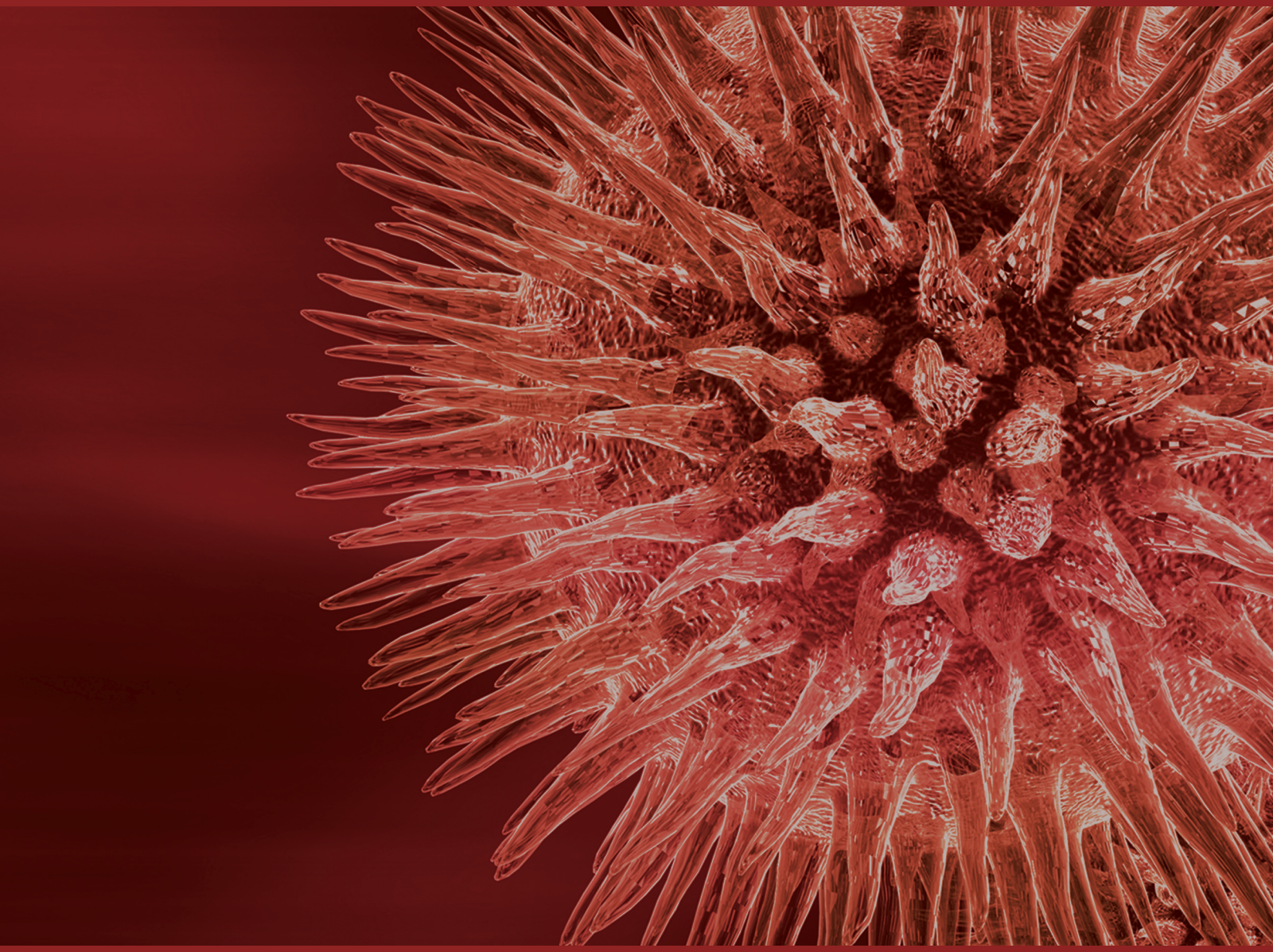
BioMed Research International

Computational Systems Biology Methods in Molecular Biology, Chemistry Biology, Molecular Biomedicine, and Biopharmacy

Guest Editors: Yudong Cai, Julio Vera González, Zengrong Liu, and Tao Huang



Computational Systems Biology Methods in Molecular Biology, Chemistry Biology, Molecular Biomedicine, and Biopharmacy

BioMed Research International

Computational Systems Biology Methods in Molecular Biology, Chemistry Biology, Molecular Biomedicine, and Biopharmacy

Guest Editors: Yudong Cai, Julio Vera González, Zengrong Liu, and Tao Huang

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 746814, 2 pages
http://dx.doi.org/10.1155/2014/746814

Editorial

Computational Systems Biology Methods in Molecular Biology, Chemistry Biology, Molecular Biomedicine, and Biopharmacy

Yudong Cai,[1] Julio Vera González,[2] Zengrong Liu,[3] and Tao Huang[4]

[1] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[2] *Department of Systems Biology, University of Rostock, Rostock 18051, Germany*
[3] *Department of Mathematics, College of Science, Shanghai University, Shanghai 200444, China*
[4] *Department of Genetics and Genomics Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA*

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

Received 25 March 2014; Accepted 25 March 2014; Published 9 April 2014

In the postgenomic era, the large-scale data, such as genome sequences, mRNA sequences, and protein sequences, increase rapidly. It is desired to develop the computational approaches that can derive and analyze useful information from them to promote the development of biomedicine and drug design. Meanwhile, in order to understand how protein-protein interactions and other complex interactions in a living system get integrated in complex nonlinear networks and regulate cell function, a new discipline, called "Systems Biology", is created.

In this special issue, 11 interesting studies were included. Several novel computational methods for systems biology were proposed for the first time and some intriguing biological findings were reported in large scale experiments.

J. Cao and L. Xiong studied the protein sequence classification using the single hidden layer feed forward neural network (SLFN). Two algorithms, the basic extreme learning machine (ELM) and the optimal pruned ELM (OP-ELM), were adopted as the learning algorithms for the ensemble based SLFNs. Their methods outperformed back propagation (BP) neural network and support vector machine (SVM).

Y. F. Gao et al. proposed a novel prediction method based on drug and compound ontology information extracted from ChEBI to identify drugs target groups, from which the kind of functions of a drug may be deduced. Their overall prediction accuracy on the training dataset was 83.12%, while it was 87.50% on the test dataset. The study may become an inspiration to solve the problems of this sort and bridge the gap between ChEBI ontology and drugs target groups.

