehydration and staining (HistoGene, LCM Frozen Section Staining Kit, Arcturus), slides were processed in an Autopix (Arcturus) instrument for laser capture microdissection (LCM) of cancer epithelial cells (CEPI), cancer stroma (CS), and normal stroma (NS) using CapSure Macro-LCM Caps (Arcturus). Approximately 30,000 cells were collected from each of the samples. Normal ovarian surface epithelial (OSE) cells were also collected from normal ovaries at the time of surgery by light brushing using a Cytobrush Plus (Medscand), immediately stabilized in RNAlater (Ambion), and subsequently stored at −20°C. Microscopic examination of all collected cells was carried out to confirm the integrity and purity of the samples.

2.2. RNA Extraction and Amplification. PicoPure RNA Isolation Kit (Arcturus) protocols were followed for RNA extraction from the LCM cells on the Macro-LCM caps in 25 μL of extraction buffer. Normal OSE cells were pelleted from RNAlater; RNA was isolated with Trizol (Invitrogen) and purified with the PicoPure RNA Isolation Kit. RNA quality was verified for all samples on the Bioanalyzer RNA Pico Chip (Agilent Technologies).

Total RNA from the above extractions was processed using the RiboAmp HS Kit (Arcturus) in conjunction with the IVT Labeling Kit from Affymetrix, to produce an amplified, biotin-labeled mRNA suitable for hybridizing to GeneChip Human Genome U133 Plus 2.0 Arrays (Affymetrix) following manufacturer's recommendations.

2.3. Microarray Data Analysis. We generated 45 individual gene expression profiles from 12 OSE brushings and 18 Cepi, 8 NS, and 7 CS patient samples isolated by laser capture microdissection (LCM). Affymetrix CEL files were processed using the Affymetrix Expression Console (EC) Software Version 1.1 with the default MAS5.0 probeset normalization algorithm. The expression values from the 12 OSE, 18 Cepi, 8 NS, and 7 CS samples were $\log_2$ transformed and then averaged for each probeset across each sample type. The microarray data were deposited in the Gene Expression Omnibus (GSE38666).

Probesets (genes) with nearly constant expression values ($\log_2$ normalized) across samples (SD < 1) were excluded from further consideration. Of the 54,675 probesets on the U133 Plus 2.0 chip, 42,698 were thus retained. A four-way ANOVA was subsequently employed to identify genes significantly differentially expressed ($P \leq 0.001$) across the four sample groups (OSE, Cepi, NS, and CS). These 6,654 genes were employed in the initial clustering analysis.

A subsequent comparison among the CS samples (CS_1 and CS_2) alone was performed using a similar approach. Specifically, genes with nearly identical expression values (SD < 1) across CS_1 and CS_2 were discarded, and the remaining 38,972 genes were subjected to an unpaired t-test to identify those genes that were significantly differentially expressed between the CS_1 and CS_2 subgroups ($P \leq 0.001$, 88 genes).

All heat maps were generated using the UPGMA (unweighted average) clustering method and the Euclidean distance similarity measure.

2.4. Ligand-Receptor Compatibility Analysis. For the ligand-receptor compatibility analysis, probesets associated with no or marginal expression across all 45 samples were discarded resulting in 5,865 differentially expressed genes. The presence or absence of the expression in samples was determined using the Affymetrix default MAS 5.0 decision algorithm. The MAS 5.0 algorithm uses Tukey's biweight estimator to provide a robust mean signal value and the Wilcoxon's rank test to calculate the significance of the signal or P value and detection call (present, marginal, or absent) for each probeset. The P values upon which the presence-absence calls for each ligand and receptor are based are presented in the appropriate Tables 1–5.

3. Results

3.1. Hierarchical Clustering Establishes Two Distinct Classes of Stroma among the Ovarian Cancer Patient Samples. Forty-five gene expression profiles were generated from 12 OSE brushings and 18 Cepi, 8 NS, and 7 CS patient samples isolated by laser capture microdissection (LCM). The relevant histopathologies of these 45 samples are listed in Table 1. Expression analysis yielded 6,654 differentially expressed probesets among the four sample types (ANOVA, $P \leq 0.001$). Hierarchical clustering of these expression data resulted in clear separations between the OSE, Cepi, NS, and CS samples (Figure 1). Interestingly, the CS samples subdivided into two distinct groups. One (CS_1) was more closely associated with the NS samples, and the other (CS_2) was more closely associated with the Cepi samples.

One possibility is that the two subclasses of CS are simply a reflection of differential responses of stroma to molecular differences in the adjacent Cepi. If this were the case, we would expect to see a correlated substructure among the molecular profiles of the Cepi samples associated with the CS_1 and CS_2 subgroups. As shown in Figure 1, no such coordinated substructure pattern exists among the Cepi samples indicating that the two subclasses of CS are not merely a reflection of differential responses of the stroma to different Cepi subtypes.

As stated above, microscopic examination of LCM collected cells was carried out to validate the integrity of our samples. As a further confirmation, we conducted an additional computational analysis. In this analysis, probesets associated with no or marginal expression across all 45 samples were discarded resulting in 5,865 differentially expressed genes. If the reason for the presence of two classes of CS samples is that the CS_2 class was a mixture of stroma and invasive Cepi cells, the gene expression levels in the putative mixed cancer stromal class (CS_2) would be expected to lie within the range of the maximum and minimum expression levels of the NS and Cepi groups (i.e., avg(CS_2) < Min and avg(CS_2) > Max). Inconsistent with this prediction, we found that 2,342 or 40% (2,342/5,865) of the differentially expressed genes making up the CS_2 class displayed values outside the predicted range of the mixed cell types. The fact that 60% of the expression values lie within the predicted range is not indicative of contamination but rather of the fact that not all genes are significantly overexpressed in the stroma relative to cancer samples. Collectively our microarray results are consistent with the microscopic examination in demonstrating the absence of infiltrating Cepi cells in the cancer stroma samples.

3.2. Gene Expression Patterns Are Consistent with the Existence of Ligand-Receptor Interactions between Cepi and CS. The significance of the presence of two distinct classes of ovarian cancer stroma may involve differential interaction between these stromal and the adjacent cancer cells. To explore this possibility, we first examined the expression levels of genes encoding signaling ligands and compatible receptors in the CS and Cepi datasets.

Two lists were established from the 5,865 differentially expressed probesets across the OSE, Cepi, NS, and CS samples. One list is comprised of all differentially expressed gene probes (note that each gene may be represented by multiple, nonoverlapping probes) encoding secreted ligands (ligand list) and the other of all expressed gene probes encoding

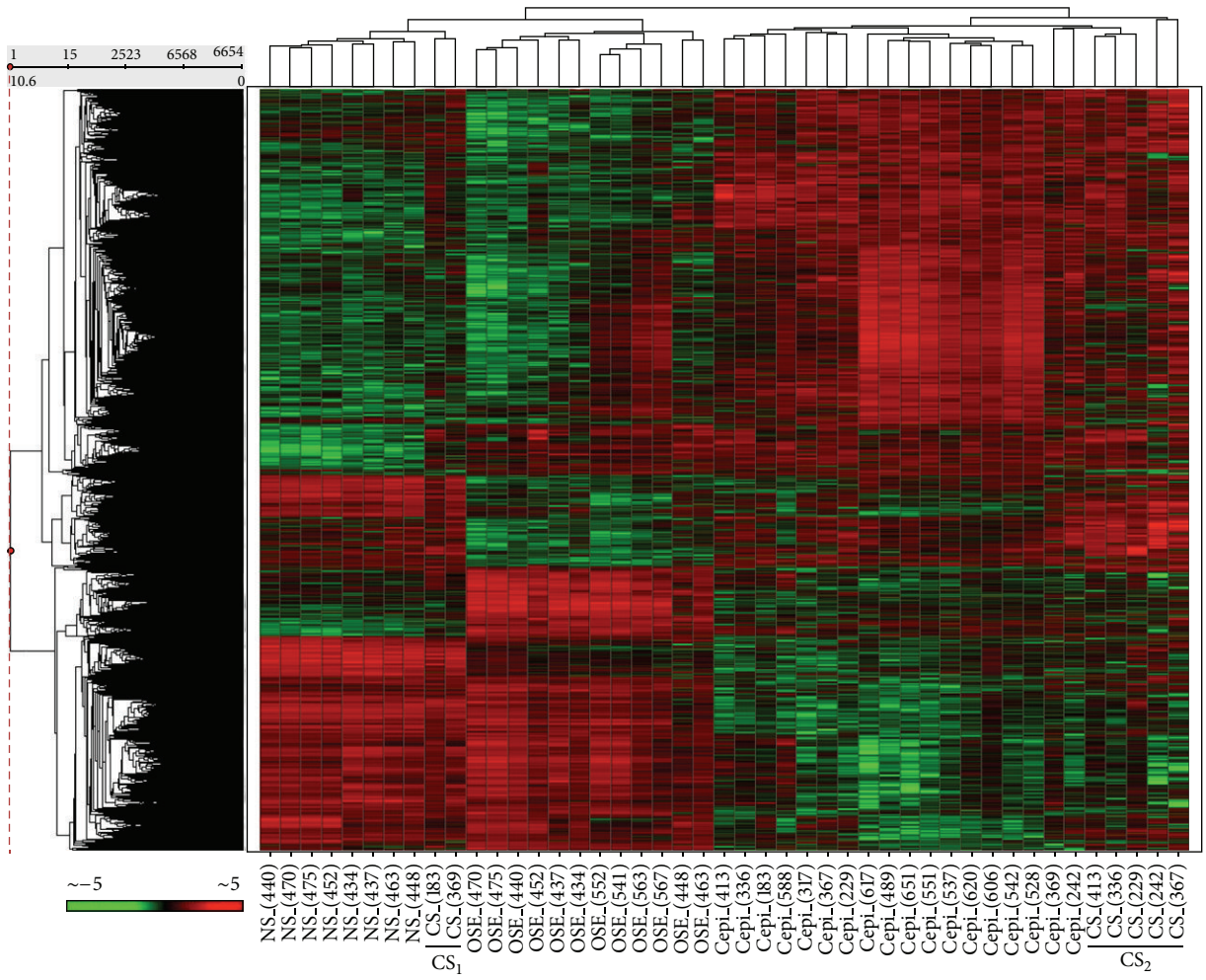


FIGURE 1: Hierarchical clustering of OSE, Cepi, NS, and CS expression profiles. The heat map was generated by z-score normalization of $\log_2$ expression values from Affymetrix HG U133 Plus 2.0. The results show that the OSE, Cepi, NS, and CS samples cluster into separate groups. The CS samples clustered into two distinct subgroups (CS_1 and CS_2).

surface receptors (receptor list) with documented binding affinity to the differentially expressed ligands (compatible ligands and receptors). The ligand list consists of 34 CS and 36 Cepi ligands while the receptor list is comprised of 20 Cepi and 21 CS receptors (Tables 2(a) and 2(b)).

We considered the expression of a ligand in CS (or Cepi) and its compatible receptor in Cepi (or CS) to be indicative of a potential CS-Cepi signaling interaction. Based on these criteria, we identified potential epithelial cancer-stroma signaling interactions (34 CS ligands and 20 Cepi receptors, see Table 2(a) and 36 Cepi ligands and 21 CS receptors, see Table 2(b)). Of these, there were 17 compatible pairs for both the CS ligands-Cepi receptors and Cepi ligands-CS receptors interactions (Table 3). Viewed from the perspective of individual genes (i.e., combining multiple probes of the same genes), there were 12 unique CS ligand-Cepi receptor pairs and 12 unique Cepi ligand-CS receptor pairs in our observed dataset (Table 4).

To determine if the observed coexpression of these 17 pairs of compatible ligands and receptors (probes) was greater than what would be expected by chance, we generated two lists. One list of observed data consisted of the expressed probes of the 17 CS ligands and 17 compatible Cepi receptors (Table 3(a)) and the other of the 17 expressed Cepi ligands and 17 CS receptors (Table 3(b)). A second list of random associations was generated using the same number of pairings as in the observed list (17 random pairs) and randomly selecting 17 pairs of ligands and receptors. One randomly selected CS (or Cepi) ligand from the pool of the 34 CS-(or 36 Cepi-) expressed ligands (Table 2(a)) was paired with one randomly selected Cepi (or CS) receptor from the pool of the 20 Cepi-(or 21 CS-) expressed receptors (Table 2(b)). These random associations were generated 100 times, and each time the number of biologically compatible ligand-receptor pairs arising by chance was counted. The number of biologically compatible interactions in the observed data (17) was then compared to the number of compatible interactions scored from the randomized associations using z-statistics. Two types of comparisons were performed, one for the pairs of CS ligands and Cepi receptors and another for the pairs of Cepi

Table 2: The 34 CS-expressed ligands with the 20 expressed Cepi receptors (a) and the 36 Cepi-expressed ligands with the 21 expressed CS receptors (b).

(a)

CS ligands		Cepi receptors	
Gene symbol	Probeset ID	Gene symbol	Probeset ID
***CXCL1	204470_at	****CXCR4	217028_at
*CXCL3	207850_at	**FGFR2	208228_s_at
***CXCL9	203915_at	***FGFR3	204379_s_at
***CXCL10	204533_at	**MET	203510_at
***CXCL11	210163_at	*TGFBR2	207334_s_at
***CXCL12	209687_at	***TGFBR2	208944_at
***CXCL12	203666_at	**TGFBR3	204731_at
***CXCL13	205242_at	***TGFBR3	226625_at
**CXCL16	223454_at	***PDGFRA	203131_at
CXCL17	226960_at	*PDGFRA	1554828_at
*FGF1	205117_at	***IL1R1	202948_at
*FGF2	204422_s_at	*IL1R1	215561_s_at
***FGF7	1554741_s_at	*IL1R2	205403_at
*FGF9	239178_at	**IL7R	226218_at
**FGF9	206404_at	**IL10RA	204912_at
***FGF13	205110_s_at	**FZD1	204451_at
*HGF	210997_at	**FZD2	210220_at
***IGF1	209540_at	**FZD7	203705_s_at
*IGF2	202409_at	***FZD7	203706_s_at
*TGFA	205016_at	**FZD10	219764_at
***TGFB2	209909_s_at		
*PDGFA	205463_s_at		
***PDGFD	219304_s_at		
*IL7	206693_at		
***IL15	205992_s_at		
**IL16	209828_s_at		
**IL17D	227401_at		
**IL18	206295_at		
**WNT2B	206458_s_at		
*WNT7A	210248_at		
***WNT5A	213425_at		
***VEGFA	210512_s_at		
*VEGFA	210513_s_at		
*VEGFA	211527_x_at		

(b)

Cepi ligands		CS receptors	
Gene symbol	Probeset ID	Gene symbol	Probeset ID
***CXCL1	204470_at	****CXCR4	217028_at
**CXCL3	207850_at	**FGFR2	208228_s_at
***CXCL9	203915_at	*FGFR3	204379_s_at
***CXCL10	204533_at	IL12RB1	1552584_at
***CXCL11	210163_at	***IL1R1	202948_at
*CXCL12	203666_at	***TGFBR2	208944_at
*CXCL12	209687_at	***TGFBR3	204731_at
**CXCL13	205242_at	***TGFBR3	226625_at
**CXCL16	223454_at	***PDGFRA	203131_at

(b) Continued.

Cepi ligands		CS receptors	
Gene symbol	Probeset ID	Gene symbol	Probeset ID
CXCL17	226960_at	**PDGFRA	215305_at
*FGF1	205117_at	***MET	203510_at
***FGF9	206404_at	**IL1R1	215561_s_at
***FGF9	239178_at	*IL1R2	205403_at
*FGF11	227271_at	**IL7R	226218_at
***FGF18	231382_at	**IL10RA	204912_at
*FGF18	211029_x_at	*IL21R	221658_s_at
*FGF18	206987_x_at	**FZD1	204451_at
**FGF18	214284_s_at	**FZD2	210220_at
**TGFA	205016_at	***FZD7	203705_s_at
**TGFB2	209909_s_at	**FZD7	203706_s_at
**PDGFA	205463_s_at	***FZD10	219764_at
***PDGFD	219304_s_at		
***IGF1	209540_at		
*IL7	206693_at		
*IL1B	39402_at		
***IL15	205992_s_at		
**IL18	206295_at		
**WNT2	205648_at		
**WNT2B	206458_s_at		
***WNT5A	213425_at		
*WNT7A	210248_at		
**WNT11	206737_at		
***VEGFA	210512_s_at		
***VEGFA	210513_s_at		
*VEGFA	211527_x_at		
**VEGFA	212171_x_at		

Significance of detection calls: $^{*}P \leq 0.05$, $^{**}P \leq 0.005$, and $^{***}P \leq 0.0005$.

ligands and CS receptors. For both comparisons, the observed number of biologically compatible ligand-receptor pairs was significantly greater than what is expected by chance (CS ligands-Cepi receptors z-score = −4.68, $P \leq 0.0002$; Cepi ligands-CS receptors z-score = −4.35, $P \leq 0.0002$). Thus, the observed coexpression of pairs of compatible ligands and receptors is biologically significant.

3.3. Specific Ligand-Receptor Pairs between Cepi and CS Show Differential Gene Expression in the Two CS Classes. Of the 24 compatible pairs of ligand- and receptor-encoding genes listed in Table 4, most display similar expression patterns between CS_1 and CS_2. However, 6 of the ligand and receptor pairs display differential patterns of expression between the two groups of CS suggesting that CS_2 may be a more conducive microenvironment for tumor growth (Table 5). For example, the FGF2 ligand, a documented inhibitor of tumor growth [9], is expressed in NS and in CS_1 but not in CS_2. Since a compatible receptor of this inhibitor (FGFR3) is expressed in Cepi, CS_2 may be a more conducive microenvironment for tumor growth than CS_1. The interleukin-7 (IL7) ligand has been previously implicated as an inducer of tumor growth

Table 3: The expressed, compatible ligands and receptors as potential interactions between the Cepi and the CS samples from Tables 2(a) and 2(b).

(a)

CS ligands	Probesets	Compatible Cepi receptors	Probesets
***CXCL12	203666_at	***CXCR4	217028_at
***CXCL12	209687_at	***CXCR4	217028_at
*FGF1	205117_at	**FGFR2	208228_s_at
*FGF1	205117_at	*FGFR3	204379_s_at
*FGF2	204422_s_at	*FGFR3	204379_s_at
**FGF9	206404_at	*FGFR3	204379_s_at
*FGF9	239178_at	*FGFR3	204379_s_at
*HGF	210997_at	*MET	203510_at
*PDGFA	205463_s_at	*PDGFRA	1554828_at
*PDGFA	205463_s_at	***PDGFRA	203131_at
***TGFB2	209909_s_at	*TGFBR2	207334_s_at
***TGFB2	209909_s_at	***TGFBR2	208944_at
*WNT2	205648_at	*FZD2	210220_at
**WNT2B	206458_s_at	*FZD10	219764_at
*WNT7A	210248_at	*FZD7	203705_s_at
*WNT7A	210248_at	**FZD7	203706_s_at
*IL7	206693_at	**IL7R	226218_at

(b)

Cepi ligands	Probesets	Compatible CS receptors	Probesets
*CXCL12	203666_at	***CXCR4	217028_at
*CXCL12	209687_at	***CXCR4	217028_at
*FGF1	205117_at	**FGFR2	208228_s_at
*FGF1	205117_at	*FGFR3	204379_s_at
*FGF9	206404_at	*FGFR3	204379_s_at
FGF9	239178_at	*FGFR3	204379_s_at
*PDGFA	205463_s_at	***PDGFRA	203131_at
*PDGFA	205463_s_at	*PDGFRA	215305_at
*TGFB2	209909_s_at	***TGFBR2	208944_at
*WNT2	205648_at	*FZD2	210220_at
*WNT2B	206458_s_at	***FZD10	219764_at
*WNT7A	210248_at	***FZD7	203705_s_at
*WNT7A	210248_at	**FZD7	203706_s_at
*IL1B	39402_at	***IL1R1	202948_at
*IL1B	39402_at	*IL1R2	205403_at
*IL1B	39402_at	*IL1R1	215561_s_at
*IL7	206693_at	*IL7R	226218_at

Significance of detection calls: $^{*}P \leq 0.05$, $^{**}P \leq 0.005$, and $^{***}P \leq 0.0005$.

in lymphoblastic leukemia [10], prostate cancer [11], breast cancer [12], and colorectal cancer [13]. IL7 is expressed in CS_2 but not in CS_1, again suggesting that CS_2 may be a more conducive microenvironment for tumor growth than CS_1.