Z. Li et al. developed a computational method to predict retinoblastoma (RB) related genes. RB is the most common primary intraocular malignancy usually occurring in childhood. Their method was based on dagging, with the maximum relevance minimum redundancy (mRMR) method followed by incremental feature selection (IFS). The RB and non-RB genes can be classified with Gene Ontology enrichment scores and KEGG enrichment scores. This method can be generalized to predict the other cancer related genes as well.

Y. Jiang et al. proposed a method to identify gastric cancer genes by first applying the shortest path algorithm to protein-protein interaction network and then filtering the shortest path genes based permutation betweenness. Many identified candidate genes were involved in gastric cancer related biological processes. Their study gives a new insight for studying gastric cancer.

T. Zhang et al. proposed a computational method for gene phenotypes prediction. Their method regarded the multiphenotype as a whole network which can rank the possible phenotypes associated with the query protein and showed more comprehensive view of the protein's biological effects. The performance of their method was better than dagging, random forest, and sequential minimal optimization (SMO).

Q. Zou et al. reviewed the network based disease gene identification methods, such as CIPHER, RWRH, Prince, Meta-path, Katz, Catapult, Diffusion Kernel, and ProDiGe and compared their performance. Some advices about software choosing and parameter setting were provided. They

also analyzed the core problems and challenges of these methods and discussed future research direction.

G. S.V. McDowell et al. developed a bioinformatics tool, Visualization and Phospholipid Identification (VaLID), to search and visualize the 1,473,168 phospholipids from the VaLID database. Each phospholipid can be generated in skeletal representation. VaLID is freely available and responds to all users through the CTPNL resources website at http://neurolipidomics.com/resources.html and http://neurolipidomics.ca/.

X. Lai et al. proposed a systems biology approach combining database-oriented network reconstruction, data-driven modeling, and model-driven experiments to study the regulatory role of miRNAs in coordinating gene expression. They illustrate the method by reconstructing, modeling and simulating the miRNA network regulating p21. Their model can be used to study the effect of different miRNA expression profiles and cooperative target regulation on p21 expression levels in different biological contexts and phenotypes.

P. Cui et al. analyzed the genome-wide relationship between chromatin features and chromatin accessibility in DNase I hypersensitive sites. They found that these features show distinct preference to localize in open chromatin. Their study provides new insights into the true biological phenomena and the combinatorial effects of chromatin features on differential DNase I hypersensitivity.

L. Zhu et al. sequenced the transcriptome of *Sophora japonica* Linn (Chinese scholar tree), a shrub species belonging to the subfamily Faboideae of the pea family Fabaceae. Approximately 86.1 million high-quality reads were generated and assembled de novo into 143010 unique transcripts and 57614 unigenes. The transcriptome data of *S. japonica* from this study represents first genome-scale investigation of gene expressions in Faboideae plants.

Y. Wang et al. characterized the microsatellite pattern in the *V. volvacea* genome and compared it with microsatellites found in the genomes of four other edible fungi: *Coprinopsis cinerea, Schizophyllum commune, Agaricus bisporus,* and *Pleurotus ostreatus*. A total of 1346 microsatellites have been identified with mononucleotides, the most frequent motif. Their analysis suggested a possible relationship between the most frequent microsatellite types and the genetic distance between the five fungal genomes.

With the current exponential increase of biological and biomedical high-throughput data generated, in the future we will see how methodologies like the ones described in this special issue become absolutely necessary. But furthermore we need to know how methodologies pertaining data analysis, network reconstruction, and modeling get together (a) to make possible the integration of massive, multiple-type quantitative high-throughput data and (b) to understand how cell phenotypes emerge from large, multilevel and structurally complex biochemical regulatory networks.

Yudong Cai
Julio Vera González
Zengrong Liu
Tao Huang

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 103054, 12 pages
http://dx.doi.org/10.1155/2014/103054

Research Article

Protein Sequence Classification with Improved Extreme Learning Machine Algorithms

Jiuwen Cao[1] and Lianglin Xiong[2,3]

[1] *Institute of Information and Control, Hangzhou Dianzi University, Zhejiang 310018, China*
[2] *School of Mathematics and Computer Science, Yunnan University of Nationalities, Kunming 650500, China*
[3] *School of Mathematics and Statistics, Yunnan University, Kunming 650091, China*

Correspondence should be addressed to Jiuwen Cao; caoj0003@e.ntu.edu.sg

Received 17 December 2013; Revised 15 February 2014; Accepted 16 February 2014; Published 30 March 2014

Academic Editor: Tao Huang

Precisely classifying a protein sequence from a large biological protein sequences database plays an important role for developing competitive pharmacological products. Comparing the unseen sequence with all the identified protein sequences and returning the category index with the highest similarity scored protein, conventional methods are usually time-consuming. Therefore, it is urgent and necessary to build an efficient protein sequence classification system. In this paper, we study the performance of protein sequence classification using SLFNs. The recent efficient extreme learning machine (ELM) and its invariants are utilized as the training algorithms. The optimal pruned ELM is first employed for protein sequence classification in this paper. To further enhance the performance, the ensemble based SLFNs structure is constructed where multiple SLFNs with the same number of hidden nodes and the same activation function are used as ensembles. For each ensemble, the same training algorithm is adopted. The final category index is derived using the majority voting method. Two approaches, namely, the basic ELM and the OP-ELM, are adopted for the ensemble based SLFNs. The performance is analyzed and compared with several existing methods using datasets obtained from the Protein Information Resource center. The experimental results show the priority of the proposed algorithms.

1. Introduction

Protein sequences (also known as polypeptides) are organic compounds made of amino acids arranged in a linear chain and folded into a globular form. The amino acids in a polymer chain are joined together by the peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The sequence of amino acids in a protein is defined by the sequence of a gene, which is encoded in the genetic code. As shown in Figure 1, a gene is any given segment along the deoxyribonucleic acid (DNA) that encodes instructions which allow a cell to produce a specific product. Typically, a protein such as an enzyme initiates one specific action.