Table 4: The unique compatible ligands and receptors as potential interactions between the Cepi and the CS samples when multiple probes from Tables 3(a) and 3(b) are combined.

CS ligands	Compatible Cepi receptors	Cepi ligands	Compatible CS receptors
***CXCL12	***CXCR4	*CXCL12	***CXCR4
*FGF1	**FGFR2	*FGF1	**FGFR2
*FGF1	*FGFR3	*FGF1	*FGFR3
*FGF2	*FGFR3	**FGF9	*FGFR3
**FGF9	*FGFR3	**PDGFA	***PDGFRA
*HGF	*MET	**TGFB2	***TGFBR2
*PDGFA	***PDGFRA	**IL7	*IL7R
***TGFB2	***TGFBR2	**IL1B	*IL1R1
*IL7	**IL7R	**IL1B	*IL1R2
*WNT2	*FZD2	**WNT2	*FZD2
**WNT2B	*FZD10	**WNT2B	***FZD10
*WNT7A	**FZD7	**WNT7A	***FZD7

Significance of detection calls: $^{*}P \leq 0.05$, $^{**}P \leq 0.005$, and $^{***}P \leq 0.0005$.

The well-documented cancer-inducing ligands FGF1 and FGF9 [14–16] are both highly expressed in Cepi. The fact that the compatible FGFR3 receptor is expressed in CS_2 but not in CS_1 again suggests that CS_2 is a more favorable microenvironment for ovarian cancer growth than CS_1.

The *WNT* family of genes is involved in a variety of developmental processes, and aberrant expression of various members of *WNT* genes has been implicated in cancer [17]. For example, WNT7A is a ligand present in the extracellular matrix that participates in the sexual development of the Mullerian ducts [18]. Recent *in vivo* mouse studies suggest that WNT7A is an inducer of ovarian cancer growth [19]. Consistent with this interpretation, WNT7A has recently been identified as a potential early stage biomarker of human ovarian cancer [20]. The fact that WNT7A is expressed in CS_2 but not in CS_1 is also consistent with the hypothesis that CS_2 may be a more conducive microenvironment for ovarian cancer growth than CS_1.

A second member of the WNT family, WNT2B, is expressed in CS_1 but not CS_2 suggesting, contrary to what is presented above, that CS_1 may be more permissive for cancer growth. However, the fact that WNT2B has been previously reported to be expressed in normal ovaries [21] coupled with our finding that it is also expressed in NS makes interpreting the significance of the dichotomy in WNT2B expression between CS_1 and CS_2 ambiguous.

4. Discussion

Cancer progression is a dynamic process involving cellular adaptation and survival that is, in part, driven by signaling interactions between participating cells. Many signaling interactions have been documented to take place between cancer epithelial cells and the surrounding stroma [22]. Early in tumor development, cancer cells produce growth factors that are believed to modulate or "activate" the surrounding

Table 5: The unique compatible ligands and receptors from Table 4 showing the expression pattern in NS, CS_1, CS_2, and Cepi. The 6 bold signals had the same expression in NS and CS_1 but different expression between CS_1 and CS_2 despite the fact that their compatible signals in Cepi were always expressed.

NS	Ligands	CS_1	CS_2	Receptors	Cepi
+	CXCL12	+	+	CXCR4	+
−	FGF1	+	+	FGFR2	−
−	FGF1	+	+	FGFR3	+
+	**FGF2**	**+**	**−**	**FGFR3**	**+**
+	FGF9	+	+	FGFR3	+
−	HGF	+	+	MET	+
+	PDGFA	+	+	PDGFRA	+
+	TGFB2	+	+	TGFBR2	+
−	**IL7**	**−**	**+**	**IL7R**	**+**
−	WNT2	+	+	FZD2	+
+	**WNT2B**	**+**	**−**	**FZD2**	**+**
−	**WNT7A**	**−**	**+**	**FZD7**	**+**
NS	Receptors	CS_1	CS_2	Ligands	Cepi
+	CXCR4	+	+	CXCL12	+
+	FGFR2	+	+	FGF1	+
−	**FGFR3**	**−**	**+**	**FGF1**	**+**
−	**FGFR3**	**−**	**+**	**FGF9**	**+**
+	PDGFRA	+	+	PDGFA	+
+	TGBFR2	+	+	TGFB2	+
+	IL1R1	+	+	IL1B	+
−	IL1R2	+	−	IL1B	+
+	IL7R	+	+	IL7	+
+	FZD2	+	+	WNT2	+
+	FZD2	+	+	WNT2B	+
+	FZD7	+	+	WNT7A	+

Expression is denoted with "+" (i.e., there is at least one Affymetrix present call with detection P value ≤ 0.05) and nonexpression with "−" (i.e., there are no Affymetrix present calls in the samples with detection P value ≤ 0.05).

stroma in order to convert the stroma into a supportive microenvironment for cancer growth [2, 14]. For example, it has been shown that growth factors secreted by macrophages can contribute to cancer progression and metastasis [23]. Other inflammatory cells such as lymphocytes, neutrophils, mast cells, T-regulatory cells, and platelets also have been shown to have the potential to support tumor progression by negatively regulating the anticancer host immune response [24–26]. Fibroblasts, the major component of the stroma, have been shown to be able to participate actively in the malignant progression of cancer by producing growth factors, various chemokines, and extra cellular matrix components that facilitate the production of endothelial cells and pericytes conducive to tumor growth [14, 27].

The purpose of this study was to investigate the process of stroma activation within the context of ovarian cancer. Toward this end, we conducted RNA microarray profiling of 45 tissue samples using the Affymetrix (U133 Plus2) gene expression platform. Laser capture microdissection (LCM) was used to isolate cancer cells from the tumors of 18 ovarian cancer patients (Cepi). For 7 of these patients, a matched set of surrounding cancer stroma (CS) was also collected. For controls, we isolated surface epithelial cells (OSE) from the normal (noncancerous) ovaries of 12 individuals including matched sets of samples of OSE and normal stroma (NS) from 8 of these patients.

Unsupervised hierarchical clustering of the microarray data resulted in the expected separation between the OSE and Cepi samples. Consistent with models of stromal activation, we also observed significant separation between the NS and CS samples. Somewhat unexpected, however, was our finding that the CS samples clustered into two distinct subgroups (CS_1 and CS_2).

Based on patterns of coexpression of ligand and receptor encoding genes, we determined that 6 biologically compatible pairs of ligands and receptors are differentially expressed between Cepi and the CS_1 and CS_2 cancer stroma. The patterns of differential expression between the compatible ligands and receptors are consistent with the hypothesis that CS_2 may be a more conducive microenvironment for tumor growth (Table 5). For example, the expression of tumor promoting ligands in Cepi is always matched with the expression of compatible receptors in CS_2 but not in CS_1.

The fact that certain tumor microenvironments are capable of inhibiting tumor growth and/or development is well established. For example, macrophages can act as anticancer agents within the context of the innate immune response [28]. Likewise, fibroblasts, in some cellular contexts, have been shown to revert tumor cells to a normal, noncancerous phenotype [9, 29]. Normal ovarian stromal cells have been shown to significantly inhibit ovarian cancer cell growth when coinjected into nude mice [30].

The apparently innate anticancer properties of normal stroma are generally considered to be transient giving way to the "activation" of procancer growth signals induced by cancer cells as the tumors progress [1]. However, since the majority of the patients associated with the CS_1 class of cancer stroma have, like the majority of the cancer patients included in our study, already progressed to advanced staged disease (Table 1), it is unlikely that the CS_1 molecular profile represents a transient condition. Rather, our results suggest that variability exists among ovarian cancer patients with respect to the propensity of normal stroma to become activated.

5. Conclusions

An understanding of the potential clinical significance of the observed molecular dichotomy between ovarian cancer stroma is beyond the scope of this present study. However, it is relevant to note that all of the cancers associated with the putatively more permissive CS_2 cancer stroma were classified as grade 3 while those associated with the putatively more resistant CS_1 cancer stroma were classified as grade 2. The fact that no distinction was apparent between the molecular profiles of grade 2 and grade 3 Cepi samples (Figure 1) suggests that cancer grade may, at least in part, be determined by the relative permissiveness of the tumor

microenvironment. Molecular profiling of larger numbers of matched sets of ovarian cancer and stroma samples will be required to adequately test this hypothesis. Nevertheless, the current results are consistent with the hypothesis that the microenvironment plays a significant role in ovarian cancer development and suggest that functionally significant variability may exist among ovarian cancer patients in the ability of the microenvironment to modulate cancer development.

Conflict of Interests

The authors declare that there is no conflict of interests.

Acknowledgments

The project was supported by grants from the Ovarian Cancer Institute (Atlanta), Northside Hospital (Atlanta), the Robinson Family Foundation, Ovarian Cycle Foundation, and the Deborah Willingham Endowment Fund. The authors also wish to thank Dr. Nathan J. Bowen for his advise in early stages of the project.

References

[1] L. A. Liotta and E. C. Kohn, "The microenvironment of the tumour—host interface," *Nature*, vol. 411, no. 6835, pp. 375–379, 2001.

[2] N. A. Bhowmick and H. L. Moses, "Tumor-stroma interactions," *Current Opinion in Genetics and Development*, vol. 15, no. 1, pp. 97–101, 2005.

[3] T. D. Tlsty and L. M. Coussens, "Tumor stroma and regulation of cancer development," *Annual Review of Pathology*, vol. 1, pp. 119–150, 2006.

[4] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, 2000.

[5] A. A. Kamat, M. Fletcher, L. M. Gruman et al., "The clinical relevance of stromal matrix metalloproteinase expression in ovarian cancer," *Clinical Cancer Research*, vol. 12, no. 6, pp. 1707–1714, 2006.

[6] L. Zhang, N. Yang, J. W. Park et al., "Tumor-derived vascular endothelial growth factor up-regulates angiopoietin-2 in host endothelium and destabilizes host vasculature, supporting angiogenesis in ovarian cancer," *Cancer Research*, vol. 63, no. 12, pp. 3403–3412, 2003.

[7] C. Porcile, A. Bajetto, F. Barbieri et al., "Stromal cell-derived factor-1α (SDF-1α/CXCL12) stimulates ovarian cancer cell growth through the EGF receptor transactivation," *Experimental Cell Research*, vol. 308, no. 2, pp. 241–253, 2005.

[8] E. Wang, Y. Ngalame, M. C. Panelli et al., "Peritoneal and subperitoneal stroma may facilitate regional spread of ovarian cancer," *Clinical Cancer Research*, vol. 11, no. 1, pp. 113–122, 2005.

[9] P. Micke and A. Östman, "Tumour-stroma interaction: cancer-associated fibroblasts as novel targets in anti-cancer therapy?" *Lung Cancer*, vol. 45, supplement 2, pp. S163–S175, 2004.

[10] A. Silva, A. Gírio, I. Cebola, C. I. Santos, F. Antunes, and J. T. Barata, "Intracellular reactive oxygen species are essential for PI3K/Akt/mTOR-dependent IL-7-mediated viability of T-cell acute lymphoblastic leukemia cells," *Leukemia*, vol. 25, no. 6, pp. 960–967, 2011.

[11] C. Schroten, N. F. Dits, E. W. Steyerberg et al., "The additional value of TGFβ1 and IL-7 to predict the course of prostate cancer progression," *Cancer Immunology, Immunotherapy*, vol. 61, no. 6, pp. 905–910, 2012.

[12] M. A. A. Al-Rawi, R. E. Mansel, and W. G. Jiang, "Interleukin-7 (IL-7) receptor (IL-7R) signalling complex in human solid tumours," *Histology and Histopathology*, vol. 18, no. 3, pp. 911–923, 2003.

[13] M. J. Maeurer, W. Walter, D. Martin et al., "Interleukin-7 (IL-7) in colorectal cancer: IL-7 is produced by tissues from colorectal cancer and promotes preferential expansion of tumour infiltrating lymphocytes," *Scandinavian Journal of Immunology*, vol. 45, no. 2, pp. 182–192, 1997.

[14] N. A. Bhowmick, E. G. Neilson, and H. L. Moses, "Stromal fibroblasts in cancer initiation and progression," *Nature*, vol. 432, no. 7015, pp. 332–337, 2004.

[15] C. Jin, F. Wang, X. Wu, C. Yu, Y. Luo, and W. L. McKeehan, "Directionally specific paracrine communication mediated by epithelial FGF9 to stromal FGFR3 in two-compartment premalignant prostate tumors," *Cancer Research*, vol. 64, no. 13, pp. 4555–4562, 2004.

[16] N. Turner and R. Grose, "Fibroblast growth factor signalling: from development to cancer," *Nature Reviews Cancer*, vol. 10, no. 2, pp. 116–129, 2010.

[17] T. Reya and H. Clevers, "Wnt signalling in stem cells and cancer," *Nature*, vol. 434, no. 7035, pp. 843–850, 2005.

[18] T. D. Bui, M. Lako, S. Lejeune et al., "Isolation of a full-length human WNT7A gene implicated in limb development and cell transformation, and mapping to chromosome 3p25," *Gene*, vol. 189, no. 1, pp. 25–29, 1997.

[19] S. Yoshioka, M. L. King, S. Ran et al., "WNT7A regulates tumor growth and progression in ovarian cancer through the WNT/β-catenin pathway," *Molecular Cancer Research*, vol. 10, no. 3, pp. 469–482, 2012.

[20] A. B. Tchagang, A. H. Tewfik, M. S. DeRycke, K. M. Skubitz, and A. P. N. Skubitz, "Early detection of ovarian cancer using group biomarkers," *Molecular Cancer Therapeutics*, vol. 7, no. 1, pp. 27–37, 2008.

[21] A. Ricken, P. Lochhead, M. Kontogiannea, and R. Farookhi, "Wnt signaling in the ovary: identification and compartmentalized expression of wnt-2, wnt-2b, and frizzled-4 mRNAs," *Endocrinology*, vol. 143, no. 7, pp. 2741–2749, 2002.

[22] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[23] A. Nowicki, J. Szenajch, G. Ostrowska et al., "Impaired tumor growth in colony-stimulating factor 1 (CSF-1)-deficient, macrophage-deficient op/op mouse: evidence for a role of CSF-1-dependent macrophages in formation of tumor stroma," *International Journal of Cancer*, vol. 65, no. 1, pp. 112–119, 1996.

[24] G. Pawelec, "Tumour escape: antitumour effectors too much of a good thing?" *Cancer Immunology, Immunotherapy*, vol. 53, no. 3, pp. 262–274, 2004.

[25] L. Yang and D. P. Carbone, "Tumor-host immune interactions and dendritic cell dysfunction," *Advances in Cancer Research*, vol. 92, pp. 13–27, 2004.

[26] L. M. Coussens and Z. Werb, "Inflammation and cancer," *Nature*, vol. 420, no. 6917, pp. 860–867, 2002.

[27] R. Kalluri and M. Zeisberg, "Fibroblasts in cancer," *Nature Reviews Cancer*, vol. 6, no. 5, pp. 392–401, 2006.

[28] S. Gillessen, Y. N. Naumov, E. E. S. Nieuwenhuis et al., "CD1d-restricted T cells regulate dendritic cell function and antitumor

immunity in a granulocyte-macrophage colony-stimulating factor-dependent fashion," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 15, pp. 8874–8879, 2003.

[29] T. A. Gonda, A. Varro, T. C. Wang, and B. Tycko, "Molecular biology of cancer-associated fibroblasts: can these cells be targeted in anti-cancer therapy?" *Seminars in Cell and Developmental Biology*, vol. 21, no. 1, pp. 2–10, 2010.

[30] J. A. Parrott, E. Nilsson, R. Mosher et al., "Stromal-epithelial interactions in the progression of ovarian cancer: influence and source of tumor stromal cells," *Molecular and Cellular Endocrinology*, vol. 175, no. 1-2, pp. 29–39, 2001.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 925373, 9 pages
http://dx.doi.org/10.1155/2013/925373

Research Article

Structural Adaptation of Cold-Active RTX Lipase from *Pseudomonas* sp. Strain AMS8 Revealed via Homology and Molecular Dynamics Simulation Approaches

Mohd. Shukuri Mohamad Ali,[1,2] Siti Farhanie Mohd Fuzi,[1,2] Menega Ganasen,[1,2] Raja Noor Zaliha Raja Abdul Rahman,[1,3] Mahiran Basri,[1,4] and Abu Bakar Salleh[1,2]

[1] *Enzyme and Microbial Technology Research Center, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia*
[2] *Department of Biochemistry, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia*
[3] *Department of Microbiology, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia*
[4] *Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia*

Correspondence should be addressed to Mohd. Shukuri Mohamad Ali; shukuri@biotech.upm.edu.my

Received 8 January 2013; Revised 6 March 2013; Accepted 29 March 2013

Academic Editor: Yudong Cai

The psychrophilic enzyme is an interesting subject to study due to its special ability to adapt to extreme temperatures, unlike typical enzymes. Utilizing computer-aided software, the predicted structure and function of the enzyme lipase AMS8 (LipAMS8) (isolated from the psychrophilic *Pseudomonas* sp., obtained from the Antarctic soil) are studied. The enzyme shows significant sequence similarities with lipases from *Pseudomonas* sp. MIS38 and *Serratia marcescens*. These similarities aid in the prediction of the 3D molecular structure of the enzyme. In this study, 12 ns MD simulation is performed at different temperatures for structural flexibility and stability analysis. The results show that the enzyme is most stable at 0°C and 5°C. In terms of stability and flexibility, the catalytic domain (N-terminus) maintained its stability more than the noncatalytic domain (C-terminus), but the non-catalytic domain showed higher flexibility than the catalytic domain. The analysis of the structure and function of LipAMS8 provides new insights into the structural adaptation of this protein at low temperatures. The information obtained could be a useful tool for low temperature industrial applications and molecular engineering purposes, in the near future.

1. Introduction

A lipase (also known as triacylglycerol acylhydrolase (E.C 3.1.1.3)) is a serine hydrolase, which acts under aqueous conditions on the carboxyl ester bond of triacylglycerol to produce fatty acids and glycerol [1]. Lipases display common α/β-hydrolase folds that are also present in other hydrolases [2]. A typical lipase consists of an active site comprised of the catalytic triad of serine, glutamine/aspartate, and histidine [3].

Lipases are widely distributed among living organism, including bacteria, eukarya, and archaea as has been reported by Jaeger et al. [4]. Recently, lipases produced by psychrophilic bacteria have been studied because of their low optimum temperatures and high activities at very low temperatures. This is reportedly due to the inherent greater flexibility compared to mesophilic and thermophilic enzymes. These enzymes are severely impaired by an excess of rigidity. Additionally, peculiar properties of psychrophilic enzymes render them particularly useful as valuable tools for

biotechnological purposes and for investigating the possible relationships between stability, flexibility, and specific activity [5, 6].

However, the adaptation of the enzyme at low temperatures is not fully understood because there has been little study on psychrophilic enzymes. Some features discovered by scientists include reduced numbers of salt bridges, slightly lower [Arg/(Arg + Lys)] ratios, reduced numbers of nonpolar residues, and higher numbers of exposed nonpolar residues [7]. In term of stability, psychrophilic enzymes are mostly unstable, as has been proven by various demonstrations, including fluorescence spectroscopy and other techniques. The enzyme tends to unfold at lower temperatures and calorimetric enthalpies [8]. Researchers suggest that one feature of these enzymes is higher numbers of nonpolar residues on their surfaces, which is responsible for the destabilization of the water structure surrounding the enzymes. There are also fewer arginine and proline residues; this may increase the backbone flexibility. Note that research regarding the flexibility of the psychrophilic enzymes (using spectroscopic analysis, dynamic fluorescence quenching, and molecular dynamics simulations) has supported the idea that increased flexibility of psychrophilic enzymes contributed to the evolution of psychrophilic enzymes [9].