Due to the wide applications in clinical proteomics and protein bioinformatics, protein sequence analyses have been comprehensively studied in recent years, such as the work presented by Barve et al. [1], Chen et al. [2], Cong et al. [3], Machado et al. [4], Carregari et al. [5], and Liu et al. [6]. Protein sequence analysis generally helps to characterize protein sequences in silico and allows the prediction of protein structures and functions. Recent research has shown that the comparative analysis of the protein sequences is more sensitive than directly comparing DNA. Hence, a number of protein sequence databases have been established in the past decades, such as Protein Information Resource (PIR) (http://pir.georgetown.edu/), Protein Data Bank (PDB) (http://www.pdb.org/pdb/home/home.do), and Universal Protein Resource (UniProt) (http://www.uniprot.org/). Hence, it becomes an important and challenging task to efficiently exploit useful information from the large protein sequence dataset for both computer scientists and biologists. As mentioned by Baldi and Brunak [7], protein sequence classification plays an important role in protein sequence analysis on the account of those protein sequence members consisting of a same protein superfamily are evolutionally related and functionally and structurally relevant to each other. Precisely classifying a member protein sequence

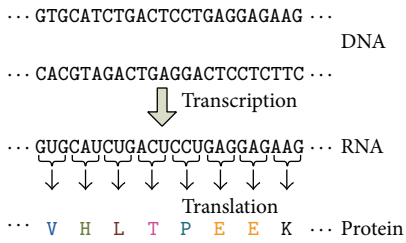


Figure 1: The DNA sequence of a gene encodes the amino acid sequence of a protein.

into a superfamily protein would show the benefit that it only needs to carry out some molecular analysis within a particular superfamily instead of the analysis on all the individual member protein sequences. Generally, two protein sequences are classified into the same category if their feature patterns extracted by sequence alignment algorithms show high homology. Lots of alignment algorithms have been proposed in the past few years to identify the class of the unseen protein sequence based on comparing it with some known protein sequences and calculating their similarities, such as iPro-Class (http://pir.georgetown.edu/), SAM (SAM: Sequence Alignment and Modeling Software System, Baskin Center for Computer Engineering and Science, http://www.cse.ucsc.edu/researchcompbio/), and MEME (MEME: Multiple Expectation Maximization for Motif Elicitation UCSD Computer Science and Engineering, http://meme.sdsc.edu). However, it is a very time-costing work to compare the testing protein sequence with all the existing identified protein sequences, especially when the database is large and the length of the unseen protein sequence is long. Therefore, establishing an efficient and intelligent classification system to exactly label the testing protein sequence in a large database becomes urgent and useful.

A number of methods have been developed for general signal classifications based on the statistical theory in the past decades, such as decision trees, statistical techniques, support vector machine (SVM), and neural networks (NN). Yang et al. [8] employed the word segmentation method for feature extraction on the protein sequence and then utilized the SVM for classification. Beside this, Caragea et al. [9] used the hashing function to reduce the dimension of the protein sequence feature vector and then performed classification with SVM. Alternative to using the SVM method, neural networks are another popular method for protein sequences classification in terms of the following two reasons: (i) as the features of protein sequences are generally distributed in a high dimensional space with complex characteristics, it is usually difficult to find a satisfactory model using the statistical or parameterized approaches, and (ii) neural networks are able to process the raw continuous values fed into the model. A lot of research works based on neural networks for protein sequences classification have been done in the last few years, such as Wang et al. [10, 11] and Wang and Huang [12]. Wang et al. [10] proposed a modular radial basis function (RBF) neural network classifier for the protein sequences with improved classification criteria, and two heuristic rules were presented for the decision-making to enhance the classification reliability. A generalized radial basis function (GRBF) neural network architecture that generates a set of fuzzy classification rules was developed for protein sequences classification by Wang et al. [11]. Most of the previous papers in protein sequence classifications usually chose the gradient based algorithm for neural networks, which are time-consuming in general. Hence, a computationally efficient tuning-free algorithm named extreme learning machine (ELM) for single hidden layer feedforward neural networks (SLFNs), which was recently proposed by Huang et al. [13, 14] and further improved by Huang et al. [15], was applied for protein sequences classification by Wang and Huang [12]. The experimental results given by Wang and Huang [12] have shown that the ELM algorithm learns thousands times faster than the conventional gradient based method (also known as backpropagation (BP), which was developed by Levenberg [16] and Marquardt [17]) with a higher classification rate in protein sequences. To enhance the classification performance and keep the training time in an acceptable level, a self-adaptive evolutionary ELM (SaE-ELM) which utilized the self-adaptive differential evolutionary to update the hidden neuron parameters in the ELM neural network has been presented in Cao et al. [18].

Although the basic ELM and its invariant SaE-ELM have been employed and discussed for protein sequence classification by Wang and Huang [12] and Cao et al. [18], respectively, there are still a lot of rooms for improvements. As presented and discussed by Wang and Huang [12], although ELM learning is much faster than the conventional BP algorithm on the protein sequence dataset, the improvement of classification rate is relatively small. With this objective, we study classification performance of protein sequence based on recent improved ELM algorithms in this paper. The contributions of the paper are threefold. First, the recent robust and generic algorithm named the optimal pruned ELM (OP-ELM) developed by Miche et al. [19] is utilized for protein sequence classification in this paper, where the multiresponse sparse regression (MRSR) technique developed by Simila and Tikka [20] and the leave-one-out (LOO) validation criterion are employed in the OP-ELM for the selection of an appropriate number of hidden neurons. Second, the ensemble based SLFNs network structure is proposed and the majority voting method is adopted to further enhance the protein sequence classification performance. Third, both the basic ELM and the OP-ELM are used as the training algorithms for each ensemble in the new structure. Thus, two algorithms named the voting based ELM (V-ELM) and the voting based OP-ELM (VOP-ELM) are developed for protein sequence classifications. The performance of all the proposed methods is analyzed using the protein sequence database from the Protein Information Resource (PIR) center. Simulations results are compared with recent state-of-art algorithms, such as SVM by Hsu and Lin [21], BP by Haykin [22], Marquardt [17], Levenberg [16], and the original ELM by Huang et al. [13, 14, 23].

Organizations of the rest of the paper are as follows. The feature extraction method for biological protein sequence data used by Wang et al. [10, 11] is reviewed in Section 2. The data description of the protein sequences downloaded from the PIR database is also introduced in this section. In Section 3, the recent ELM and its improved method OP-ELM for protein sequence classification using single hidden layer feedforward neural network are first given. The ensemble based SLFNs structure combined with the majority voting method is then proposed for protein sequence classification. The original ELM and the OP-ELM are used as the learning algorithms for each ensemble. Experimental results and performance comparisons are given in Section 4. Discussions and conclusions are drawn in Section 5.

2. Feature Extraction

As described by Wang et al. [10, 11], a protein sequence is generally made from various combinations of 20 amino acids with notations as $\Sigma = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$. For protein sequence classifications, the n-gram features with a pair of values (v_m, c_m) are extracted as the input signals fed to a classifier, where v_m is the feature m and c_m is the count of this feature in a protein sequence with $m = 1, 2, \ldots, 20^n$. For instance, the 2-gram features from the set Σ are all the combinations as $(AA, AC, \ldots, AY, CA, CC, \ldots, CY, \ldots, YA, YC, \ldots, YY)$. A feature is the number of occurrences of an amino within a protein sequence. Taking a protein sequence *VAAGTVAGT* as an example, the 2-gram features can be extracted and presented as $\{(VA, 2), (AA, 1), (AG, 2), (GT, 2), (TV, 1)\}$. Another commonly used information for protein sequence feature extraction is the 6-letter exchange group. That is, the 6 combinations of the letters from the whole set Σ are formed as $\mathbf{A} = \{H, R, K\}$, $\mathbf{B} = \{D, E, N, Q\}$, $\mathbf{C} = \{C\}$, $\mathbf{D} = \{S, T, P, A, G\}$, $\mathbf{E} = \{M, I, L, V\}$, and $\mathbf{F} = \{F, Y, W\}$. Therefore, using the 6-letter group, the above protein sequence *VAAGTVAGT* can be transformed as **EDDDDEDDD** and its 2-gram features are $\{(\mathbf{DE}, 1), (\mathbf{ED}, 2), (\mathbf{DD}, 5)\}$. Similar to the feature extraction done by Wang et al. [10, 11] and Wang and Huang [12], we use $\mathbf{e}_n$ and $\mathbf{a}_n$ to represent n-gram features from the 6-letter group and the 20-letter set, respectively. Each set of n-gram features from a protein sequence, that is, $\mathbf{e}_n$ and $\mathbf{a}_n$, is scaled separately to avoid skew in the counts value using the formula: $\overline{x} = (x/(\mathbf{l}-n+1))$, where x is the count of the generic gram feature, $\overline{x}$ is the normalized x, which is the inputs of the classifiers, $\mathbf{l}$ is the length of the protein sequence, and n is the size of the n-gram features.

A 56-dimensional feature vector, extracted from a protein sequence and comprised from n-gram features of the 6-letter group represented as $\mathbf{e}_2$ and the 20 letters set represented as $\mathbf{a}_1$, is used as the input information of the classifiers. Two protein sequence datasets with ten-super families (classes) were obtained from the PIR databases and denoted as PIR1 and PIR2, respectively. There are 949 protein sequences samples in PIR1 and 534 protein sequences samples in PIR2. The details of the ten superfamilies to be classified are Cytochrome c (113/17), Cytochrome c6 (45/14), Cytochrome b (73/100), Cytochrome b5 (11/14), Triosephosphate isomerase (14/44), Plastocyanin (42/56), Photosystem II D2 protein (30/45), Ferredoxin (65/33), Globin (548/204), and Cytochrome b6-f complex 4.2K (8/6), where the first digit in the bracket denotes the number of the protein sequences in the PIR1 database and the second digit represents the number of the protein sequences in the PIR2 database, respectively.

3. Methodologies

3.1. SLFN for Protein Sequence Classification

3.1.1. Model Description of SLFN. For the supervised learning in SLFNs, a dataset with input signal features and their associated class category is generally available to train the network parameters. Assuming that the available dataset is $\mathscr{A} = \{(\mathbf{x}_i, t_i)\}_{i=1}^{\mathbf{N}}$, where $\mathbf{x}_i$, t_i, and $\mathbf{N}$ represent the feature vector of the ith protein sequence, its corresponding category index, and the number of protein sequences, respectively; a single hidden layer feedforward neural network (SLFN) with $\mathbf{J}$ nodes in the hidden layer can be expressed as

$$\mathbf{o}_i = \sum_{j=1}^{\mathbf{J}} \mathbf{w}_j g\left(\mathbf{a}_j, b_j, \mathbf{x}_i\right), \quad i = 1, 2, \ldots, \mathbf{N}, \tag{1}$$

where $\mathbf{o}_i$ is the output obtained by the SLFN associated with the ith input protein sequence, and $\mathbf{a}_j \in \mathscr{R}^d$ and $b_j \in \mathscr{R}$ $(j = 1, 2, \ldots, \mathbf{J})$ are parameters of the jth hidden node, respectively. The variable $\mathbf{w}_j \in \mathscr{R}^m$ is the link connecting the jth hidden node with the output layer and $g(\cdot)$ is the hidden node activation function. With all training samples, (1) can be expressed in the compact form as

$$\mathscr{O} = \mathscr{H}\mathscr{W}, \tag{2}$$

where $\mathscr{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{\mathbf{J}})$ and $\mathscr{O}$ are the output weight matrix and the network outputs, respectively. The variable $\mathscr{H}$ denotes the hidden layer output matrix with the entry $\mathscr{H}_{ij} = g(\mathbf{a}_j, b_j, \mathbf{x}_i)$. To perform multiclasses classification, SLFNs generally utilize the One-Against-All (OAA) method to transform the classification application to a multioutput model regression problem. That is, for a $\mathscr{C}$-categories classification application, the output label t_i of the protein sequence feature vector $\mathbf{x}_i$ is encoded to a $\mathscr{C}$-dimensional vector $\mathbf{t}_i = (t_{i1}, t_{i2}, \ldots, t_{i\mathscr{C}})^T$ with $t_{i\mathbf{c}} \in \{1, -1\}$ $(\mathbf{c} = 1, 2, \ldots, \mathscr{C})$. If the category index of the protein sequence $\mathbf{x}_i$ is $\mathbf{c}$, then $t_{i\mathbf{c}}$ is set to be 1 while the rest entries in $\mathbf{t}_i$ are set to be -1. Hence, the objective of training phase for the SLFN in (1) becomes finding the best network parameters set $S = \{(\mathbf{a}_j, b_j, \mathbf{w}_j)\}_{j=1,\ldots,\mathbf{J}}$ such that the following error cost function is minimized:

$$\min_S E = \min_S \|\mathscr{O} - \mathscr{T}\|, \tag{3}$$

where $\mathscr{T} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{\mathbf{N}})$ is the target output matrix. The most popular algorithm is the gradient descent based method where the network back-forward errors are used to iteratively update the network parameters. However, the slow learning

Given:
Training dataset $\mathscr{A}$; number of hidden nodes $\mathbf{J}$, activation function $g(\cdot)$
(1) Randomly generate $\mathbf{a}_j$ and b_j for $j = 1, \ldots, \mathbf{J}$;
(2) Calculate $\mathscr{H}$;
(3) Find $\mathscr{W}$ according to (4).