The importance of understanding the structural adaptations of extremozymes is underscored by their usefulness in various industrial applications. Until now, only a few extremophilic organisms, particularly psychrophiles, have been characterized and used as enzyme sources for industrial processes. Previously, a new strain of psychrophilic bacteria (designated strain AMS8) from Antarctic soil was screened for extracellular lipase activity and further analyzed using a molecular approach. Analysis of 16S rDNA showed that the strain AMS8 was similar to *Pseudomonas* sp. A lipase gene named *LipAMS8* was successfully isolated from strain AMS8 with an open reading frame of 1,431 bp that encoded a polypeptide consisting of 476 amino acids. This crude lipase exhibited maximum activity at 20°C. Additional genetic studies revealed that *LipAMS8* lacked an N-terminal signal peptide and contained a glycine- and aspartate-rich nonapeptide sequence at the C-terminus (experimental data).

In this study, the structure and function of lipase isolated from *Pseudomonas* sp. strain AMS8 are studied by using the structure predicted by using appropriate software. The structural adaptation of the enzyme at low temperatures is also studied using molecular dynamic simulation (MD simulation). Because it is a newly isolated enzyme, further study is needed to provide further understanding and reveal the potential of this psychrophilic enzyme, which may be used for the industrial, biotechnological, and fundamental purposes.

2. Materials and Methods

2.1. Software. The modeling and simulation of the enzyme's predicted structure was run on a single PC (Intel (R) Core RM i5 CPU, 650 @ 3.2 GHz Co, 4.0 GB RAM) with the Windows 7 Ultimate operating system. The Yet Another Scientific Artificial Reality Application (YASARA) software [10] program was installed on the PC and was used for the molecular modeling and molecular dynamics (MD) simulation of the LipAMS8 predicted molecular structure.

2.2. Sequence Alignment of LipAMS8. The amino acid sequence of the LipAMS8 enzyme obtained from NCBI with Accession Number of ADM87309 consists of 476 amino acid residues; a weight of 50 kDa is used for sequence analysis and modeling. The BLAST [11] program identified the homologous sequence that has high sequence identity with the LipAMS8 enzyme. The sequence with the highest score of sequence identity was chosen based on certain characteristics, including the type of origin and the availability of the solved 3D structures. Subsequently, multiple sequence alignment was carried out using the Biology Workbench [12] open software with the protein sequences of *Serratia marcescens* [13] and *Pseudomonas* sp. MIS38 [14] as both of these templates fulfilled the criteria needed as mentioned above.

2.3. Comparative Modeling and Validation. The templates used for the modeling were the crystal structures of the lipases obtained from *Serratia marcescens* [13] and *Pseudomonas* sp. MIS38 [14]. The atomic coordinates for the lipases *Serratia marcescens* (PDB ID: 2qua) and *Pseudomonas* sp. MIS38 (PDB ID: 2z8x) were obtained from the Protein Data Bank. The 3D model was generated using the YASARA [10]. The validation was performed with VERIFY3D [15] (to evaluate the fitness of the protein sequence in its current 3D structure) and Ramachandran plot [16] (to evaluate the geometrical aspects of the structure).

2.4. Molecular Dynamics (MD) Simulations at Various Temperatures. An MD simulation provided more information for detailed microscopic modeling on the molecular scale. The method follows the constructive approach by mimicking the behavior of molecules with the use of model systems. More powerful computers make it possible to study greater complexity with a realistic expectation of obtaining meaningful and useful information [17].

In this study, the MD simulation was performed in water. This involved the simulation of predicted model inside a trajectory box filled with 6940 molecules of solvent (including NaCl and water) and 467 LipAMS8 amino acid residues at the temperatures of 0°C, 5°C, 25°C, 37°C, 50°C, and 100°C. The density of water varies with temperature because the theoretical density of water depends on temperature.

The AMBER03 [18] force field parameter which implemented in the YASARA software was used for MD simulation. In each simulation, the initial model was minimized in order to reduce the contact area difference (CAD) between the protein model and the solvent molecule. During the minimization, conjugate-descent and steepest-descent algorithms were employed.

2.5. Simulation Analysis. The enzyme was studied using 240 saved steps for each simulation, which represents up

to 6 nanoseconds of the production period. The analysis provides better understanding of the dynamic properties of the enzyme in water at different temperatures. The root mean square deviation (RMSd) was computed for the protein backbone and residues in order to check the stability of the trajectories. Additionally, the root mean square fluctuation (RMSf) was computed per residue in order to study the flexibility of the trajectories. Further analysis was performed by calculating the radius of gyration (Rgyration) and solvent accessible surface area (SASA) of the enzyme within the 6 nanoseconds (ns) of production time.

3. Results and Discussion

3.1. Comparative Modeling of LipAMS8

3.1.1. Modeling of LipAMS8. The sequence alignment searches for suitable templates to construct the 3D structure of the lipase AMS8 (LipAMS8), using comparative modeling. In this study, the crystal structures from *Pseudomonas* sp. MIS38 lipase and *Serratia marcescens* LipA (which score 80% and 69% for sequence identity) were chosen as the templates for modeling because both have highest scores of sequence identity when aligned with the LipAMS8 sequence. Both templates also have solved structures that can be obtained from the RSCB PDB Data Bank, which is important for predicting the 3D structure of the enzyme.

Two templates were used in the modeling of LipAMS8 in such a way that a model was formed from each template, as well as hybrid model which formed based on the best configuration of protein using the two templates chosen. However, the Z-score of each model is the most crucial parameter as the best Z-scores value obtained from each model obtained will denote for the accuracy and the quality of the model itself. Subsequently, the hybrid structure which is expected to be the best model for LipAMS8 is rejected due to poor Z-scores compared to the one obtained using only *Pseudomonas* sp. MIS38 as template.

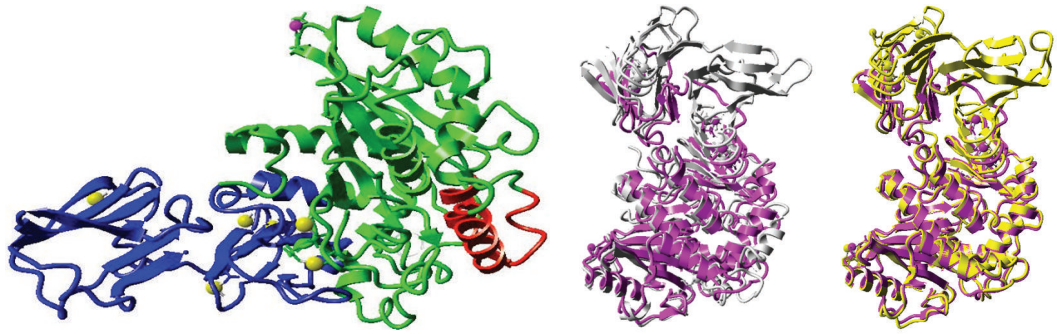
The model was validated using the Ramachandran plot (Figure 2). The model had 89.5% of the residues residing in the most favored allowed region. Although the best scores are 90.0% and higher, the score obtained is considered to be acceptable because the model is a prediction model and not a crystal structure (i.e., the crystal structure is a fully solved structure compared to the predicted one). In a previous study, the serine of the catalytic triad, which resides in the negative/disallowed region of the Ramachandran plot, was proposed for the active conformation of the enzyme [19]. However, from this study, the Ser^{207} of the catalytic triad of the LipAMS8 resides in the allowed region of the Ramachandran plot. This suggests that the enzyme is in the nonactive conformation, which is supported by the observable lid structure of the enzyme (closed conformation). Prediction of the active conformation of the enzyme also can be performed by observing the lid structure.

As compared to the template structure used to model LipAMS8 as shown in superimposed image in Figures 1(b) and 1(c), the predicted model of LipAMS8 also displays two main regions: the catalytic (green) and noncatalytic (blue) domains, as shown in Figure 1(a). The catalytic domain at the N-terminal is rich in α helices, while the noncatalytic domain at the C-terminal is dominated by β-strands. The catalytic domain also exhibits the presence of an α/β-hydrolase fold and catalytic triad, which includes Ser^{207}, His^{255}, and Asp^{313} residues. The Ser^{207} appears in the pentapeptide of G-X-S-X-G motifs (where X represents His^{206} and Leu^{208}) and is located at the sharp turn between the β-strand and α-helix that resembles the nucleophilic elbow (normally present in the structural family of α/β-hydrolases) [20]. This suggests that the catalytic serine is aided by an oxyanion hole that stabilizes the negative charge generated during a nucleophilic attack by Ser Oγ [19].

Generally, a lipase can exist in 2 conformational states: active and inactive. The lid conformation (open or closed) determines whether the enzyme is in the active or inactive conformation. The active conformation of the lipase is necessary for catalytic activity [19]. In this study, there is a lid-like structure that covers the catalytic site on the catalytic domain, which suggests the inactive conformation of the enzyme. Previous studies suggested that the active form of the enzyme has the lid open, allowing for the entrance of the substrate into the binding site for catalysis [21]. Thus, the closed conformation of the lid in this study meant that the nucleophilic Ser^{207} could not attack the substrate. To open the lid, a water-oil interface is required so the lid can be modulated to uncover the catalytic site, providing access to the catalytic pocket for the substrates [19]. This enhances the activity of LipAMS8. Lid number 1 of LipAMS8 has high numbers of hydrophobic residues, including Ala^{51}, Leu^{53}, Val^{54}, Val^{57}, and Val^{58}. This may allow efficient interaction between the hydrophobic lid residues with the lipid interface and contribute to easier lid opening.

From this study, the LipAMS8 consists of RTX motifs at the noncatalytic domain, suggesting the LipAMS8 is one of the RTX lipases that belong to the I.3 subfamily. This is further proven by the absence of cysteine residues [22]. Note that, the RTX motif is present in a variety of Gram-negative microorganisms [6]. This motif (comprised of glycine-rich nonapeptide sequences) is usually located at the carboxy-terminal portion of an enzyme [22]. In the 3D structure of LipAMS8, the sequence of RTX motifs constituted the parallel β-roll, which the first 6 residues of each motif form to attach calcium ions (Ca^{2+}). The remaining 3 residues build short β-strands, which result in the right-handed helix of the β-strand on the noncatalytic domain. The exact function of the RTX motifs remains obscure. However, they could be receptor binding domains, enhancers of secretion, and internal chaperones [6].

In the predicted model of LipAMS8, there are metal ions, including 6 atoms of Ca^{2+} and 1 atom of Zn^{2+}. Note that metal ions are required by a substantial fraction of enzymes to perform catalytic activity. Metal ions may also contribute to substrate activation and electrostatic stabilization of enzyme structure. In this study, Asp^{128}, Asp^{130}, and the ligand interact in the Zn^{2+} binding sites. The Zn^{2+} present in this LipAMS8 is not located in the catalytic site, so it may not contribute to the catalytic activity of the enzyme. Along with their contribution



(a) (b) (c)

Figure 1: (a) LipAMS8 predicted 3D structure. The structure is composed of catalytic (green) and noncatalytic (blue) domains. The lid is colored in (red). (b) and (c) are superimposition of LipAMS8 structure (purple) with 2QUA (silver), and 2Z8X (yellow).

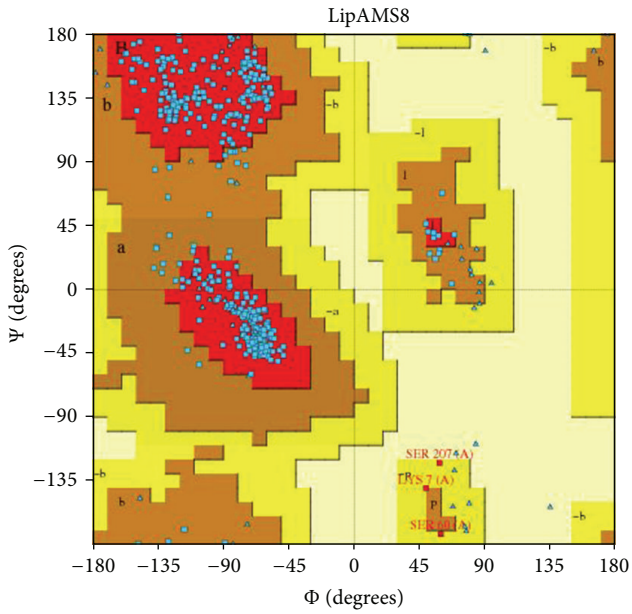


Figure 2: Ramachandran plot of the LipAMS8 3D structure. The structure scores 89.5%, meaning that 89.5% of the residues reside in the most favored region.

to catalytic activity, ions may ensure the local and overall structural stability, similar to the function of disulfides [23]. To conclude, there are metal ions in the predicted structure of LipAMS8, which may contribute to the overall structural stability. However, the enzyme is observed to have few or no arginines compared to lysines, low proline content, and a lack of a salt bridge, which reportedly contributes to the adaptation of psychrophilic enzymes at low temperatures. Additionally, the enzyme is predicted to have higher flexibility compared to mesophilic or thermophilic enzymes, allowing psychrophilic enzymes to be active at low energy costs [7].

3.1.2. Molecular Dynamic (MD) Simulations. MD simulation was performed to reveal changes in the structure, flexibility, and dynamics of LipAMS8 when simulated in elevated temperatures. The conformational sampling was limited to 12 ns. It has been suggested that LipAMS8 is a cold-active enzyme. Therefore, the structure should denature or unfold as the temperature increases from to 25°C to 100°C because of the disruption of the intermolecular forces, due to the increase in kinetic energy at elevated temperatures. However, in this study, there was no much unfolding of the secondary structure even when the enzyme is simulated at a higher temperature. This might be due to the limitation of conformational sampling, which is a maximum of 12 ns. The simulation time may need to be prolonged in order to see the structural changes.

3.1.3. Molecular Dynamics Simulation Data Analysis. The root mean square deviation (RMSd) values of the backbone atoms in the initial models assess the convergence of the protein system. In this study, the RMSd values from the minimized predicted model structure during MD simulation at 0°C, 5°C, 25°C, 37°C, 50°C, and 100°C are shown in Figure 3. At 5°C, the protein remains native-like and equilibrates with an average of 1.83 Å from the minimized structure. This indicates the stability of the enzyme at 5°C. Compared to 5°C, RMSd value at 0°C demonstrated that the 3D structure of the protein becomes unstable when it reaches 1000 ps. The RMSd value begins to fluctuate to values higher than 2 Å and even reached the value of 3.45 Å at 2525 ps. The value dropped to 1.807 Å at 3675 ps and rose to higher than 2 Å at 4000 ps. The unstable fluctuation of the RMSd is consistent with the difference in secondary structure elements observed during the simulation. However, the structural stability is observed after 5500 ps as the pattern of RMSd fluctuation is maintained and does not diverge more than 2 Å towards the end of simulation at 12 ns.

The fluctuation trend for the stability of the enzyme at 25°C, 37°C, 50°C, and 100°C increased in value throughout simulation with RMSd values remaining higher than 4.0 Å at 3000 ps for the rest of the simulation period. The unstable state of this enzyme is supported by experimental data, which found that the enzyme activity of LipAMS8 was reduced when the temperature was increased above 25°C. From the

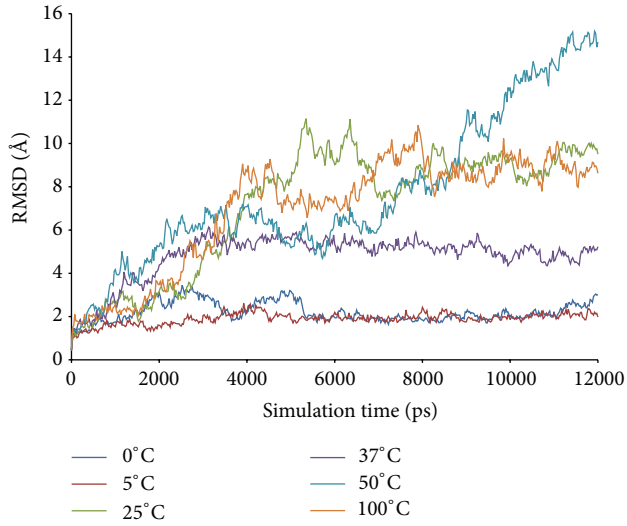


Figure 3: LipAMS8 root mean square deviations (RMSd) of the backbone atoms as functions of time.

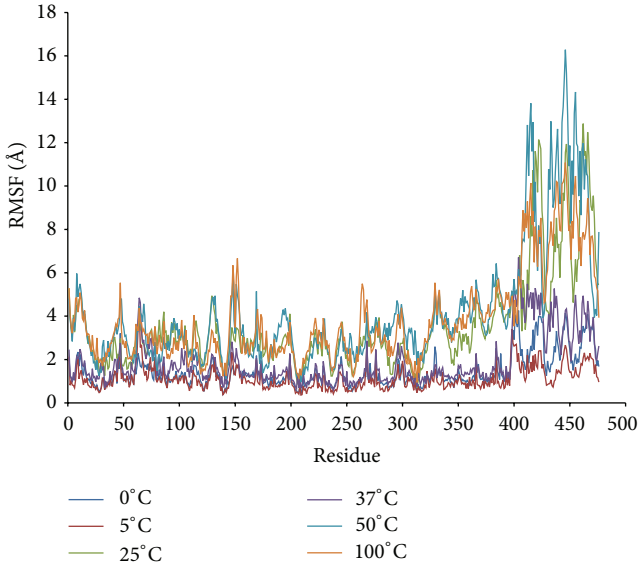


Figure 4: LipAMS8 atoms root mean square fluctuations (RMSf) as functions of time.

study, we deduce that at 25°C, 37°C, 50°C, and 100°C, the global 3D structure of the protein loses its native structure in order to adapt the molecule to changes in temperature and water density (the solvent). Thus, this result indicates that changes in the geometry coordinates and unfolding of the protein result from the reduced stability of the system.

We deduced that the most stable temperatures for LipAMS8 simulated in water are 0°C and 5°C as low temperature promotes less conformational movement maintaining structural integrity and stability. On the other hand, for temperatures between 25°C and 100°C, the enzyme structure is unstable and greatly deviates from its initial structure. The higher deviation of the enzyme structure when simulated in water at 25°C to 100°C may be related to the disruption of the molecular forces, which leads to higher kinetic activities of the molecules. However, the structural changes of the enzyme simulated in water are not that critical when compared to the structural changes that occur when the enzyme is simulated in an organic solvent [24].

Interestingly, average RMSd scores for 0°C and 5°C was seen to be stable throughout simulation. However, the stability was only adapted to catalytic domain of enzyme which is believed to be promoted by the presence of metal ions such as Zn^{2+} and Ca^{2+}. In other hands, the unstable state of this enzyme which is contributed from high flexibility of the noncatalytic domain at low temperature such as 0°C and 5°C however is somehow predictable since the enzyme has to adjust itself towards the low temperature whereby the kinetic energy is slowly decrease. If the enzyme does not increase their flexibility at this point, the structure may become too rigid and could not compensate for the decrease in catalysis temperature.

Figure 4 shows the root mean square fluctuations (RMSf) per residue for LipAMS8 simulated in water at temperatures from 0°C to 100°C. The average RMSf scores per residue for all temperatures vary from 1.12 Å to 4.1 Å. The highest fluctuation score is at 100°C. This result is equivalent to the effect of higher temperature on the flexibility of the enzyme. However, when comparing the flexibility of the enzyme at 0°C and 5°C, the RMSf value decreased from 1.43 to 1.12 Å. The higher flexibility of the enzyme at the lower temperature may suggest an adaptation of the enzyme to counteract the "freezing effect" as the temperatures dropped.

At the catalytic site, the RMSf scores do not indicate higher flexibility of the residues simulated in water at different temperatures. The flexibility is maintained at a value below the average RMSf score at different temperatures. This suggests the stability of the catalytic triad is supported by the figure of RMSd per residue.

Overall, the pattern of fluctuation per residue was similar to the pattern shown in the figure of RMSd per residue. Both figures show that higher fluctuation occurs at the noncatalytic domain. The most consistent fluctuation across temperatures occurred at residues 393–476, which reside in the noncatalytic domain whereby the RTX repeats are present. The flexibility of the domain may occur because of the presence of high glycine residues, which are known to introduce flexibility in protein structure because they lack side chains [25]. The increased flexibility of LipAMS8 (which originated from the low stability of the noncatalytic domain) may imply that the noncatalytic domain is the crucial part of the LipAMS8 molecular structure. This domain may contribute to high enzyme activity at low temperature.