Algorithm 1: ELM [14, 23].

speed and poor learning scalability limit their applications in large datasets and high dimensional signals, such as the protein sequence classification.

3.1.2. ELM. Alternative to iteratively tuning the network parameters, extreme learning machine (ELM), which was recently developed by Huang et al. [13, 14, 23], claims that random hidden node parameters can be utilized for SLFNs and the hidden node parameters may not need to be tuned. It was stated in Huang et al. [14] that a standard SLFN with $\mathbf{N}$ hidden nodes using random input weights and biases and the infinitely differentiable activation function can exactly learn $\mathbf{N}$ arbitrary distinct samples. In such case, the system (2) becomes a linear model and the network parameter matrix can be analytically solved by using the least-square method. That is,

$$\mathscr{W} = \mathscr{H}^{\dagger}\mathscr{T}, \tag{4}$$

where $\mathscr{H}^{\dagger}$ is the Moore-Penrose generalized inverse of the hidden layer output matrix $\mathscr{H}$ given by Serre [24]. The universal approximation property of the ELM theory by using random input weights and biases is also presented in Huang et al. [14] to support the algorithm.

In the following, a brief summary of the ELM algorithm is given in Algorithm 1.

As illustrated through simulations given by Huang et al. [14, 23, 25], Liang et al. [26], and Zhang et al. [27], utilizing random hidden parameters in the training phase, ELM not only learns thousand times faster than the BP algorithm and its variants but also has a comparable generalization performance as the conventional gradient descent based methods and SVM. The protein sequence classification based on ELM has been first studied by Wang and Huang [12]. The experimental results obtained by Wang and Huang [12] have shown that ELM performs slight better than the BP algorithm where the improvement of the successful testing classification rate obtained by ELM is around 1%. But ELM runs thousand times faster than BP.

3.1.3. OP-ELM. However, using random hidden node parameters and the tuning-free learning framework in ELM may also bring some issues. For example, the parameters may not be the optimal one and redundant nodes may exist. To address these issues and enhance the performance, many invariants of ELM have been developed in the past several years, such as considering the distribution of input dataset done by Cao et al. [28] and utilizing the differential evolutionary method for parameter optimization by Zhu et al. [29] and Cao et al. [30]. One of the recent representative improvement methods named the optimal pruned ELM (OP-ELM) is developed by Miche et al. [19] to prune the related neurons with irrelevant variables. To get rid of unuseful hidden nodes of ELM, OP-ELM combines the multiresponse sparse regression (MRSR) by Simila and Tikka [20] with the leave-one-out (LOO) validation method to rank hidden neurons and to select the actual best number of neurons for the model.

The MRSR technique is described as follows. Suppose $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_{\mathbf{m}}] \in \mathscr{R}^{\mathbf{n}\times\mathbf{m}}$ is the regressor and MRSR adds each column of the regressor matrix one by one to the model $\mathbf{Y}^l = \mathbf{X}\boldsymbol{\beta}^l$, where $\mathbf{Y}^l = [\mathbf{y}_1^l \ldots \mathbf{y}_{\mathscr{C}}^l]$ is the target approximation of the model. At the lth step, the weight matrix $\boldsymbol{\beta}^l$ has l nonzero rows. A new nonzero row and a new column of the regressor matrix are added to the model when increasing the steps. With the MRSR method, the hidden neurons $\mathbf{h}_i$ which are the rows of the hidden layer output matrix $\mathscr{H}$ are ranked. For the details of MRSR, one can refer to the work presented by Simila and Tikka [20].

To select the best number of neurons, the LOO validation method is introduced in OP-ELM. Employing the PREdiction Sum of Squares (PRESS) statistics (presented by Bontempi et al. [31] and Myers [32]), the LOO validation error $\varepsilon^{\text{PRESS}}$ is calculated as

$$\varepsilon^{\text{PRESS}} = \frac{t_i - \mathbf{h}_i\boldsymbol{\beta}}{1 - \mathbf{h}_i\mathbf{P}\mathbf{h}_i^T}, \tag{5}$$

where $\mathbf{P} = (\mathscr{H}^T\mathscr{H})^{-1}$. Then, the appropriate number of neurons for the model is taken by evaluating the LOO error versus the used number of neurons. The OP-ELM can be summarized in three steps as found in Algorithm 2.

3.2. Ensemble Based SLFNs for Protein Sequence Classification. For both ELM and OP-ELM, we can find that the hidden node parameters are randomly assigned once and never updated. With a single SLFN training by ELM or OP-ELM on the protein sequence dataset, the misclassification number of samples may be high. To address this issue, we propose an ensemble based structure of SLFNs for protein sequence classification. Since the original ELM enjoys a much fast training speed for SLFN, it is feasible to employ multiple independent SLFNs to predict the category index with the majority voting method. Rather than relying on single realization of random parameters in ELM, employing the ensemble method would

(1) Utilize ELM to train the SLFN;
(2) Rank the hidden neurons of ELM by the MRSR method;
(3) Decide the proper number of neurons using the LOO validation error in terms of using (5).

Algorithm 2: OP-ELM [19].

Given:
Protein sequence dataset $\mathscr{A}$; number of ensembles $\mathbf{K}$
(1) Randomly generate $\mathbf{K}$ sets of hidden node parameters, train each SLFN via ELM and obtain the corresponding output weight matrix;
(2) For each testing protein sequence $\mathbf{x}^{\text{test}}$, obtain the estimated category indexes for all $\mathbf{K}$ SLFNs constructed in *Step 1*, update (6);
(3) Decide the class label of $\mathbf{x}^{\text{test}}$ via (7).