An examination of the structural flexibility of the lid structure reveals that lid number 1, comprised of residues 51–58, does not have a higher than average RMSf score at any temperature. However, lid number 2, which comprised of residues 148–167, does have fluctuation at residues 148–153 at most temperatures. This proves the flexibility of the lid residue when simulated in water at various temperatures. This result led to the hypothesis that lid number 2 of LipAMS8

may be able to undergo a conformational transition under the right conditions, such as the presence of substrates. The residues 148–153 on lid number 2, which is more flexible, may acts like a holder that opens up the lid structure to expose the binding site in interfacial activation. In contrast to lid number 1, lid number 2 is more flexible. Thus, lid number 2 may be the first lid to open when substrates are present. The water-oil interface around the opening of first lid is less flexible for binding the substrate to the catalytic site. Note that the lid opening is a crucial step in a lipase with lid-like structure. The flexibility of the residues that reside in the lid structure is important for determining the motion rate of the lid, which plays an important role in the adaptation of enzyme function at low temperatures [7].

The structural flexibility observed on the structure of LipAMS8, simulated in water at various temperatures, may lessen if simulated in organic solvent. A previous study suggests higher rigidity of the enzyme in organic solution [26]. However, there is no much information available at the molecular level to accept or reject the proposal. To check the structural flexibility of the enzyme in both aqueous and nonaqueous environments, further simulation is needed to compare the structural basis of the enzyme.

The root mean square fluctuation (RMSf) analysis per residue and root mean square deviation per residue (RMSd) of the backbone atoms are used to analyze the atomic fluctuations of the predicted model of LipAMS8. Figure 5 shows the RMSd per residue at 0°C, 5°C, 25°C, 37°C, 50°C, and 100°C. The average RMSd for all residues, including the terminal residue, qualitatively measures the protein flexibility for the relationship between protein conformational flexibility and dynamics.

From the data, the RMSd average values of residues are almost the same between 0°C and 5°C. However, at 25°C, 37°C, 50°C, and 100°C, the RMSd average values increase to 6.29 Å. The pattern of the RMSd average values of residues is similar to the average values of root mean square fluctuation (RMSf) at all temperatures.

The deviation of the structural geometry of the enzyme mainly occurred at the noncatalytic domain. The RMSd values per residue at residues 412–425 and 436–470 are higher than the average score of RMSd at each temperature. From this, the destabilization of the enzyme does not involve the whole protein as proposed by one study [19]. Instead, the destabilization of the enzyme may involve a portion of the enzyme (as observed in this study). Additionally, the data also rejected the postulate that the destabilization of the enzyme mostly occurs at the catalytic domain [8]. The destabilization of the enzyme mostly occurs on the noncatalytic domain; this may be caused by the presence of residues that cause a loss in stability of the enzyme 3D structure. Interestingly, β-roll is proposed to maintain the stability of *Pseudomonas* sp. MIS38. Deletion, alteration, and mutation on this β-roll which is formed by the presence of RTX motifs and Ca^{2+} would cause the structural conformation changes and denaturation of the enzyme. However, we proposed that actually shorter/lesser number of RTX repeats, less number of metal ions in this counterpart of *Pseudomonas* sp. MIS38, may be the reason why LipAMS8 is able to adapt in extremely cold environment.

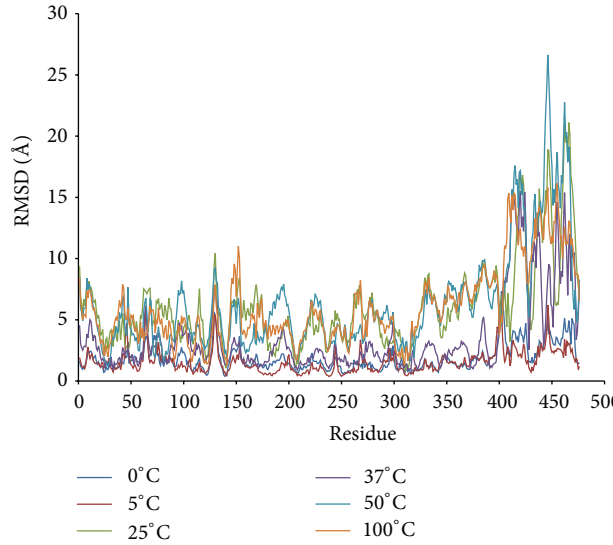


Figure 5: LipAMS8 atoms root mean square deviations (RMSd) per residue.

Study of the catalytic domain shows that Asp^{128} and Asp^{130}, which interact with the Zn^{2+}, have RMSd per residue values higher than the average scores of RMSd at each temperature. Thus, Asp^{128} and Asp^{130} are among those residues that fluctuate and mobilize the adaptation to temperature. However, the ratios of fluctuation of Asp^{128} and Asp^{130} to the average scores of the RMSd at each temperature decrease as the temperature increases. This implies higher interaction between metals and Asp^{128} and Asp^{130} at higher temperatures. This result strongly agrees with a previous study, which suggested that Zn^{2+} promotes structural stabilization in an active conformation of the enzyme at elevated temperatures [27]. This result suggests a higher probability of adaptation of the enzyme at higher temperatures, especially in the catalytic domain because Zn^{2+} stabilizes the enzyme structure.

As depicted from Figure 4, the scores for those residues that interact with Ca^{2+} in the catalytic domain are maintained. This indicated that the metal contributes to stabilizing the catalytic domain so that the structure does not deviate from its initial structure. This result agrees with the function of Ca^{2+} in the structural stability of the lipase proposed in a previous study of the *B. glumae* lipase [28]. Experimental data, on the effect of calcium on LipAMS8 activity, also supported the idea that Ca^{2+} may promote structural stability of the enzyme. Thus, the enzymatic properties are maintained and the catalytic activity is improved to provide a greater yield. In contrast to the residues that interact with Ca^{2+} at the catalytic domain, Ca^{2+} in the noncatalytic domain has higher values of RMSd and RMSf, which indicates destabilization and flexibility of the residue. The Ca^{2+} contribution on the stability of the residue does not agree with the suggestion that it promotes stability. The scores of RMSd and RMSf of the residue that interacts with Ca^{2+} in the noncatalytic domain may be due to the process of adaptation from Ca^{2+} so that the catalytic domain remains stable. This prevents the catalytic

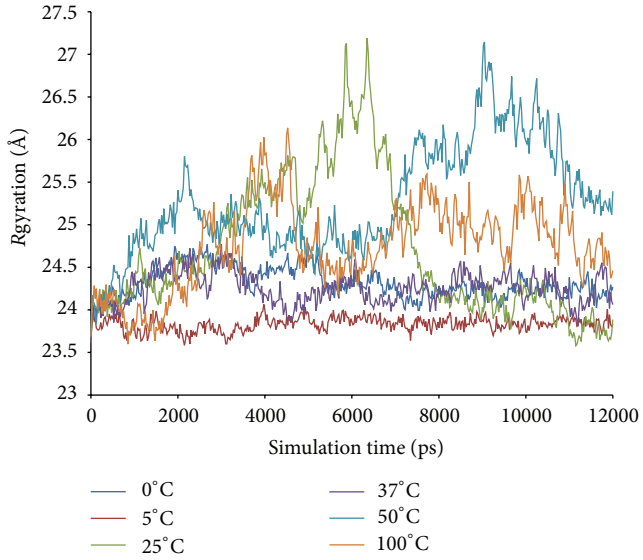


Figure 6: LipAMS8 radius of gyration (Rgyration) scores as functions of time.

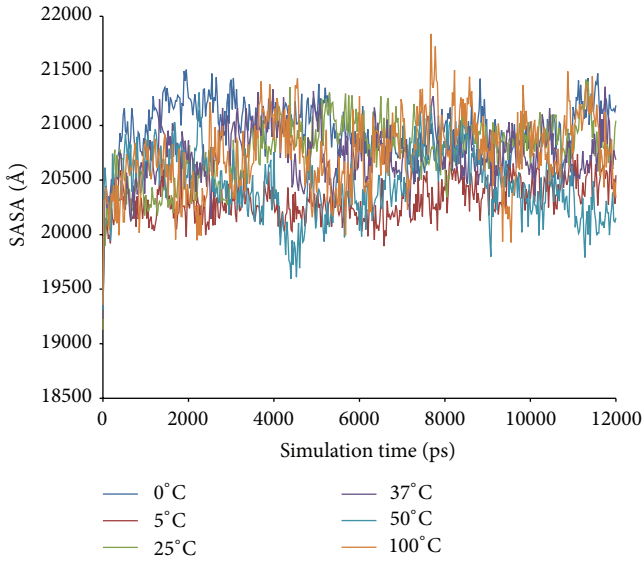


Figure 7: LipAMS8 solvent accessible surface area (SASA) scores as functions of time.

domain, which is responsible for catalytic activity, from being unfolded and dysfunction due to changes in the structural stability and configuration.

The radius of gyration (Rgyration) of the enzyme at different temperatures within the trajectories at 12 ns is shown in Figure 6. The parameter provides information on the tendency of the protein structures to expand during a dynamic simulation. At 5°C, the score of Rgyration is maintained throughout the simulation compared to the Rgyration score of the enzyme when simulated at higher and lower temperatures. The highest score of radius of gyration of LipAMS8 is at 25°C; the score increased to 27.128 Å within the 12 ns of simulation. This is followed by other temperatures' scores

Table 1: Ramachandran plot scores for LipAMS8 after simulation for 12 ns at different temperatures.

Temperature	0°C	5°C	25°C	37°C	50°C	100°C
Most favored regions (%)	85.1	83.1	87.2	81.3	84.4	82.1
Additional allowed regions (%)	13.1	15.9	12.3	17.7	14.9	16.2
Generously allowed regions (%)	1.5	0.8	0.3	0.3	0.0	1.3
Disallowed regions (%)	0.3	0.3	0.3	0.8	0.8	0.5
Nonglycine and nonproline residues (%)	100.0	100.0	100.0	100.0	100.0	100.0
Endresidues (excl. Gly and Pro)	2	2	2	2	2	2
Glycine residues	69	69	69	69	69	69
Proline residues	15	15	15	15	15	15
Total number of residues	476	476	476	476	476	476

except for the score at 5°C. This indicates the adaptation of the enzyme structure to prevent loss of the native compactness of the structure at low temperature. In general, the result is proportional to the deviation of the enzyme structure, as indicated in Figure 3. While SASA indicates the transfer of free energy required to move a protein from aqueous to nonpolar solvent [29], as exhibited in Figure 7, the SASA scores at 5°C and 0°C show the fluctuation pattern, which is maintained throughout the 12 ns of simulation. Compared to 25°C, the figure increases, which indicates unfolding of enzyme within the simulation period. The unfolding of enzyme at this temperature is proportional to the structural changes of enzyme's secondary structure. The increase in SASA score is also observed at the temperatures of 37°C, 50°C, and 100°C.

From further analysis done, it is observed that the scores obtained from the Ramachandran plot may be able to provide insight about the relationship between enzyme activity and the stability. It is suggested by previous studies from researchers worldwide that 20°C to 30°C is the range for the optimum temperature of cold-active enzyme. Suggesting that, in order for the enzyme to perform the catalytic activity optimally at particular temperature, the enzyme should be in good conformation whereby the degree of quality of the enzyme may be judged based on the scores obtained from the Ramachandran plot. As stated in Table 1, structure obtained from 25°C scores for the highest percentage of Ramachandran plot compared to others indicates the best geometrical conformation of the enzyme at this temperature. On the contrary the stability judged from the figure indicates the instability of enzyme at 25°C. The contrast result from Ramachandran plot and average RMSD scores is found to be the key to explain how the structural stability and conformation could give impact to the catalytic activity of the enzyme.

Noting that, in cold-active enzyme the issue on the enzyme dynamic stability and activity remained as hot issue to be discussed among cold-active enzyme researchers. As observed, at 25°C the enzyme is in a good geometrical

conformation which may allow for the best activity at this temperature even though the enzyme is seen to be unstable at this temperature. From these, we deduce that flexibility of the enzyme is crucial to allow the catalytic activity of the enzyme. In contrary to mesophilic and thermophilic counterparts, if the cold-active enzyme is too rigid or too stable with low flexibility, the low activation free energy value ($\Delta G^{\#}$) may not be able to be overcome so that the activity cannot be promoted and thus leads to inability of the enzyme to perform the catalytic activity.

As observed in all lipases, the common α/β-hydrolase fold which is normally seen in lipases structure is also observable in the structure of both LipAMS8 and T1 lipases [30]. Both have lid structures that proposed to function like doors that will open up during interfacial activation. Noting that both LipAMS8 and T1 differ in terms of their philicity towards temperature, significance difference in stability towards certain temperature is observed. LipAMS8, due to its psychrophilic properties, is suggested to be stable at 0°C and 5°C with optimum temperature for activity around 25°C as annotated by the pattern of average Cα atom root mean square deviation throughout 12 ns simulation period. Compared to T1 lipase which is stable at 60°C with optimum temperature of 70°C [31], LipAMS8 has the ability to function at low temperature whereby no other thermophilic enzyme is able to function. The mechanism of adaptation for LipAMS8 enables it to withstand and maintain the stability of the enzyme even at low temperature. However, to concise that eventhough LipAMS8 is able to maintain the stability at low temperature, LipAMS8 also is prone to cold denaturation due to broken bond and interaction of molecules. In other hands, T1 lipase which is thermoalkalophilic enzyme is more stable at higher temperature due to the presence of numbers of metal ions such as Zn^{2+} and Ca^{2+} that are proposed to promote stability towards the T1 lipase.

4. Conclusions

The structural and dynamic study of LipAMS8 provides insights on the structural properties of the enzyme when simulated in water at various temperatures. This RTX (repeat in toxin) lipase might be the first psychrophilic lipase studied in terms of structural stability and flexibility using MD simulation. The enzyme is in a closed conformation and inactive as indicated by the presence of closed lids. From the MD simulation, the LipAMS8 has higher stability at 0°C and 5°C; the catalytic domain exhibits higher stability but lower flexibility compared to the noncatalytic domain. The structural change mainly occurs on the noncatalytic domain of the enzyme, including the unfolding of the secondary structures.

There has not been much work on the physicochemical properties of the enzyme. Therefore, further study, including crystallization of the enzyme, simulation of the enzyme in organic solvent, and docking, should be performed to provide information and understanding of the cold adaptation and structural insights of LipAMS8.

Abbreviations

Å:	Angstrom
ns:	Nanoseconds
ps:	Picoseconds
LipAMS8:	Lipase AMS8
Rgyration:	Radius of gyration
RMSd:	Root mean square deviation
RMSf:	Root mean square fluctuation
RTX:	Repeat in toxin
SASA:	Solvent accessible surface area
YASARA:	Yet Another Scientific Artificial Reality Application.

Acknowledgment

This project was supported by the Science fund research grant, Ministry of Science, Technology and Innovation, Malaysia (Project no. 02-01-04-SF1133).

References

[1] R. Gupta, N. Gupta, and P. Rathi, "Bacterial lipases: an overview of production, purification and biochemical properties," *Applied Microbiology and Biotechnology*, vol. 64, no. 6, pp. 763–781, 2004.

[2] D. L. Ollis, E. Cheah, M. Cygler et al., "The alpha/beta hydrolase fold," *Protein Engineering*, vol. 5, no. 3, pp. 197–211, 1992.

[3] T. Norin and F. Hæffner, "Molecular modelling of lipase catalysed reactions. Prediction of enantioselectivities," *Chemical and Pharmaceutical Bulletin*, vol. 47, no. 5, pp. 591–600, 1999.

[4] K. E. Jaeger, S. Ransac, B. W. Dijkstra, C. Colson, M. Van Heuvel, and O. Misset, "Bacterial lipases," *FEMS Microbiology Reviews*, vol. 15, no. 1, pp. 29–63, 1994.

[5] S. D'Amico, P. Claverie, T. Collins et al., "Molecular basis of cold adaptation," *Philosophical Transactions of the Royal Society B*, vol. 357, no. 1423, pp. 917–925, 2002.

[6] H. Lilie, W. Haehnel, R. Rudolph, and U. Baumann, "Folding of a synthetic parallel β-roll protein," *FEBS Letters*, vol. 470, no. 2, pp. 173–177, 2000.

[7] V. Spiwok, P. Lipovová, T. Skálová et al., "Cold-active enzymes studied by comparative molecular dynamics simulation," *Journal of Molecular Modeling*, vol. 13, no. 4, pp. 485–497, 2007.

[8] T. Collins, F. Roulling, F. Piette et al., "Fundamentals of cold-adapted enzymes," in *Psychrophiles: From Biodiversity to Biotechnology*, R. Margesin, F. Schinner, J. C. Marx, and C. Gerday, Eds., pp. 211–227, Springer, Berlin, Germany, 2008.

[9] D. I. Paredes, K. Watters, D. J. Pitman, C. Bystroff, and J. S. Dordick, "Comparative void-volume analysis of psychrophilic and mesophilic enzymes: structural bioinformatics of psychrophilic enzymes reveals sources of core flexibility," *BMC Structural Biology*, vol. 11, no. 1, p. 42, 2011.

[10] E. Krieger, G. Vriend, and C. Spronk, "YASARA-Yet Another Scientific Artificial Reality Application".

[11] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[12] S. Subramaniam, "The Biology Workbench—a seamless database and analysis environment for the biologist," *Proteins*, vol. 32, no. 1, pp. 1–2, 1998.

[13] R. Meier, T. Drepper, V. Svensson, K. E. Jaeger, and U. Baumann, "A calcium-gated lid and a large β-roll sandwich are revealed by the crystal structure of extracellular lipase from *Serratia marcescens*," *Journal of Biological Chemistry*, vol. 282, no. 43, pp. 31477–31483, 2007.

[14] C. Angkawidjaja, D. J. You, H. Matsumura et al., "Crystal structure of a family I.3 lipase from *Pseudomonas* sp. MIS38 in a closed conformation," *FEBS Letters*, vol. 581, no. 26, pp. 5060–5064, 2007.

[15] D. Eisenberg, R. Lüthy, and J. U. Bowie, "VERIFY3D: assessment of protein models with three-dimensional profiles," *Macromolecular Crystallography B*, vol. 277, pp. 396–404, 1997.

[16] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of polypeptide chain configurations," *Journal of molecular biology*, vol. 7, pp. 95–99, 1963.

[17] D. C. Rapaport, *The Art of Molecular Dynamics Simulation*, Cambridge University Press, Cambridge, UK, 2004.

[18] Y. Duan, C. Wu, S. Chowdhury et al., "A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations," *Journal of Computational Chemistry*, vol. 24, no. 16, pp. 1999–2012, 2003.

[19] T. Xu, B. Gao, L. Zhang, J. Lin, X. Wang, and D. Wei, "Template-based modeling of a psychrophilic lipase: conformational changes, novel structural features and its application in predicting the enantioselectivity of lipase catalyzed transesterification of secondary alcohols," *Biochimica et Biophysica Acta*, vol. 1804, no. 12, pp. 2183–2190, 2010.

[20] D. Pascale, A. M. Cusano, F. Autore et al., "The cold-active Lip1 lipase from the Antarctic bacterium Pseudoalteromonas haloplanktis TAC125 is a member of a new bacterial lipolytic enzyme family," *Extremophiles*, vol. 12, no. 3, pp. 311–323, 2008.

[21] J. Pleiss, M. Fischer, and R. D. Schmid, "Anatomy of lipase binding sites: the scissile fatty acid binding site," *Chemistry and Physics of Lipids*, vol. 93, no. 1-2, pp. 67–80, 1998.

[22] I. Linhartová, L. Bumba, J. Mašín et al., "RTX proteins: a highly diverse family secreted by a common mechanism," *FEMS Microbiology Reviews*, vol. 34, no. 6, pp. 1076–1112, 2010.

[23] K. A. McCall, C. C. Huang, and C. A. Fierke, "Function and mechanism of zinc metalloenzymes," *Journal of Nutrition*, vol. 130, no. 5, pp. 1437S–1446S, 2000.

[24] B. A. Tejo, A. B. Salleh, and J. Pleiss, "Structure and dynamics of Candida rugosa lipase: the role of organic solvent," *Journal of Molecular Modeling*, vol. 10, no. 5-6, pp. 358–366, 2004.

[25] M. Giovanola, F. D'Antoni, M. Santacroce et al., "Role of a conserved glycine triplet in the NSS amino acid transporter KAAT1," *Biochimica Et Biophysica Acta*, vol. 1818, no. 7, pp. 1737–1744, 2012.

[26] G. Colombo, G. Ottolina, and G. Carrea, "Modelling of enzyme properties in organic solvents," *Monatshefte fur Chemie*, vol. 131, no. 6, pp. 527–547, 2000.

[27] W. C. Choi, H. K. Myung, H. S. Ro, R. R. Sang, T. K. Oh, and J. K. Lee, "Zinc in lipase L1 from *Geobacillus stearothermophilus* L1 and structural implications on thermal stability," *FEBS Letters*, vol. 579, no. 16, pp. 3461–3466, 2005.

[28] M. El Khattabi, P. Van Gelder, W. Bitter, and J. Tommassen, "Role of the calcium ion and the disulfide bond in the *Burkholderia glumae* lipase," *Journal of Molecular Catalysis B*, vol. 22, no. 5-6, pp. 329–338, 2003.

[29] A. B. Salleh, A. S. M. A. Rahim, R. N. Z. R. A. Rahman, T. C. Leow, and M. Basri, "The role of Arg157Ser in improving the compactness and stability of ARM lipase," *Journal of Computer Science Systems Biology*, vol. 5, no. 2, pp. 039–046, 2012.

[30] H. Matsumura, T. Yamamoto, T. C. Leow et al., "Novel cation-π interaction revealed by crystal structure of thermoalkalophilic lipase," *Proteins*, vol. 70, no. 2, pp. 592–598, 2008.

[31] T. C. Leow, R. N. Z. R. A. Rahman, M. Basri, and A. B. Salleh, "A thermoalkaliphilic lipase of Geobacillus sp. T1," *Extremophiles*, vol. 11, no. 3, pp. 527–535, 2007.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 414327, 8 pages
http://dx.doi.org/10.1155/2013/414327

Research Article

A Novel Method of Predicting Protein Disordered Regions Based on Sequence Features

Tong-Hui Zhao,[1,2] Min Jiang,[3] Tao Huang,[4] Bi-Qing Li,[5] Ning Zhang,[6] Hai-Peng Li,[7] and Yu-Dong Cai[1,2]

[1] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[2] *Department of Mathematics, College of Science, Shanghai University, Shanghai 200444, China*
[3] *State Key Laboratory of Medical Genomics, Institute of Health Sciences, Shanghai Jiaotong University School of Medicine and Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200025, China*
[4] *Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA*
[5] *Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[6] *Department of Biomedical Engineering, Tianjin University, Tianjin Key Lab of BME Measurement, Tianjin 300072, China*
[7] *CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*

Correspondence should be addressed to Hai-Peng Li; lihaipeng@picb.ac.cn and Yu-Dong Cai; cai_yud@yahoo.com.cn

Received 3 March 2013; Accepted 26 March 2013

Academic Editor: Bin Niu

With a large number of disordered proteins and their important functions discovered, it is highly desired to develop effective methods to computationally predict protein disordered regions. In this study, based on Random Forest (RF), Maximum Relevancy Minimum Redundancy (mRMR), and Incremental Feature Selection (IFS), we developed a new method to predict disordered regions in proteins. The mRMR criterion was used to rank the importance of all candidate features. Finally, top 128 features were selected from the ranked feature list to build the optimal model, including 92 Position Specific Scoring Matrix (PSSM) conservation score features and 36 secondary structure features. As a result, Matthews correlation coefficient (MCC) of 0.3895 was achieved on the training set by 10-fold cross-validation. On the basis of predicting results for each query sequence by using the method, we used the scanning and modification strategy to improve the performance. The accuracy (ACC) and MCC were increased by 4% and almost 0.2%, respectively, compared with other three popular predictors: DISOPRED, DISOclust, and OnD-CRF. The selected features may shed some light on the understanding of the formation mechanism of disordered structures, providing guidelines for experimental validation.

1. Introduction

The protein structure-function paradigm has been believed as a dogma in the 20th century. However, the discovery of intrinsically disordered proteins, which have regions devoid of stable secondary structures or have a large number of conformations [1], challenges the traditional view and calls for reassessment of the paradigm.

Eukaryotic proteins apparently have more intrinsic disordered regions than those of bacteria or archaea [2], suggesting also more important functions such as being involved in signaling and regulation of gene expression [3]. Lack of intrinsic structures could render protein additional functions, including binding to different targets [4], transcriptional regulation, translational regulation, and cellular signal transduction regulation [5].

Although there is a growing amount of disordered proteins discovered or shown to have disordered regions under physiological conditions [6], most of them were poorly detected by experimental approaches [2, 4, 5, 7–9]. Firstly, such experimental methods are often time consuming and expensive. Furthermore, it is believed, in X-ray crystallography, that regions missing electron density were related to disorder in many protein structures [6]. However, without

additional experiments, it is not sure whether a low electron-density region is intrinsically disordered or is a wobbly domain, or just the result of technical difficulties [2]. NMR spectroscopy, one of the most readily suited techniques for detecting disordered proteins in solution, could also underrepresent a native molten globular domain, which is one of the types of disordered regions [2].

Generally speaking, intrinsically disordered proteins have a biased amino acid composition. Weathers and colleagues reported that amino acid composition was sufficient to be used to accurately recognize disorder [10]. Several algorithms for predicting intrinsically disordered proteins have been developed, such as DISOPRED [11], DISOclust [12], and OnD-CRF [13]. DISOPRED is a web service, which is trained on high resolution X-ray crystal structures and identifies disorder when the electron density map of a residue misses coordinates. It is initially generated sequence profile by a PSI BLAST [14] searching. After being trained using a support vector algorithm, the classifier can output a probability estimate. However, a limitation of this algorithm is that coordinates missing may be caused by the artifact of the crystallization process rather than disorder. DISOclust, based on analysis of three-dimensional structure models, identifies disorder when residues change or are consistently missing. OnD-CRF is a method for predicting the transition between structured and disordered regions. The approach uses conditional random fields relying on features derived from amino acid sequences and secondary structure prediction results.

In the present study, we developed a new strategy for analyzing and predicting protein disordered regions by means of Random Forest (RF), Maximum Relevancy Minimum Redundancy (mRMR), Incremental Feature Selection (IFS), and a scanning and modification strategy. Optimal feature set was selected from candidate features, containing Position Specific Scoring Matrix (PSSM) conservation score features and secondary structure features. Our method outperformed other three existing disorder predictors achieving the highest ACC and MCC values.

2. Materials and Methods

2.1. Benchmark Dataset. In this study, disordered proteins were downloaded from the Database of Protein Disorder (DisProt) (version 4.9) [8], which is constructed based on literature description, providing structured and functional information for intrinsically disordered proteins. Ordered proteins were collected from DisProt database and PDB-Select-25 (the October 2008 version) [15]. PDB-Select-25 is a representative subset of the Protein Data Bank (PDB), containing protein families less than 25% sequence identity [16]. Data was preprocessed according to the following criteria. (i) Only disordered protein chains having more than 50 residues and only proteins with low resolution (≥2 Å) were retained. (ii) Only chains having no missing backbones or side chain coordinates were retained. Finally, 960 protein chains containing 293,780 residues were obtained, in which 55,637 residues were in disordered regions. All protein chains were divided randomly into training set and test set.

A 21-residue sliding window approach was employed along each of the protein sequence, containing the center ordered or disordered residue and 10 residues upstream and downstream of the center residue. Since the dataset used in this study was an unbalanced dataset with much more ordered samples than disordered ones, for the training set, we randomly selected the equal number of ordered samples to match the disordered ones. Finally, 43,903 ordered samples and 43,903 disordered samples from 753 proteins were obtained in the training set, which can be found in Online Supporting Information S1 available online at http://dx.doi.org/10.1155/2013/414327. The test set contained 54,582 ordered samples and 11,734 disordered samples from 192 proteins, which were given in Online Supporting Information S2.

2.2. Feature Extraction

2.2.1. Feature of PSSM Conservation Scores. Evolutionary conservation is considered important in biological sequence analysis. A more conserved residue within a protein sequence indicates that it is under stronger selective pressure and hence more important for the protein function. Mutations on such residues may cause significant changes of the protein. In view of this, we used conservation scores to encode peptides.

Herein, the Position Specific Iterative BLAST (PSI BLAST) was employed to measure the conservation status for a specific residue. For each residue, a 20-dimensional vector was calculated to denote the conservation probabilities of mutations to 20 basic amino acids. For a given protein sequence, a Position Specific Scoring Matrix (PSSM [17]) was obtained, which was constructed by all vectors of all residues in the sequence.

2.2.2. Feature of Secondary Structures. Intrinsically disordered proteins are devoid of well-defined tertiary structures under physiological conditions; however, generally speaking, they often display signs of local secondary structures [18, 19]. After statistical analysis of complex of 24 intrinsically disordered proteins, Fuxreiter et al. [20] found that some regions in disordered proteins had strong preference for helical structures. Therefore, in this study, each amino acid was encoded as three types of secondary structures: helix, strand, or coil, as predicted by SSPro [21]. Helix, strand, and coil are the three major kinds of protein secondary structures. Helix is the protein region with spiral conformation. Strand is a protein structural unit of twisted, pleated sheet. The coil region is the region that does not belong to helix or strand. SSPro predicts the protein secondary structures based on PSI-BLAST profiles with an ensemble neural network model [21].

Thus, each 21-residue peptide was encoded into a vector containing $(20+3)\times 21 = 483$ features. The features are named with following rules: first, the amino acid position ("AA" with position), then, feature types ("PSSM" and "SS"), and last, detail information. For PSSM features, it is the amino acid type. For secondary structure (SS) features, it is the secondary structure code. In secondary structure code, H, E, and C strand for helix, strand, and coil, respectively.

2.3. Maximum Relevancy Minimum Redundancy (mRMR). A classification model containing more features may not have more discriminating power. Additional features may have detrimental effects on the classification such as slowing down the learning process and causing overfitting the training data. It is believed that feature selection is an effective way of reducing the dimension of the feature space to improve the prediction performance.

The Maximum Relevancy Minimum Redundancy (mRMR) method was used in this study to select an optimal feature subset. The mRMR was originally developed by Peng et al. [22] to deal with the microarray data processing. If a feature had better tradeoff between maximum relevance to the target and minimum redundancy among other features, it was deemed as a better feature and would be ranked first (with a smaller index) in the final ordered list. The algorithm is described briefly below.

To determine the relevance properties of the feature space, the mutual information (MI), denoted as I, is defined as

$$I(x, y) = \iint P(x, y) \log \frac{P(x, y)}{P(x)\,P(y)} dx\, dy, \quad (1)$$

where x and y are two random variables. $P(x, \mathrm{y})$ is the joint probabilistic density function of x and y. $P(x)$ and $P(y)$ are the margin probabilistic density functions of x and y, respectively. To calculate MI, the joint probabilistic density function $P(x, y)$ and the margin probabilistic density functions $P(x)$ and $P(y)$ should be given in advance.

Suppose G denotes the entire feature space; we aim to find a subset S of the features to satisfy both maximum relevance and minimum redundancy.

Based on MI, the following mRMR function is constructed:

$$\max_{f_j \in \Omega_t} \left[I(f_j, c) - \frac{1}{m} \sum_{f_i \in \Omega_s} I(f_j, f_i) \right] \quad (j = 1, 2, \ldots, n), \quad (2)$$

where Ω_s is the already selected feature set and Ω_t is the to-be-selected feature set, and m and n are the sizes of these two feature sets, respectively. The higher the ordered rank is, the more important the feature is.

A parameter is introduced here to deal with the continuous variables. In our study, t was assigned to be 1. Finally, an ordered feature list was obtained in which better features had smaller indexes.

The mRMR software could be obtained from http://penglab.janelia.org/proj/mRMR/.

2.4. The Random Forest (RF) Method. The Random Forest (RF) algorithm, firstly introduced by Svetnik [23] in 2003, is a combining ensemble tree-structured classifier. The individual decision tree in the forest depends on a random vector and has independent identically distribution. The Random Forest has been widely used in various fields such as economics and medical and text categorization. It has been also successfully employed in biological prediction problems [24–26] and even can efficiently handle large-scale dataset.

In our research, we use the Random Forest (RF) algorithm to construct a prediction model to predict whether an amino acid is in disordered region or not. The method is briefly introduced as follows.

Firstly, 10 decision trees are grown according to the following criteria.

(a) Suppose the number of cases in the training data is M; sample M cases randomly with replacement from the original data to keep the size of the original data not changing.

(b) When dealing with each note, n predictors are selected randomly in terms of N features (where $n \ll N$). The split on the n predictors is also implemented to split the corresponding note. The m value is set to constant.

(c) Each tree is grown as large as possible and unnecessarily pruned. Then each tree gives the queried input a classification. Finally, the forest will choose the one that has the most votes among the trees.

In this study, the Random Forest classifier in Weka was employed with default parameters. The WEKA program is available at http://www.cs.waikato.ac.nz/ml/weka/index_downloading.html.

2.5. The Cross-Validation Method. In the literature, cross-validation methods are used to evaluate the stability of a predictor. The independent dataset test, subsampling (k-fold cross-validation), and jackknife analysis are the three methods generally used [27]. For a given benchmark dataset, the jackknife test generates a unique outcome and is deemed as the most objective one compared to other two methods, as elucidated in [28, 29] and demonstrated by [30, equations (28)–(32)] in. However, to reduce the computational time, in this study, 10-fold cross-validation test was used instead of jackknife test. During the 10-fold cross-validation, the whole dataset is divided into 10 equal parts. Each part is in turn used as test set and the remaining 9 parts as training set. We introduced prediction accuracy (ACC), specificity (SP), sensitivity (SN), and Matthews correlation coefficient (MCC) to evaluate the performance of the predictor, which are calculated as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}},$$

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}},$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}}, \quad (3)$$

where TP, TN, FP, and FN stand for the number of true positive, true negative, and false positive, false negative samples, respectively.

2.6. Incremental Feature Selection (IFS). The incremental feature selection (IFS) [31–33] procedure was used to find an optimal subset from the mRMR feature list generated above. Suppose the total number of the features is N; we can obtain N feature subsets which are initiated from a subset containing one feature and generated by adding them one by one from the mRMR feature list.

The ith subset is denoted by

$$S_i = \{f_1, f_2, \ldots, f_i\} \quad (1 \le i \le N). \tag{4}$$

Based on the N feature subsets, N Random Forest predictors were constructed with 10-fold cross-validation evaluating its performance. Then the IFS curve of MCC to the feature subset index i was plotted, in which the peak point was noted as h. Finally an optimal feature subset was obtained with which the corresponding predictor yields the best MCC.

3. Results and Discussion

3.1. Feature Reduction. We calculated the Cramer's V coefficient [34, 35] between features and targets. The Cramer's V coefficient is a statistical measurement derived from the Pearson chi-square test [36]. It ranges from 0 to 1 with smaller value indicating weaker association. Features with Cramer's V coefficient less than 0.1 were removed. After this procedure, 175 features remained containing 112 PSSM conservation features and 63 secondary structure features, which can be found in Online Supporting Information S3.

3.2. The mRMR Result. Two kinds of outcomes were obtained after executing the mRMR program. One was called "MaxRel feature list" that ranked the features according to the relevance to the target; the other was named "mRMR feature list" that ranked the features based on the criteria of maximum relevance and minimum redundancy. In our research, only the "mRMR feature list" was used to select optimal feature subset in the IFS procedure. It was listed in Online Supporting Information S4.

3.3. IFS and Optimal Feature Subset. 175 predictors were constructed based on the 175 feature subsets in the IFS procedure. Prediction results of the predictors were listed in Online Supporting Information S5 and the IFS curve was plotted in Figure 1 in which the MCC reached the topmost 0.3895 with 128 features on the training set. Thus, the top 128 features were considered as the optimal feature subset and were used to construct the final predictor. The 128 features were given in Online Supporting Information S6. The MCC of the predictor on independent test set was 0.2791.

3.4. Feature Analysis. The distribution of the feature types in the final optimal feature set was shown in Figure 2. In the 128 optimal features, 92 were from PSSM conservation scores and 36 from secondary structure features (Figure 2(a)). The two types of features contributed to the prediction. It can be seen from the site-specific distribution of the optimal feature set (Figure 2(b)) that features at sites 8–14 played important roles. In addition, features at sites 1-2, sites 5-6, and sites 15–17, 19, and 21 also had considerable impacts on the prediction of disordered protein.

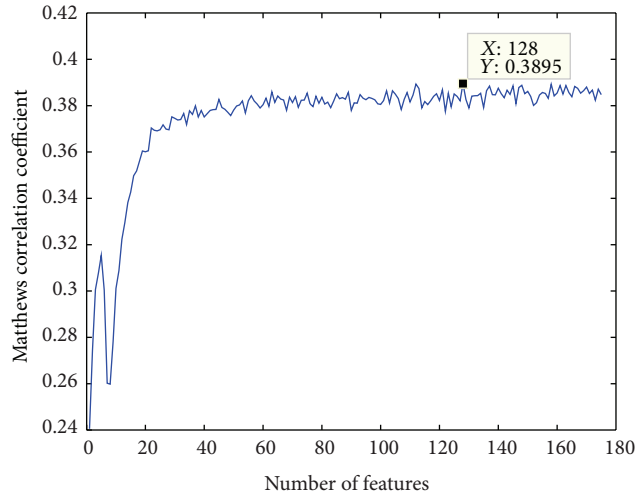


Figure 1: The IFS curve showing the Matthews correlation coefficient (MCC) against the number of features. The details were given in Online Supporting Information S5. With the top 128 features, the MCC on training set by 10-fold cross-validation takes the peak 0.3895.

3.5. PSSM Conservation Score Feature Analysis. As mentioned above, among the 128 optimal features, 92 belonged to the PSSM conservation scores, accounting for the most. Mutations to 20 different amino acids could have different impacts in determining the disordered regions. It can be clearly seen from Figure 3(a) that only 8 out of 20 amino acid mutations were affected. In this regard, the amino acid P (Proline) or S (Serine) could impact most, successively followed by K (Lysine), Q (Glutamine), and so forth. Interestingly, it has been reported that Q was overrepresented in protein interaction domains [37]. It was recently reported that the Ure2p prion and other Q/N-rich yeast prion proteins, which were completely disordered, were driven to format amyloid primarily by intermolecular interactions [38]. Meanwhile, as shown in Figure 3(b), for the 21-length peptides, PSSM conservation scores at sites 8–15 played the most important role. Furthermore, 6 out of the top 10 features in the optimal feature list were PSSM conservation features. The first one was the conservation feature against residue K (Lys) at site 6 (index 2, "AA6_PSSM-12-K"). The other 5 were conservation features against residues E, P, and K at sites 7, 1, and 8, respectively (index 5, "AA7_PSSM-7-E", index 6, and index 7, "AA1_PSSM-15-P" and "AA8_PSSM-12-K") and conservation features against residue E, D at site 21 and site 15 (index 6 and index 7, "AA21_PSSM-7-E" and "AA15_PSSM-4-D").