Algorithm 3: V-ELM.

be able to reduce the misclassification number. The proposed structure of the ensemble based SLFNs for protein sequence classification is given in Figure 2.

As shown in Figure 2, instead of using a single SLFN, multiple independent SLFNs where each one has the same structure and the same hidden node activation function are used. The protein sequence feature vectors are fed to train each SLFN separately. Then, the final category index of the testing sample is decided by majority voting among all the results obtained by these SLFNs. It is apparent that the gradient descent based methods are not suitable to adopt here as the training algorithm for each SLFN. In general, the training time cost by ensemble based SLFNs is linearly increased with a proportion to the number of independent ensembles. Since the conventional gradient descent methods normally suffer from a long training time for the high dimensional feature vectors and large sample sizes, the training time increases dramatically when employing it for multiple ensembles. Therefore, in this paper, we propose to use the basic ELM and the recent OP-ELM as the training algorithm for each SLFN for protein sequence classification as follows.

3.2.1. The Proposed Ensemble Based ELM. The voting based ELM (V-ELM) developed by Cao et al. [33] is proposed for protein sequence classification in this section. The details are described as follows. We assume that $\mathbf{K}$ SLFNs are employed as ensembles. In the first stage, $\mathbf{K}$ sets of hidden parameters $\{(\mathbf{a}_j^k, b_j^k)_{j=1,\ldots,\mathbf{J}}\}_{k=1,\ldots,\mathbf{K}}$ are randomly generated and the corresponding $\mathbf{K}$ sets of output weight matrix $\{\mathscr{W}^k\}_{k=1,\ldots,\mathbf{K}}$ are obtained with (4) using the protein sequence training dataset $\mathscr{A}$. In the second stage, for each trained ELM denoted as $S^k = \{(\mathbf{a}_j^k, b_j^k)_{j=1,\ldots,\mathbf{J}}, \mathscr{W}^k\}_{k=1,\ldots,\mathbf{K}}$, the estimated category index of the testing protein sequence sample $\mathbf{x}^{\text{test}}$ is obtained. For all $\mathbf{K}$ ensembles, a $\mathscr{C}$-dimensional vector $\mathscr{L}_{\mathbf{K},\mathbf{x}^{\text{test}}}$ is used to store all the results where if the category index derived by the kth ($k \in [1, \ldots, \mathbf{K}]$) ELM is $\imath$, the value of the corresponding entry $\imath$ in the vector $\mathscr{L}_{\mathbf{K},\mathbf{x}^{\text{test}}}$ is increased by one; that is,

$$\mathscr{L}_{\mathbf{K},\mathbf{x}^{\text{test}}}(\imath) = \mathscr{L}_{\mathbf{K},\mathbf{x}^{\text{test}}}(\imath) + 1. \tag{6}$$

In the third stage, the final category index of $\mathbf{x}^{\text{test}}$ is determined via

$$\mathbf{c}^{\text{test}} = \arg \max_{\imath \in [1,\ldots,\mathscr{C}]} \left\{\mathscr{L}_{\mathbf{K},\mathbf{x}^{\text{test}}}(\imath)\right\}. \tag{7}$$

A proposition is also given by Cao et al. [33] to illustrate the superiority of the V-ELM over the original ELM as follows.

Proposition 1 (see Cao et al. [33]). *Given a standard SLFN with* $\mathbf{J}$ *hidden nodes, an output function* $g(\cdot)$ *and a set of training samples* $\{(\mathbf{x}_i, \mathbf{t}_i)\}_{i=1}^{\mathbf{N}}$ *assume that the probability of correctly predicting the testing sample* $\mathbf{x}^{test}$ *using ELM under all different possible hidden node parameters* $\mathbf{a}$ *and* b *is* $\mathscr{P}_{ELM}(\mathbf{c} \mid \mathbf{x}^{test})$. *If the following inequality holds*

$$\mathscr{P}_{ELM}\left(\mathbf{c} \mid \mathbf{x}^{test}\right) > \max\left\{\mathscr{P}_{ELM}\left(\imath \mid \mathbf{x}^{test}\right)\right\}_{\imath \in [1,\ldots,\mathscr{C}], \imath \neq \mathscr{C}}, \tag{8}$$

where $\mathscr{P}_{ELM}(\imath \mid \mathbf{x}^{test})$ *is the probability that ELM classifies* $\mathbf{x}^{test}$ *to category* $\imath$ *that is different from the class* $\mathbf{c}$, *then, with a sufficiently large independent training number* $\mathbf{K}$, *V-ELM is able to correctly classify* $\mathbf{x}^{test}$ *with probability one.*

A brief summary of the proposed V-ELM for protein sequence classification is described in Algorithm 3.

3.2.2. The Proposed Ensemble Based OP-ELM. Besides using ELM as the training algorithm for each ensemble, we propose the voting based OP-ELM (VOP-ELM) for the protein sequence classification in this section. For each ensemble, the OP-ELM algorithm is incorporated as the training algorithm instead of using the original ELM. Similar to V-ELM, each SLFN ensemble is trained using the OP-ELM method on the

Given:
Protein sequence dataset $\mathscr{A}$; number of ensembles **K**
(1) Utilize OP-ELM to train each SLFN ensemble independently;
(2) For each testing protein sequence $\mathbf{x}^{\text{test}}$, obtain the estimated category indexes for all **K** SLFNs in *Step 1*, update (9);
(3) Decide the final class label of $\mathbf{x}^{\text{test}}$ using (10).

Algorithm 4: VOP-ELM.