3.6. Secondary Structure Feature Analysis. The feature subtypes and site-specific distributions of the secondary structure features in the optimal feature set were plotted in Figure 4. From Figure 4, it can be seen that features of "coil"

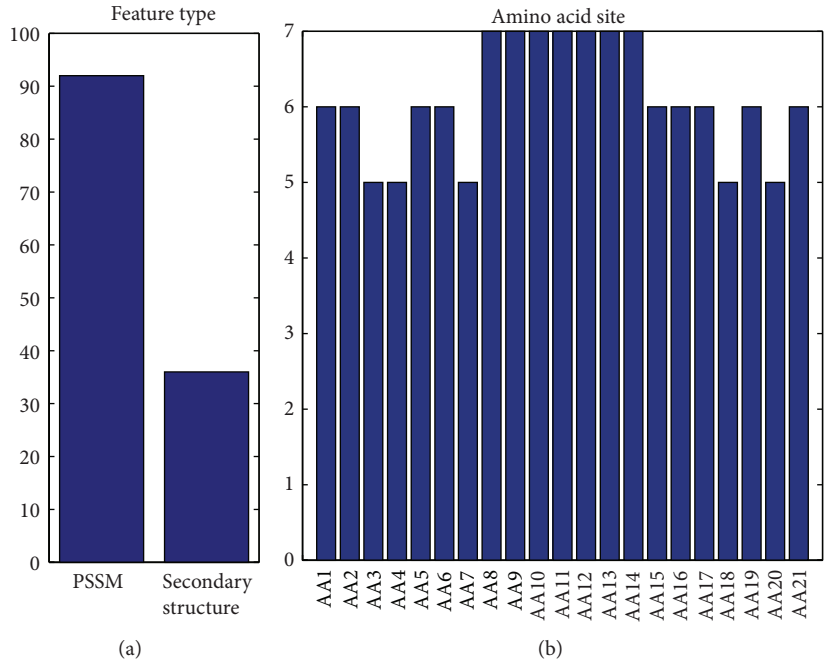


Figure 2: The distribution of feature types and amino acid sites in optimal feature subset. The histograms show the number of each type and each site of features in optimal feature subset. In (a), there are 92 PSSM features and 36 secondary structure features. (b) provides the site distributions of the features in the optimal feature set.

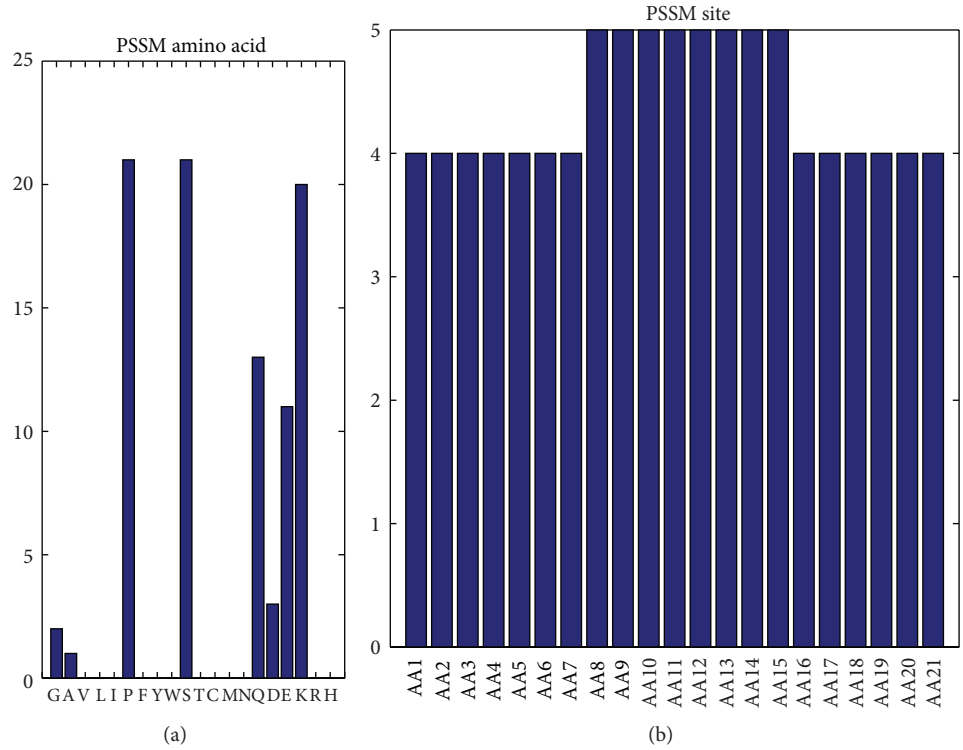


Figure 3: The distribution of amino acid compositions and sites on PSSM conservation feature. The histograms reveal the types and site distributions of PSSM features in the optimal feature set. (a) indicates the effects on prediction of mutations to 20 different amino acids. (b) provides the site distributions of the PSSM features in the optimal feature set.

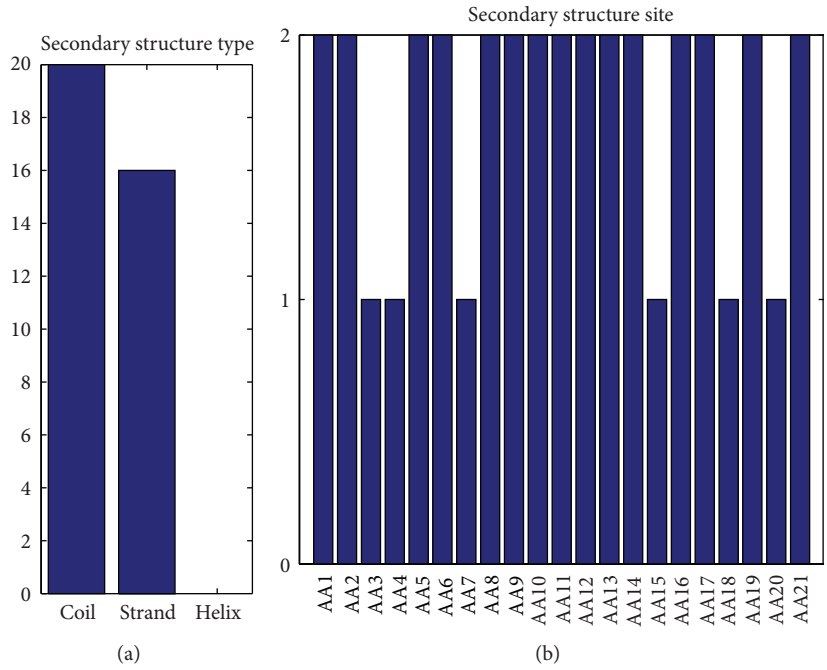


Figure 4: The distribution of secondary structure types and amino acid sites on secondary structure feature. The histograms give the types and site distributions of secondary structure features in the final optimal feature set. (a) indicates the effects on prediction of three different types of secondary structures: coil, strand, and helix. (b) provides the site distributions of the secondary structure features in the optimal feature set.

and "strand" did affect the disorder (Figure 4(a)). And the "coil" feature was affected the most, followed by the "strand" feature. The secondary structure features at 15 out of 21 sites had relatively more impact than the left 5 (Figure 4(b)). Intrinsically disordered and aggregation-prone domains exist within the very diverse set of human extracellular matrix protein [39]. Recently Evans reported that the aragonite extracellular matrix proteins (AECMPs) had evolved signature molecular traits of intrinsically disordered and aggregation-prone "interactive" sequences that enabled matrix assembly [40]. It was also reported that cyclization of the skeletal DHPR II-III loop affected the secondary structures and the dynamic properties of the helical A/B region as well as the critical C region. These structural effects were correlated with a change in vitro activation profile of the RyR1 and with an interaction with DHPR II-III loop α-helical recognition sites in the SPRY2 domain of RyR1 [41]. So it is believed the sequence location and number of intrinsically disordered and secondary motifs may be important for aggregation, protein orientation, and assembly stability and may also play a role in the recognition and interaction between proteins with other specific component(s).

3.7. Scan the Entire Protein Sequence to Refine the Disordered Region Prediction. The prediction result of the predictor constructed based on the 128 optimal features on the independent test was shown in Online Supporting Information S7. The third column, *predicted*, was the prediction result where "1" indicated the residue was in ordered region while "2" denoted the residue was in disordered region. It can be seen that many ordered (1) sites were wrong predicted as disordered (2), resulting in short disordered segments (2) being inserted in an ordered segment (1), and vice versa. Therefore, we used a scanning method to refine the prediction results according to the following criteria [42]. (i) Any predicted disordered sites (2) were refined to ordered (1) if there were more than 4 continuous "1s" upstream of the site but less than 4 continuous "2s" downstream of it. (ii) Any predicted ordered sites (1) were changed to disordered (2) if there were more than 4 continuous "2s" upstream of the site but less than 3 continuous "1s" downstream of it. After the refinement procedure, the performance improved much as shown in Table 1. The scanning results can be found in the last column, *scanning*, also can be found in Online Supporting Information S7.

3.8. Comparison with the Existing Methods. Our method was compared with three other existing methods, DISOPRED, DISOclust, and OnD-CRF. The DISOPRED server allows users to submit a protein sequence and returns a probability estimate of being disordered of each residue in the sequence [11]. In the prediction results by DISOPRED, disordered residues were marked with asterisks ($*$) and ordered residues were marked with dots ($\cdot$). The prediction results by DISOclust were formulated by a series of "D" and "O," denoting the residues being in disordered region and ordered region, respectively. The predicting result by Ond-CRF only delivers users the information of disordered regions. As a result, our method outperformed the other three existing methods. As shown in Table 1, the ACC and MCC were improved to

Table 1: The evaluation of prediction result on independent test set by different methods.

Method	Accuracy (ACC)	Matthews correlation coefficient (MCC)	Sensitivity (SN)	Specificity (SP)
Before scanning	0.7028	0.2791	0.7189	0.6281
After scanning	0.7508	0.3304	0.7806	0.6118
DISOPRED	0.7173	0.3285	0.7239	0.6864
DISOclust	0.6650	0.3105	0.6453	0.7570
OnD-CRF	0.6562	0.3228	0.6265	0.7941

a certain extent, at least 4% increase on ACC and almost 0.2% increase on MCC. It is suggested that our method is pretty more effective than other methods on prediction of intrinsically disordered protein region.

3.9. Useful Insights for Guiding Experiments or Being Validated by Experiments. About 50% of human proteins were previously predicted to contain at least one larger disordered region, and it was shown that the main reason for the existence of such regions was to harbor binding sites [43]. In this study, the selected features at different sites could provide insights for researchers to find or validate new disordered protein or disordered regions, as can be seen from the following two aspects. (i) PSSM feature: it was found from the results that the PSSM conservation score that mutates to amino acid P or S had the most impact. Besides, mutations to amino acids K, Q, and E also had more impacts than others. However mutations to other 12 amino acids were affected little. For example, phosphorylation of Ser66 in the intrinsically disordered N-terminal region of AtREM1.3 weakened the interaction strength with importin α proteins, indicating a regulatory domain in the N-terminal region stabilizing the interactions [44]. (ii) Secondary structure feature: it was found in our optimal feature set that the second structure feature at site 11 had the ranking index 1, implying the most important role to the prediction. Interestingly, it has been reported that disorder regions often correlated domain boundaries where usually harbor some coil structures [45]. Accordingly, other features in the optimal feature set are certainly worth being further investigated by future experiments.

4. Conclusion

The plasticity of disordered regions provides interaction capacity. In this study, we investigated important features for predicting disordered protein regions. As a result, the PSSM conservation scores and the second structures are two types of important features, which play key important roles in determining disordered regions. Among these, only 8 amino acids play major roles. The coil and strand structures also affected the prediction. These may provide additional insight into disordered proteins.

Authors' Contribution

T.-H. Zhao, M. Jiang and T. Huang contributed equally to this work.

Acknowledgments

This work was supported by Grants from the National Basic Research Program of China (2011CB510102, 2011CB510101), and the Innovation Program of Shanghai Municipal Education Commission (12ZZ087) and the grant of "The First-class Discipline of Universities in Shanghai."

References

[1] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell, "Protein disorder prediction: implications for structural proteomics," *Structure*, vol. 11, no. 11, pp. 1453–1459, 2003.

[2] A. K. Dunker, J. D. Lawson, C. J. Brown et al., "Intrinsically disordered protein," *Journal of Molecular Graphics and Modelling*, vol. 19, no. 1, pp. 26–59, 2001.

[3] K. Nishikawa, "Natively unfolded proteins: an overview," *Biophysics*, vol. 5, pp. 53–58, 2009.

[4] P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm," *Journal of Molecular Biology*, vol. 293, no. 2, pp. 321–331, 1999.

[5] H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 3, pp. 197–208, 2005.

[6] V. N. Uversky, "Natively unfolded proteins: a point where biology waits for physics," *Protein Science*, vol. 11, no. 4, pp. 739–756, 2002.

[7] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradović, "Intrinsic disorder and protein function," *Biochemistry*, vol. 41, no. 21, pp. 6573–6582, 2002.

[8] M. Sickmeier, J. A. Hamilton, T. LeGall et al., "DisProt: the database of disordered proteins," *Nucleic Acids Research*, vol. 35, supplement 1, pp. D786–D793, 2007.

[9] D. Eliezer, "Biophysical characterization of intrinsically disordered proteins," *Current Opinion in Structural Biology*, vol. 19, no. 1, pp. 23–30, 2009.

[10] E. A. Weathers, M. E. Paulaitis, T. B. Woolf, and J. H. Hoh, "Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein," *FEBS Letters*, vol. 576, no. 3, pp. 348–352, 2004.

[11] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, and D. T. Jones, "The DISOPRED server for the prediction of protein disorder," *Bioinformatics*, vol. 20, no. 13, pp. 2138–2139, 2004.

[12] L. J. Mcguffin, "Intrinsic disorder prediction from the analysis of multiple protein fold recognition models," *Bioinformatics*, vol. 24, no. 16, pp. 1798–1804, 2008.

[13] L. Wang and U. H. Sauer, "OnD-CRF: predicting order and disorder in proteins conditional random fields," *Bioinformatics*, vol. 24, no. 11, pp. 1401–1402, 2008.

[14] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[15] U. Hobohm and C. Sander, "Enlarged representative set of protein structures," *Protein Science*, vol. 3, no. 3, pp. 522–524, 1994.

[16] T. Huang, Z. S. He, W. R. Cui et al., "A sequence-based approach for predicting protein disordered regions," *Protein and Peptide Letters*, vol. 20, no. 3, pp. 243–248, 2013.

[17] S. Ahmad and A. Sarai, "PSSM-based prediction of DNA binding sites in proteins," *BMC Bioinformatics*, vol. 6, no. 1, article 33, 2005.

[18] I. Radhakrishnan, G. C. Pérez-Alvarado, H. J. Dyson, and P. E. Wright, "Conformational preferences in the Ser133-phosphorylated and non-phosphorylated forms of the kinase inducible transactivation domain of CREB," *FEBS Letters*, vol. 430, no. 3, pp. 317–322, 1998.

[19] M. K. Yoon, V. Venkatachalam, A. Huang, B. S. Choi, C. M. Stultz, and J. J. Chou, "Residual structure within the disordered C-terminal segment of p21 Waf1/Cip1/Sdi1 and its implications for molecular recognition," *Protein Science*, vol. 18, no. 2, pp. 337–347, 2009.

[20] M. Fuxreiter, I. Simon, P. Friedrich, and P. Tompa, "Preformed structural elements feature in partner recognition by intrinsically unstructured proteins," *Journal of Molecular Biology*, vol. 338, no. 5, pp. 1015–1026, 2004.

[21] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles," *Proteins: Structure, Function and Genetics*, vol. 47, no. 2, pp. 228–235, 2002.

[22] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[23] V. Svetnik, A. Liaw, C. Tong, J. Christopher Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[24] A. Dehzangi, S. Phon-Amnuaisuk, and O. Dehzangi, "Using random forest for protein fold prediction problem: an empirical study," *Journal of Information Science and Engineering*, vol. 26, no. 6, pp. 1941–1956, 2010.

[25] K. K. Kandaswamy, K. C. Chou, T. Martinetz et al., "AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties," *Journal of Theoretical Biology*, vol. 270, no. 1, pp. 56–62, 2011.

[26] Z. P. Liu, L. Y. Wu, Y. Wang, X. S. Zhang, and L. Chen, "Prediction of protein-RNA binding sites by a random forest method with combined features," *Bioinformatics*, vol. 26, no. 13, pp. 1616–1622, 2010.

[27] K. C. Chou and C. T. Zhang, "Prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 4, pp. 275–349, 1995.

[28] K. C. Chou and H. B. Shen, "Recent progress in protein subcellular location prediction," *Analytical Biochemistry*, vol. 370, no. 1, pp. 1–16, 2007.

[29] K. C. Chou and H. B. Shen, "Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms," *Nature Protocols*, vol. 3, no. 2, pp. 153–162, 2008.

[30] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[31] Z. He, J. Zhang, X. H. Shi et al., "Predicting drug-target interaction networks based on functional groups and biological features," *PLoS ONE*, vol. 5, no. 3, Article ID e9603, 2010.

[32] B. Q. Li, L. L. Hu, S. Niu, Y. D. Cai, and K. C. Chou, "Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches," *Journal of Proteomics*, vol. 75, no. 5, pp. 1654–1665, 2011.

[33] T. Huang, L. Chen, Y. D. Cai, and K. C. Chou, "Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property," *PloS ONE*, vol. 6, no. 9, Article ID e25297, 2011.

[34] H. Cramér, *Methods of Mathematical Statistics*, vol. 23, Princeton University Press, Princeton, NJ, USA, 1946.

[35] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics. Volume 2: Inference and: Relationship*, Griffin Printing, Santa Ana, Calif, USA, 1973.

[36] K. M. D. Harrison, T. Kajese, H. I. Hall, and R. Song, "Risk factor redistribution of the national HIV/AIDS surveillance data: an alternative approach," *Public Health Reports*, vol. 123, no. 5, pp. 618–627, 2008.

[37] S. Roy, D. Martinez, H. Platero, T. Lane, and M. Werner-Washburne, "Exploiting amino acid composition for predicting protein-protein interactions," *PLoS ONE*, vol. 4, no. 11, Article ID e7813, 2009.

[38] S. Ngo, V. Chiang, E. Ho, L. Le, and Z. Guo, "Prion domain of yeast Ure2 protein adopts a completely disordered structure: a solid-support EPR study," *PloS ONE*, vol. 7, no. 10, Article ID e47248, 2012.

[39] F. Peysselon, B. Xue, V. N. Uversky, and S. Ricard-Blum, "Intrinsic disorder of the extracellular matrix," *Molecular BioSystems*, vol. 7, no. 12, pp. 3353–3365, 2011.

[40] J. S. Evans, "Aragonite-associated biomineralization proteins are disordered and contain interactive motifs," *Bioinformatics*, vol. 28, no. 24, pp. 3182–3185, 2012.

[41] H. S. Tae, Y. Cui, Y. Karunasekara, P. G. Board, A. F. Dulhunty, and M. G. Casarotto, "Cyclization of the intrinsically disordered α1S dihydropyridine receptor II-III loop enhances secondary structure and in vitro function," *The Journal of Biological Chemistry*, vol. 286, no. 25, pp. 22589–22599, 2011.

[42] B. Q. Li, L. L. Hu, L. Chen, K. Y. Feng, Y. D. Cai, and K. C. Chou, "Prediction of protein domain with mRMR feature selection and analysis," *PloS One*, vol. 7, no. 6, Article ID e39308, 2012.

[43] B. Meszaros, I. Simon, and Z. Dosztanyi, "Prediction of protein binding regions in disordered proteins," *PLoS Computational Biology*, vol. 5, no. 5, Article ID e1000376, 2009.

[44] M. Marin, V. Thallmair, and T. Ott, "The intrinsically disordered N-terminal region of AtREM1. 3 remorin protein mediates protein-protein interactions," *The Journal of Biological Chemistry*, vol. 287, no. 47, pp. 39982–39991, 2012.

[45] N. Nagarajan and G. Yona, "Automatic prediction of protein domains from sequence information using a hybrid learning system," *Bioinformatics*, vol. 20, no. 9, pp. 1335–1360, 2004.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 287019, 9 pages
http://dx.doi.org/10.1155/2013/287019

Research Article

Signal Propagation in Protein Interaction Network during Colorectal Cancer Progression

Yang Jiang,[1] Tao Huang,[2] Lei Chen,[3] Yu-Fei Gao,[1] Yudong Cai,[4] and Kuo-Chen Chou[5,6]

[1] *Department of Surgery, China-Japan Union Hospital of Jilin University, Changchun 130033, China*
[2] *Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA*
[3] *College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*
[4] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[5] *King Abdulaziz University, Jeddah, Saudi Arabia*
[6] *Gordon Life Science Institute, 53 South Cottage Road, Belmont, MA 02478, USA*

Correspondence should be addressed to Yu-Fei Gao; gaoyufei1975@yahoo.com.cn and Yudong Cai; cai_yud@yahoo.com.cn

Received 16 January 2013; Accepted 18 February 2013

Academic Editor: Bin Niu

Colorectal cancer is generally categorized into the following four stages according to its development or serious degree: Dukes A, B, C, and D. Since different stage of colorectal cancer actually corresponds to different activated region of the network, the transition of different network states may reflect its pathological changes. In view of this, we compared the gene expressions among the colorectal cancer patients in the aforementioned four stages and obtained the early and late stage biomarkers, respectively. Subsequently, the two kinds of biomarkers were both mapped onto the protein interaction network. If an early biomarker and a late biomarker were close in the network and also if their expression levels were correlated in the Dukes B and C patients, then a signal propagation path from the early stage biomarker to the late one was identified. Many transition genes in the signal propagation paths were involved with the signal transduction, cell communication, and cellular process regulation. Some transition hubs were known as colorectal cancer genes. The findings reported here may provide useful insights for revealing the mechanism of colorectal cancer progression at the cellular systems biology level.

1. Background

Cancer is a complex system disease [1]. The complexity reflects in many ways. First, it is a network disease that involves the changes of many genes and these genes are connected in a certain way. Second, the disease network is evolving all the time during the progression. Some efforts have been made to understand such dynamic network [2–6].

As the third most common cancer worldwide [7], colorectal cancer develops via a progressive accumulation of genetic mutations and pathway dysfunctions [6]. It has the following four stages from early to late [8]: Dukes A, B, C, and D. In the stage of Dukes A, the cancer is only limited to the innermost layer. In Dukes B stage, the cancer has grown through the muscle layer. In Dukes C stage, the cancer has spread to the lymph nodes nearby. In Dukes D stage, the cancer is widely spread. The stage of Dukes D is the most advanced stage of colorectal cancer. Understanding the underlying molecular mechanisms of the pathological changes in colorectal cancer progression will facilitate the development of therapeutic treatments.

In the study of prion disease, it was found that during different stages of the disease, different regions of the network were activated and they formed a clear disease aggravation pattern on the network [2]. However, it is still not clear how one activated region is connected with another and how they can transit into one another.

To investigate the transition processes of different network states, we analyzed the gene expression profiles of 290 colorectal cancer patients, who were at different stages of Dukes A, B, C, and D. Using the Maximum Relevance and Minimum Redundancy (mRMR) [9] and Incremental Feature Selection (IFS) methods [10, 11] to compare the gene expressions among the patients of Dukes A, B, C, and D

stages, we obtained 158 early stage biomarkers and 284 late stage biomarkers, respectively. Subsequently, the early stage biomarkers and the late stage biomarkers were mapped onto the protein interaction network. If the early stage biomarker and the late stage biomarker were close to each other in the network, and also their expression levels were correlated with the patients of the Dukes B and C stages, then we assume that a signal propagation path may exist from the early stage biomarker to the late stage biomarker. Thus, by screening all the possible signal propagation paths from the early stage biomarkers to the late stage biomarkers, we have identified 632 signal propagation paths that contained 473 transition genes.

According to the Gene Ontology (GO) [12] enrichment analysis, many of the transition genes that transmitted the disease signal from the early stage biomarkers to the late stage biomarkers were involved into the signal transduction, cell communication, and cellular process regulation. Some transition hub genes were known colorectal cancer genes. They helped the transduction of the disease signal and the aggravation of colorectal cancer.

One signal propagation path from early stage biomarker MAVS to late stage biomarker GFPT1 was shown as an example. MAVS is an important immune protein and signaling protein in mitochondria [13–15] and GFPT1 is a rate-limiting enzyme of metabolism [16, 17]. It was suggested through our signal propagation analysis that MAVS responded to colorectal cancer in the early stage and then transmitted the disease signal to GFPT1 whose dysfunction further accelerated the colorectal cancer patients into late stage. This kind of in-depth analysis on the signal propagation path may provide useful insights into, or enrich, the understanding of the mechanism of colorectal cancer at the cellular or system biology level.

2. Methods

2.1. Benchmark Dataset. We downloaded the expression profiles of 19,621 genes in 290 colorectal cancer patients [18] from Gene Expression Omnibus (GEO) under accession number GSE14333. Of the 290 colorectal cancer patients, 44 were Dukes stage A, 94 Dukes stage B, 91 Dukes stage C, and 61 Dukes stage D. From Dukes A stage to Dukes D, the colorectal cancer gets more and more severe.

The protein interaction network we used was STRING v9.0 (http://string-db.org/) [19]. Each protein interaction in STRING has a confidence score, varying from 0.150 to 1. The confidence score is calculated by integrating the functional associations from genomic context, experiments, conserved coexpression, and previous knowledge with Bayesian method [19]. Suppose the interaction confidence score is denoted by $\mathbb{I}_{\text{score}}$, it follows according to the original definition

$$\mathbb{I}_{\text{rank}} = \begin{cases} \text{low confidence}, & \text{if } I_{\text{score}} > 0.150 \\ \text{mediam confidence}, & \text{if } I_{\text{score}} > 0.400 \\ \text{high confidence}, & \text{if } I_{\text{score}} > 0.700 \\ \text{highest confidence}, & \text{if } I_{\text{score}} > 0.900, \end{cases} \quad (1)$$

where $\mathbb{I}_{\text{rank}}$ represents the rank of protein interaction.

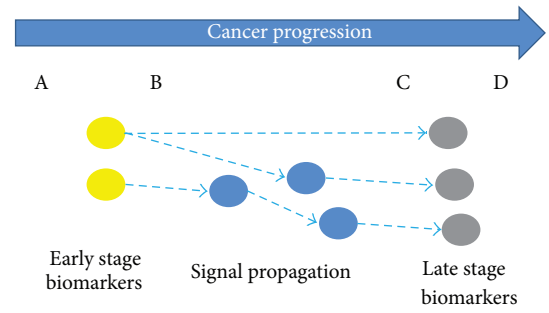


Figure 1: The diagram of signal propagation analysis during cancer progression. The blue arrow represents cancer progression. The colorectal cancer has four stages: Dukes A, B, C, and D. From A to D, the cancer gets worse and worse. Yellow nodes and grey nodes represent the biomarkers in the early and late stage, respectively. The goal of signal propagation analysis is to understand the transition from early stage biomarkers to late stage biomarkers by analyzing the signal propagation in the protein interaction network.

2.2. The Diagram of Signal Propagation Analysis during Cancer Progression. In studying or analyzing complex biological systems, it is quite helpful to introduce graphs or diagrams since they can provide an overall view or intuitive insights for the systems investigated, as demonstrated by a series of studies on various important biological topics (see, e.g., [20–29]). In this study, we first constructed a graph $\mathbb{G}$ with the PPI data from STRING. In the graph, an edge was assigned for each pair of proteins if they were in interaction with each other. There were 1375295 interaction edges among 15240 proteins. The "intimate degree" between two interacting proteins was defined by

$$\mathbb{I}_{\text{intimate}} = 1000 \times \left(1 - \mathbb{I}_{\text{score}}\right), \quad (2)$$

where $\mathbb{I}_{\text{score}}$ is the confidence score between two proteins concerned [19]. Thus, the higher the interaction confidence score between two proteins is, the closer their "interactive distance," and hence more intimate between them.

Shown in Figure 1 is an illustration for analyzing the signal propagation during the cancer progression. The colorectal cancer has four stages: Dukes A, B, C, and D. From Dukes A to Dukes D, the cancer gets worse and worse. The blue arrow represents the cancer progression. Below, we are to identify the biomarkers in the early stage (yellow nodes) and biomarkers in the late stage (grey nodes). Subsequently, we try to understand the transition from early stage biomarkers to late stage biomarkers by analyzing the signal propagation in the protein interaction network. This kind of analysis may provide useful insights for us to in-depth understand how the signal is propagated through the network.

2.3. Identification of Biomarkers in the Early and Late Stage. The following methods were used to identify the genes between different Dukes stages. First, the Maximum Relevance and Minimum Redundancy (mRMR) [9] method was applied to select the genes that has both maximum relevance with the cancer stages and minimum redundancy to each other. The mRMR program was downloaded from

http://penglab.janelia.org/proj/mRMR/. Second, the mRMR ranked genes were optimized with the Incremental Feature Selection (IFS) method [8, 30–35]. During the IFS operation, the accuracies of all possible top gene sets were calculated and the gene set that had the highest prediction accuracy was chosen as the optimal gene set, that is, the biomarkers. The accuracy was examined by the jackknife test, also known as Leave-One-Out Cross Validation (LOOCV) [36–39] and the prediction model was Nearest Neighbor Algorithm (NNA) [40]. The prediction accuracy was defined as the number of correctly predicted samples divided by the number of total samples.

The early stage biomarkers were selected from the Dukes A patients and Dukes B patients with mRMR and IFS methods. The late stage biomarkers were selected from the Dukes C patients and Dukes D patients.

2.4. The Transition from the Early Stage Biomarkers to the Late Stage Biomarkers. The early stage biomarkers and late stage biomarkers were mapped onto weighted protein interaction network graph $\mathbb{G}$. We identified the shortest paths between them using Dijkstra's algorithm [41–43]. The path length was the sum of edge weights through which the path passed. If the path length was smaller than $1000 \times (1 - 0.700) = 300$, it had high confidence to happen.

Meanwhile, we also tested the correlation between early stage biomarkers and late stage biomarkers in Dukes B patients and Dukes C patients. The Pearson correlation test P values were adjusted with false discovery rate (FDR) [44]. The cutoff of Pearson correlation test FDR was set to 0.001.

Included were those transitions that had the length shorter than 300 and the correlation test FDR smaller than 0.001. The shortest paths from the early stage biomarkers to the late stage biomarkers in the protein interaction network were deemed as the signal propagation paths for the transition.

2.5. Statistical Significance of Signal Propagation Path Identification. To evaluate the statistical significance of the identified signal propagation paths, we estimated the FDR of the signal propagation path based on the permutation [45]. We permuted the gene symbols in protein interaction network and gene expression profiles by 20,000 times. For each of the permutations, we calculated the length of the shortest path based on the weighted protein interaction network and the Pearson correlation test P value adjusted with the FDR method based on the gene expression profiles. The FDR of the signal propagation path was defined as

$$\mathrm{FDR}_{\text{signal-path}} = \frac{N_1}{N_2}, \tag{3}$$

where N_1 was the number of permutations in which the permuted shortest path length is shorter than the actual shortest path length and the permuted Pearson correlation test FDR is smaller than the actual Pearson correlation test FDR, while N_2 the total number of permutations which was 20,000 in this study.

2.6. The Transition Hubs in the Signal Propagation Paths. For each of the transition genes, we calculated the number of shortest paths that crossed it. Those genes that were crossed by more signal propagation paths were deemed more important transition hubs.

3. Results

3.1. Early and Late Stage Biomarkers. By selecting discriminative genes between the Dukes A patients and the Dukes B patients with mRMR and IFS methods, we identified the early stage biomarkers. Similarly, we obtained the late stage biomarkers from the Dukes C patients and the Dukes D patients. The IFS curves of early and late stage biomarker selection were shown in Figures 2(a) and 2(b), respectively. In Figure 2(a), the highest accuracy was 0.891 with 158 genes of the early stage biomarkers. In Figure 2(b), the highest accuracy was 0.855 with 284 genes of the early stage biomarkers. The 158 early stage biomarkers and 284 late stage biomarkers can be found in Supplemental Tables S1 and S2, available online at http://dx.doi.org/10.1155/2013/287019 respectively.

3.2. Comparison of Early and Late Stage Biomarkers. Now let us compare the early stage biomarkers with the late stage ones. It was observed between the two kinds of biomarkers there was only one gene, RNF4, in common. The expected number of overlap genes should be 2.29 and the odds ratio was 0.432. In other words, there was less overlap than expected. It was reported that in different stages of disease, different regions of the biological network are activated [2] and the dynamics of the biological network reflects the histopathology and clinical changes [6, 46]. The shifting from the activated region of early stage biomarkers to the activated region of late stage biomarkers in the biological network explains the under overlap between the early and late stage biomarkers, which may also help understand the colorectal cancer progression. In the following section, we are to study the transition processes in which the early stage biomarkers propagate the disease-aggravating signal to the late stage biomarkers, triggering the patients to develop into the most severe condition.

3.3. From Early Stage Biomarkers to Late Stage Biomarkers: The Transition. There were 136 early stage biomarkers and 230 late stage biomarkers that could be mapped onto the STRING network. The number of all possible combination pairs between the early and late stage biomarkers was $136 \times 230 = 31{,}280$, for each of which we calculated their shortest path length that was the sum of the edge weights in the shortest path. Furthermore, we calculated the Pearson correlation test FDR between them in Dukes B patients and Dukes C patients. Two criteria were applied to get the signal propagation path from early stage biomarkers to late stage biomarkers: the path length should be shorter than 300 and the correlation test FDR should be smaller than 0.001. There were 632 such signal propagation paths, as given in Table S3. Such 632 signal propagation paths linked 76 early stage biomarkers and 109 late stage biomarkers. Shown in Figure 3

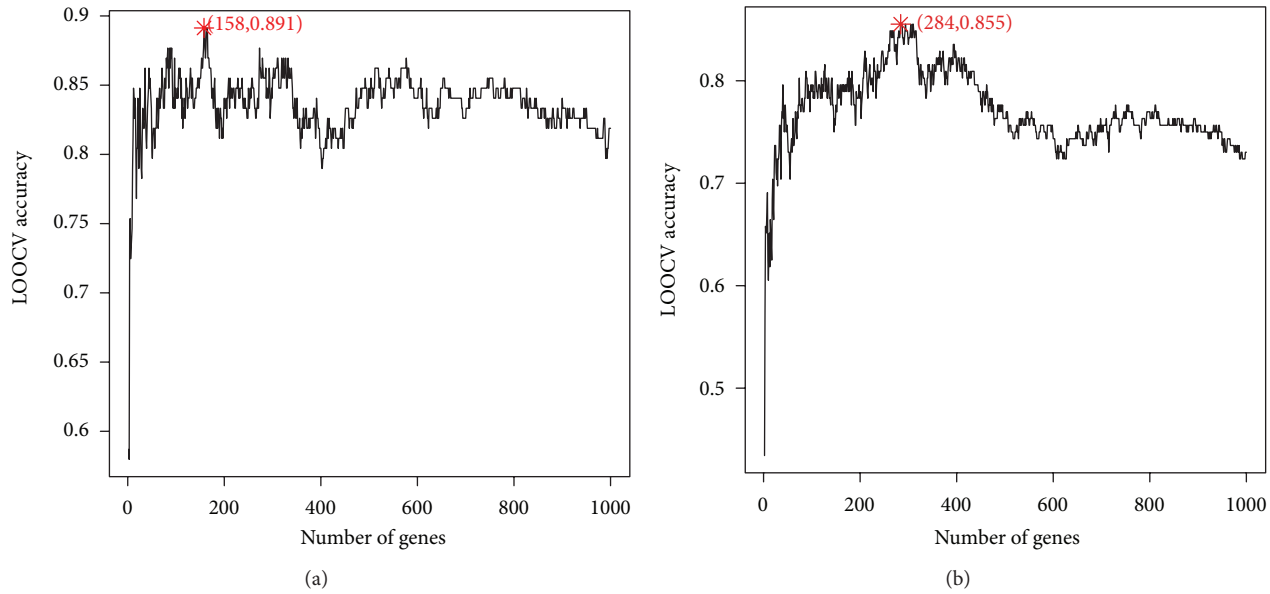


Figure 2: The IFS curves of early stage biomarkers and late stage biomarker. (a) The IFS curves of early stage biomarker selection. The highest accuracy was 0.891 with 158 genes which were the early stage biomarkers. (b) The IFS curves of late stage biomarker selection. The highest accuracy was 0.855 with 284 genes which were the late stage biomarkers.

are the transition networks from early stage biomarkers to late stage biomarkers.

Meanwhile, the values of FDR for the identified signal propagation paths were also calculated by first permuting the gene symbols in the protein interaction network and gene expression profiles and then comparing the permuted shortest path length and Pearson correlation FDR with the actual ones. Based on the results of the 20,000 permutations, the statistical significance of each identified signal propagation path was evaluated. It was found that all the 632 identified signal propagation paths were with FDR less than 0.05 and 81.3% of them had FDR less than 0.01.

3.4. The Transition Hubs in Signal Propagation. The 632 signal propagation paths crossed 473 genes. We ranked each of the 473 transition genes based on the number of signal propagation paths that had crossed it. The genes crossed by more signal propagation paths were regarded as more important transition hubs. The detailed results of the 473 transition genes as well as the numbers of signal propagation paths that had crossed them can be found in Table S4. The top three transition hubs were TP53 (tumor protein 53), CTNNB1 (cadherin-associated protein, beta 1), and EP300 (E1A binding protein p300). Interestingly, two of them, TP53 and EP300, were colorectal cancer genes, fully consistent with the reports in the Online Mendelian Inheritance in Man [47] (OMIM, http://omim.org/entry/114500).

4. Discussion

4.1. The Biological Functions of Early Stage Biomarkers, Late Stage Biomarkers, and Transition Genes. We used GATHER [48] (http://gather.genome.duke.edu/) to investigate the biological functions of the 158 early stage biomarkers, the 284 late stage biomarkers, and the 473 transition genes. The Gene Ontology (GO) enrichment results thus obtained are shown in Tables 1, 2, and 3, respectively. Since the 473 transition genes were enriched into too many GO terms, only the five enriched GO terms with the highest Bayes factor [49] were shown in Table 3. It is instructive to point out that the late stage biomarkers had more enriched GO terms than the early stage biomarkers. Also, the late stage biomarkers were more enriched in the common GO terms than the early stage biomarkers, such as "GO:0009607: response to biotic stimulus," "GO:0006952: defense response," and "GO:0006955: immune response." The roles of defense response and immune response in colorectal cancer [50, 51] have been widely studied. Many of the transition genes were involved in the signal transduction, cell communication, and cellular process regulation. These kinds of functions played important roles in transducing the disease signal and aggravating the colorectal cancer.

4.2. The Overlapped Gene between Early Stage Biomarkers and Late Stage Biomarkers. One overlapped gene, RNF4 (RING finger protein 4), was observed between the early stage biomarkers and the late stage biomarkers. As reported in [52], RNF4 was a patented biomarker gene of colorectal cancer. Also, as reported in [53], downregulation of RNF4 was related to the colorectal cancer risk (http://www.wipo.int/patentscope/search/en/WO2010033371).

Since RNF4 plays a unique role in ubiquitylation [54], DNA demethylation [35], and DNA repair [35], the colorectal cancer progression may involve the abnormal ubiquitylation and demethylation.