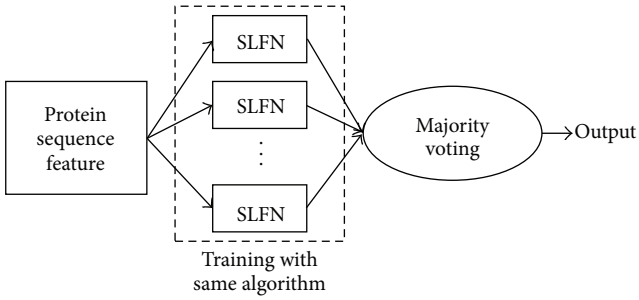


Figure 2: Ensemble structure of SLFNs for protein sequence classification.

protein sequence dataset $\mathscr{A}$ independently. After training, the category index of each testing sample $\mathbf{x}^{\text{test}}$ by all the **K** ensembles is obtained. Similar to (6), the $\mathscr{C}$-dimensional vector $\mathscr{L}^{\text{OP}}_{\mathbf{K},\mathbf{x}^{\text{test}}}$ is updated via the following equation:

$$\mathscr{L}^{\text{OP}}_{\mathbf{K},\mathbf{x}^{\text{test}}}(\iota) = \mathscr{L}^{\text{OP}}_{\mathbf{K},\mathbf{x}^{\text{test}}}(\iota) + 1. \tag{9}$$

Then, the final category index of the testing protein sequence is obtained by

$$\mathbf{c}^{\text{test}} = \arg\max_{\iota\in[1,\ldots,\mathscr{C}]} \left\{\mathscr{L}^{\text{OP}}_{\mathbf{K},\mathbf{x}^{\text{test}}}(\iota)\right\}. \tag{10}$$

A brief summary of the proposed VOP-ELM for protein sequence classification is illustrated in Algorithm 4.

We can find that the proposed VOP-ELM would have a better performance than the V-ELM for each ensemble. Compared with V-ELM, the hidden neurons of each ensemble in VOP-ELM are optimized using the OP-ELM given in Section 3.1. However, the drawback of VOP-ELM is that the cost training time may increase as for each SLFN in VOP-ELM, the MRSR technique, and the LOO validation are performed for hidden neurons selection.

4. Experimental Results

In this section, the classification results and discussions on the protein sequence dataset from the Protein Information Resource center are presented. We study the performance in two scenarios. In the first scenario, the PIR1 dataset with 949 samples is fixed as the training dataset while the PIR2 dataset with 534 samples is used as the testing dataset. In the second scenario, the PIR1 and PIR2 datasets are first mixed together and then the training and testing datasets with 949 and 534 samples, respectively, are randomly generated from the mixed dataset. For each protein sequence, a 56-dimensional feature vector is extracted as introduced in Section 2 and hence, the number of nodes in the input layer of the SLFN is 56. The proposed ensemble based VOP-ELM and V-ELM are implemented for the protein sequence classifications. The OP-ELM developed by Miche et al. [19] is also used for comparisons. To compare the classification performance, the original ELM by Huang et al. [14, 23], one of the fastest algorithms of the BP's variants, namely, the Levenberg-Marquardt method by Haykin [22], Levenberg [16], and the SVM by Hsu and Lin [21], is employed for the protein sequence classifications. Three different kernel functions are used as the activation function in the SLFN and SVM, which are the linear kernel function (denoted as **L**), the sigmoid kernel function (denoted as **S**), and the Gaussian kernel function (denoted as **G**), respectively. To maintain the training time under an acceptable level, only $\mathbf{K} = 3$ ensembles of SLFNs are used for VOP-ELM. However, for V-ELM, $\mathbf{K} = 7$ ensembles of SLFNs are employed due to that the V-ELM utilizes the original ELM as training algorithm and it learns much faster than OP-ELM. For the fairness of comparison, all these simulations are executed in the MATLAB 7.4 environment running on an ordinary PC with 3.0 GHz CPU and 4.G RAM memory. Simulations with SVM are carried out using the compiled C-code SVM package: Libsvm (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) running in the same PC. For SVM, the cost parameter c and the kernel parameter γ are searched in a grid as proposed by Hsu and Lin [21] and the best combination of these parameters is then obtained in terms of the generalized performance. For all these methods, multiple independent trials of simulation are tested and the average results are reported in the paper. For the BP method, 10 trials are used due to its long training time while, for the rest approaches, 50 trials are utilized. The number of hidden nodes for the BP method is 80. When further increasing the number of the nodes, it usually runs out of memory in our PC, which means the computational complexity of the BP algorithm is very high when processing the protein sequence. All the features of the protein sequence are normalized within the region $[-1, 1]$.

4.1. Performance on Fixed Training and Testing Datasets. In this section, we fix the protein sequence PIR1 as the training dataset while the protein sequence PIR2 is used as the testing dataset. Table 1 shows the comparisons of the average successful testing classification rate (Rate) and its standard

Table 1: Successful testing classification rate and standard derivation (rate (%) ± Dev (%)) comparisons of different classifiers: fixed protein sequence training dataset (pir1) and testing dataset (pir2), where **L**, **S**, and **G** stand for the linear kernel, the sigmoid kernel, and the Gaussian kernel, respectively.

Methods	L	S	G
VOP-ELM	**92.51** ± 0	**92.64** ± 1.6	**93.69** ± 0.89
V-ELM	87.45 ± 0.01	90.14 ± 0.71	90.46 ± 0.69
OP-ELM	91.39 ± 0.01	91.51 ± 1.0	92.28 ± 0.70
SVM	90.63 ± 0.01	90.63 ± 0.01	90.63 ± 0.01
BP	85.77 ± 0.42	88.39 ± 1.69	88.77 ± 1.35
ELM	85.39 ± 0	87.66 ± 1.2	87.80 ± 1.4

derivation (Dev) among multiple trials for all the classifiers. Table 2 gives the corresponding training time cost by all the classifiers and their comparisons.

As highlighted in the boldface in Table 1, the proposed ensemble based VOP-ELM has the highest successful classification rate among all 6 approaches. It can be seen that the proposed VOP-ELM offers an improvement of 7.12% and 1.12%, an improvement of 4.98% and 1.13%, an improvement of 5.89% and 1.41% over the original ELM and the OP-ELM by using the linear kernel **L**, the sigmoid kernel **S**, and the Gaussian kernel **G**, respectively. The V-ELM method performs better than the original ELM and BP in general. The improvement of the classification rate obtained by V-ELM over the original ELM is 2.06%, 2.48%, and 2.66% with the linear kernel, the sigmoid kernel, and the Gaussian kernel, respectively. But the performance of V-ELM is slightly worse than VOP-ELM, OP-ELM, and SVM. It is also worth pointing out that, for all these approaches, using nonlinear kernels (the **S** and **G** functions) generally achieves a higher recognition rate than using the linear kernel. As expected, the original ELM has the fastest training speed. As shown in Table 2, the training phase of the original ELM can be finished less than 1 second (s). It learns ten thousand times faster than BP, thousand times faster than SVM and VOP-ELM (the **S** and **G** kernels), and hundred times faster than OP-ELM (the **S** and **G** kernels). In addition, the training time cost by V-ELM is linearly proportional to the number of ensembles we have used as compared to the training time by the original ELM. With only 7 ensembles, the training phase of V-ELM still can be finished within 1 second, as underlined in Table 1.