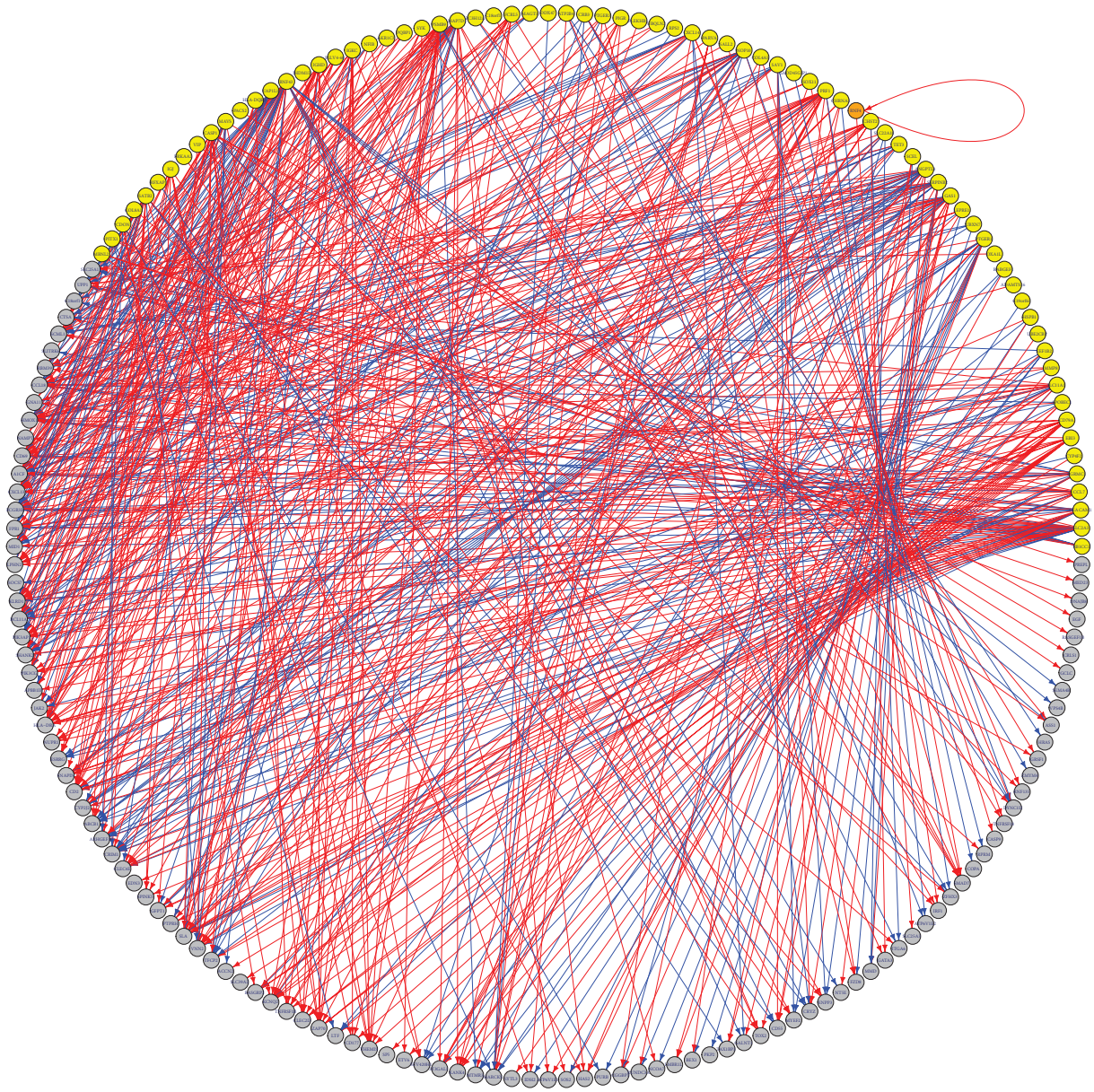


Figure 3: The transition network from early stage biomarkers to late stage biomarkers. The yellow and grey nodes were early and late stage biomarkers, respectively. The orange node, RNF4, was both early and late stage biomarker. The red and blue edges indicated that the early and late stage biomarkers were positively and negatively correlated.

4.3. The Signal Propagation Path from the Early Stage Biomarker MAVS to the Late Stage Biomarker GFPT1. It is interesting to see that GFPT1 was ranked no. 1 among the late stage biomarkers although it was even not a biomarker in the early stage. We traced back in the signal propagation paths and found GFPT1 was the downstream of the following seven early stage biomarkers: MAVS, TET3, GAS1, ANGPTL4, MAP7D1, CEACAM1, and PGRMC1. Among the 158 early stage biomarkers, MAVS was ranked no. 4, but MAVS was not a late stage biomarker. The Pearson correlation test P value and Pearson correlation coefficient between the expression levels of MAVS and GFPT1 in the Dukes B patients and the Dukes C patients were $1.09e-05$ and 0.317, respectively. Shown in Figure 4 is the signal propagation path from MAVS to GFPT1 in the STRING network: MAVS → IRF3 → CREBBP → TP53 → ATF3 → ATF4 → ASNS → GLUL → GFPT1.

Mitochondrial antiviral signaling (MAVS) protein is important in innate immunity [13–15]. The antibody able to induce immune responses can be used to treat cancer [55]. Immune responses usually occur early in the cancer progression stage but later the cancer cells may develop an ability to escape the immune-mediated lysis [56]. This might explain why MAVS was an early stage biomarker, but not a late stage biomarker.

GFPT1 is the key enzyme in hexosamine synthesis pathway whose products have been implicated in O-linked N-acetylglucosamine (O-GlcNAc) protein modification, insulin

Table 1: The enriched GO terms of the 158 early stage biomarkers with adjusted P value less than 0.01.

Gene ontology	Number of input genes with the annotation	Adjusted P value
GO:0009607: response to biotic stimulus	16	0.001
GO:0006952: defense response	14	0.004
GO:0006955: immune response	13	0.004

Table 2: The enriched GO terms of the 284 late stage biomarkers with adjusted P value less than 0.01.

Gene ontology	Number of input genes with the annotation	Adjusted P value
GO:0006952: defense response	25	0.0002
GO:0006955: immune response	23	0.0002
GO:0016064: immunoglobulin mediated immune response	8	0.0006
GO:0006959: humoral immune response	9	0.001
GO:0009607: response to biotic stimulus	25	0.002
GO:0019730: antimicrobial humoral response	6	0.005

Table 3: The most enriched five GO terms of the 473 transition genes.

Gene ontology	Number of input genes with the annotation	Adjusted P value	Bayes factor
GO:0008283: cell proliferation	107	<0.0001	47
GO:0007154: cell communication	219	<0.0001	43
GO:0007165: signal transduction	191	<0.0001	43
GO:0051244: regulation of cellular physiological process	71	<0.0001	40
GO:0050794: regulation of cellular process	84	<0.0001	38

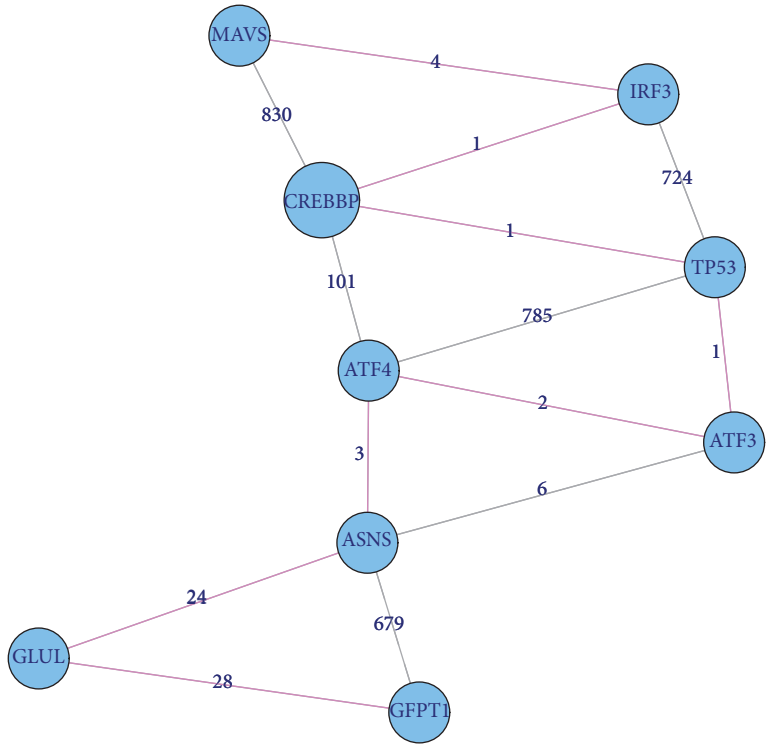


Figure 4: The signal propagation path from MAVS to GFPT1. The signal propagation path from MAVS to GFPT1 was MAVS → IRF3 → CREBBP → TP53 → ATF3 → ATF4 → ASNS → GLUL → GFPT1. The genes in the signal propagation path were mapped onto STRING network. The number on the edge was the edge weight. The edges on the signal propagation path were highlighted with pink color.

resistance, and glucose toxicity [16, 17]. It is a molecular therapeutic target for type-2 diabetes [57, 58]. As a metabolic disease, cancer is always accompanied with impaired mitochondrial function and dysfunctional energy metabolism [59].

Accordingly, it is rational to deduce the signal propagation from MAVS to GFPT1 as follows: in mitochondria, as an important innate immunity protein, MAVS may response to colorectal cancer in a very early stage. Then as a signaling protein, it transmits its signal to GFPT1 that has close relationship with mitochondria. The perturbation of GFPT1 may cause the dysfunction of mitochondria in the energy metabolism. The fates of the cells may be doomed by the collapse of their energy systems.

5. Conclusions

Our results indicated that the strong signals of early stage biomarkers would not necessarily disappear during the colorectal cancer progression, but might be transferred to other late stage biomarkers. This finding may provide useful insights for in-depth analyzing the signal propagation paths and helping to reveal the cellular mechanism of colorectal cancer aggravation.

Authors' Contribution

Y. Jiang and T. Huang contributed equally to this work.

Acknowledgments

This work was supported by Grants from National Basic Research Program of China (2011CB510102, 2011CB510101), Innovation Program of Shanghai Municipal Education Commission (no. 12YZ120, no. 12ZZ087), Natural Science Fund Projects of Jilin province (201215059), Development of Science and Technology Plan Projects of Jilin province (20100733, 201101074), and SRF for ROCS, SEM (2009-36), Scientific Research Foundation (Jilin Department of Science & Technology, 200705314, 20090175, 20100733), Scientific Research Foundation (Jilin Department of Health, 2010Z068), SRF for ROCS (Jilin Department of Human Resource & Social Security, 2012-2014).

References

[1] I. G. Khalil and C. Hill, "Systems biology for cancer," *Current Opinion in Oncology*, vol. 17, no. 1, pp. 44–48, 2005.

[2] D. Hwang, I. Y. Lee, H. Yoo et al., "A systems approach to prion disease," *Molecular Systems Biology*, vol. 5, p. 252, 2009.

[3] T. Huang, Y.-D. Cai, L. Chen et al., "Selection of reprogramming factors of induced pluripotent stem cells based on the protein interaction network and functional profiles," *Protein and Peptide Letters*, vol. 19, no. 1, pp. 113–119, 2012.

[4] T. Huang, J. Zhang, L. Xie et al., "Crosstissue coexpression network of aging," *OMICS A Journal of Integrative Biology*, vol. 15, no. 10, pp. 665–671, 2011.

[5] T. Huang, L. Liu, Q. Liu et al., "The role of Hepatitis C Virus in the dynamic protein interaction networks of Hepatocellular cirrhosis and Carcinoma," *International Journal of Computational Biology and Drug Design*, vol. 4, no. 1, pp. 5–18, 2011.

[6] E. R. Fearon and B. Vogelstein, "A genetic model for colorectal tumorigenesis," *Cell*, vol. 61, no. 5, pp. 759–767, 1990.

[7] N. Ismaili, "Treatment of colorectal liver metastases," *World Journal of Surgical Oncology*, vol. 9, article 154, 2011.

[8] Drug-and-Therapeutics-Bulletin, "Population screening for colorectal cancer," *Drug and Therapeutics Bulletin*, vol. 44, no. 9, pp. 65–68, 2006.

[9] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[10] T. Huang, J. Zhang, Z.-P. Xu et al., "Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches," *Biochimie*, vol. 94, no. 4, pp. 1017–1025, 2012.

[11] T. Huang, Z. S. He, W. R. Cui et al., "A sequence-based approach for predicting protein disordered regions," *Protein and Peptide Letters*, vol. 20, no. 3, pp. 243–248, 2013.

[12] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[13] M. Papatriantafyllou, "Innate immunity: MAVS build-ups for defence," *Nature Reviews Immunology*, vol. 11, no. 9, pp. 570–571, 2011.

[14] D. Arnoult, F. Soares, I. Tattoli, and S. E. Girardin, "Mitochondria in innate immunity," *EMBO Reports*, vol. 12, no. 9, pp. 901–910, 2011.

[15] F. Hou, L. Sun, H. Zheng, B. Skaug, Q.-X. Jiang, and Z. J. Chen, "MAVS forms functional prion-like aggregates to activate and propagate antiviral innate immune response," *Cell*, vol. 146, no. 3, pp. 448–461, 2011.

[16] K. Liu, G. Wang, S. H. Zhao et al., "Molecular characterization, chromosomal location, alternative splicing and polymorphism of porcine GFAT1 gene," *Molecular Biology Reports*, vol. 37, no. 6, pp. 2711–2717, 2010.

[17] T.-J. Hsieh, T. Lin, P.-C. Hsieh, M.-C. Liao, and S.-J. Shin, "Suppression of Glutamine:Fructose-6-phosphate amidotransferase-1 inhibits adipogenesis in 3T3-L1 adipocytes," *Journal of Cellular Physiology*, vol. 227, no. 1, pp. 108–115, 2012.

[18] R. N. Jorissen, P. Gibbs, M. Christie et al., "Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer," *Clinical Cancer Research*, vol. 15, no. 24, pp. 7642–7651, 2009.

[19] L. J. Jensen, M. Kuhn, M. Stark et al., "STRING 8—a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Research*, vol. 37, no. 1, pp. D412–D416, 2009.

[20] G. P. Zhou and M. H. Deng, "An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways," *Biochemical Journal*, vol. 222, no. 1, pp. 169–176, 1984.

[21] K. C. Chou, "Graphic rules in steady and non-steady state enzyme kinetics," *Journal of Biological Chemistry*, vol. 264, no. 20, pp. 12074–12079, 1989.

[22] K. C. Chou, "Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady-state systems," *Biophysical Chemistry*, vol. 35, no. 1, pp. 1–24, 1990.

[23] I. W. Althaus, A. J. Gonzales, J. J. Chou et al., "The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase," *Journal of Biological Chemistry*, vol. 268, no. 20, pp. 14875–14880, 1993.

[24] K. C. Chou, F. J. Kezdy, and F. Reusser, "Kinetics of processive nucleic acid polymerases and nucleases," *Analytical Biochemistry*, vol. 221, no. 2, pp. 217–230, 1994.

[25] J. Andraos, "Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws—new methods based on directed graphs," *Canadian Journal of Chemistry*, vol. 86, no. 4, pp. 342–357, 2008.

[26] K. C. Chou, "Graphic rule for drug metabolism systems," *Current Drug Metabolism*, vol. 11, no. 4, pp. 369–378, 2010.

[27] G. P. Zhou, "The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism," *Journal of Theoretical Biology*, vol. 284, no. 1, pp. 142–148, 2011.

[28] K. C. Chou, W. Z. Lin, and X. Xiao, "Wenxiang: a web-server for drawing wenxiang diagrams," *Natural Science*, vol. 3, pp. 862–865, 2011.

[29] G. P. Zhou, "The structural determinations of the leucine zipper coiled-coil domains of the cGMP-dependent protein kinase Iα and its interaction with the myosin binding subunit of the myosin light chains phosphase," *Protein and Peptide Letters*, vol. 18, no. 10, pp. 966–978, 2011.

[30] B. Q. Li, T. Huang, L. Liu, Y. D. Cai, and K. C. Chou, "Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network," *PLoS ONE*, vol. 7, Article ID e33393, 2012.

[31] H. Huang, J. Wang, Y. D. Cai, H. Yu, and K. C. Chou, "Hepatitis C virus network based classification of Hepatocellular cirrhosis and carcinoma," *PLoS ONE*, vol. 7, Article ID e34460, 2012.

[32] T. Huang, Z. Xu, L. Chen, Y. D. Cai, and X. Kong, "Computational analysis of HIV-1 resistance based on gene expression profiles and the virus-host interaction network," *PLoS ONE*, vol. 6, no. 3, Article ID e17291, 2011.

[33] T. Huang, S. Wan, Z. Xu et al., "Analysis and prediction of translation rate based on sequence and functional features of the mRNA," *PLoS ONE*, vol. 6, no. 1, Article ID e16036, 2011.

[34] T. Huang, W. Cui, L. Hu, K. Feng, Y. X. Li, and Y. D. Cai, "Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles," *PLoS ONE*, vol. 4, no. 12, Article ID e8126, 2009.

[35] X. V. Hu, T. M. A. Rodrigues, H. Tao et al., "Identification of RING finger protein 4 (RNF4) as a modulator of DNA demethylation through a functional genomics screen," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 34, pp. 15087–15092, 2010.

[36] T. Huang, C. Wang, G. Zhang, L. Xie, and Y. Li, "SySAP: a system-level predictor of deleterious single amino acid polymorphisms," *Protein and Cell*, vol. 3, no. 1, pp. 38–43, 2012.

[37] T. Huang, S. Niu, Z. Xu et al., "Predicting transcriptional activity of multiple site P53 mutants based on hybrid properties," *PLoS ONE*, vol. 6, no. 8, Article ID e22940, 2011.

[38] K. C. Chou, "Prediction of protein cellular attributes using pseudo amino acid composition," *Proteins*, vol. 43, pp. 246–255, 2001, Erratum: vol. 44, p. 60, 2001.

[39] K. C. Chou and C. T. Zhang, "Review: prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, pp. 275–349, 1995.

[40] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review)," *Journal of Theoretical Biology*, vol. 273, pp. 236–247, 2011.

[41] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[42] G. Chartrand and O. R. Oellermann, *Applied and Algorithmic Graph Theory*, Mcgraw-Hill College, 1992.

[43] T. H. Cormen, C. E Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press and Mcgraw-Hill, 2nd edition, 2001.

[44] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, vol. 57, pp. 289–300, 1995.

[45] Y. Xie, W. Pan, and A. B. Khodursky, "A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data," *Bioinformatics*, vol. 21, no. 23, pp. 4280–4288, 2005.

[46] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, and Y. Li, "Prediction of lysine ubiquitination with mRMR feature selection and analysis," *Amino Acids*, vol. 42, no. 4, pp. 1387–1395, 2011.

[47] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 33, pp. D514–D517, 2005.

[48] J. T. Chang and J. R. Nevins, "GATHER: a systems approach to interpreting genomic signatures," *Bioinformatics*, vol. 22, no. 23, pp. 2926–2933, 2006.

[49] S. N. Goodman, "Toward evidence-based medical statistics. 2: the Bayes factor," *Annals of Internal Medicine*, vol. 130, no. 12, pp. 1005–1013, 1999.

[50] M. Czéh, C. Loddenkemper, S. Shalapour et al., "The immune response to sporadic colorectal cancer in a novel mouse model," *Oncogene*, vol. 29, no. 50, pp. 6591–6602, 2010.

[51] C. S. D. Roxburgh and D. C. McMillan, "The role of the in situ local inflammatory response in predicting recurrence and survival in patients with primary operable colorectal cancer," *Cancer Treatment Reviews*, vol. 38, no. 5, pp. 451–466, 2012.

[52] Q. Chen, Z. Ye, S. C. Lin, and B. Lin, "Recent patents and advances in genomic biomarker discovery for colorectal cancers," *Recent Patents on DNA and Gene Sequences*, vol. 4, no. 2, pp. 86–93, 2010.

[53] H.-J. Terng, W.-J. Lee, and C.-Y. Chen, "Molecular markers for lung and colorectal carcinomas," WO2010033371, 2010.

[54] A. Plechanovová, E. G. Jaffray, S. A. McMahon et al., "Mechanism of ubiquitylation by dimeric RING ligase RNF4," *Nature Structural and Molecular Biology*, vol. 18, no. 9, pp. 1052–1059, 2011.

[55] J. Hess, P. Ruf, and H. Lindhofer, "Cancer therapy with trifunctional antibodies: linking innate and adaptive immunity," *Future Oncology*, vol. 8, no. 1, pp. 73–85, 2012.

[56] I. B. . Barsoum, T. K. Hamilton, X. Li et al., "Hypoxia induces escape from innate immunity in cancer cells via increased expression of ADAM10: role of nitric oxide," *Cancer Research*, vol. 71, no. 24, pp. 7433–7441, 2011.

[57] K. C. Chou, "Molecular therapeutic target for type-2 diabetes," *Journal of Proteome Research*, vol. 3, no. 6, pp. 1284–1288, 2004.

[58] K. C. Chou, "Structural bioinformatics and its impact to biomedical science," *Current Medicinal Chemistry*, vol. 11, no. 16, pp. 2105–2134, 2004.

[59] T. N. Seyfried and L. M. Shelton, "Cancer as a metabolic disease," *Nutrition and Metabolism*, vol. 7, article 7, 2010.