In general, a large number of SLFNs used as ensembles usually guarantees a higher classification rate than using a small number of SLFNs as ensembles, but the training time also linearly increases. To illustrate this, three different numbers of ensembles with $\mathbf{K} = 5$, $\mathbf{K} = 7$, and $\mathbf{K} = 15$ are tested on different hidden nodes in each SLFN. The performance is compared with the original ELM. Figures 3 and 4 depict the classification rate and training time of ELM and V-ELM on using the protein sequence PIR1 as training dataset and the PIR2 as the testing dataset. As demonstrated in these two figures, the classification rate increases along with the number of ensembles while the training time also increases gradually. Even though, we can still verify from Figure 4 that V-ELM can finish the training phase within

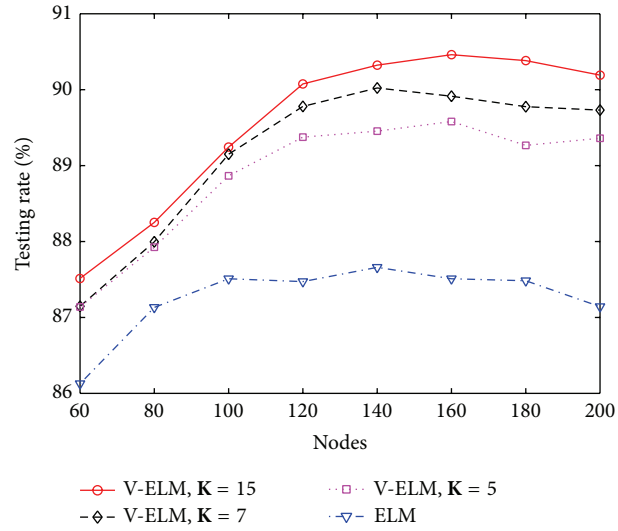


Figure 3: The classification rates of ELM and V-ELM w.s.t. different nodes on fixed training and testing protein sequence datasets.

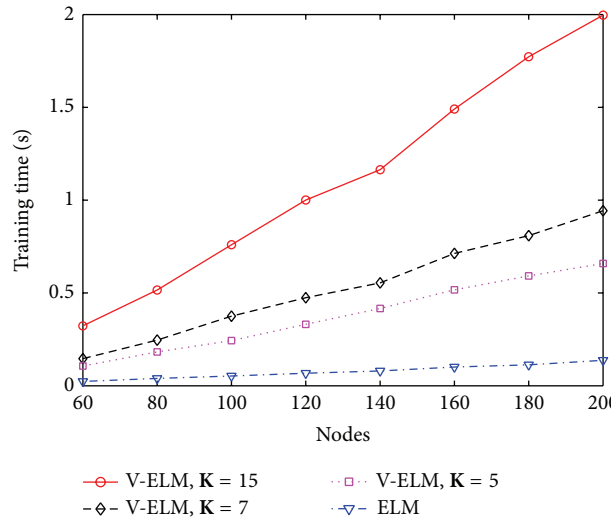


Figure 4: The training time of ELM and V-ELM w.s.t. different nodes on fixed training and testing protein sequence datasets.

several seconds in general, which makes it acceptable for large sample sizes protein sequence applications. The performance of VOP-ELM with respect to (w.s.t) different numbers of ensembles is not shown here. This is because when further increasing the numbers of ensembles in VOP-ELM to 7 or more, the training time jumps to hundreds or thousands seconds. It will affect its feasibility of applications on large protein datasets.

4.2. Performance on Randomly Generated Training and Testing Datasets. In this experiment, the protein sequence datasets PIR1 and PIR2 are first mixed into one file after the feature extraction. Then, for each trial, 939 samples are randomly generated from the whole dataset as the training dataset and the rest samples are assigned to the testing dataset. Table 3

Table 2: Training time comparisons of different classifiers: fixed protein sequence training dataset (pir1) and testing dataset (pir2), where **L**, **S**, and **G** stand for the linear kernel, the sigmoid kernel, and the Gaussian kernel, respectively.

Methods	L		S		G	
	Training time (s)	Speedup	Training time (s)	Speedup	Training time (s)	Speedup
VOP-ELM	5.5685	105.4	86.80	13.37	95.22	13.93
V-ELM	*0.4590*	1279.3	*0.9020*	1286.14	*0.8680*	1528.34
OP-ELM	1.8588	315.90	30.11	38.53	30.45	43.57
SVM	236.16	2.486	236.67	4.90	243.10	5.46
BP	587.20	1	1160.1	1	1326.6	1
ELM	**0.0612**	9594.8	**0.0799**	14519	**0.0792**	16750

Table 3: Successful testing classification rate and standard derivation (rate (%) ± Dev (%)) comparisons of different classifiers: randomly generated training and testing datasets from the mixed protein sequences, where **L**, **S**, and **G** stand for the linear kernel, the sigmoid kernel, and the Gaussian kernel, respectively.

Methods	L	S	G
VOP-ELM	**97.30** ± 0.73	**98.19** ± 0.70	**98.68** ± 0.71
V-ELM	96.91 ± 0.71	97.75 ± 0.64	97.74 ± 0.59
OP-ELM	95.95 ± 0.29	96.42 ± 0.45	97.55 ± 0.55
SVM	97.17 ± 0.49	97.28 ± 0.63	97.33 ± 0.66
BP	94.93 ± 0.95	96.29 ± 0.85	95.54 ± 0.91
ELM	94.94 ± 0.75	96.72 ± 0.85	96.65 ± 0.63

lists the average successful testing classification rate and the standard derivation of the 6 classifiers among all trials. The training time for the 6 classifiers and their comparisons is shown in Table 4.

From Table 3, it can be seen that the proposed VOP-ELM wins the highest classification rate among all classifiers for all 3 kernels. The VOP-ELM achieves the classification rates of 97.30%, 98.19%, and 98.68% by using the linear, sigmoid, and Gaussian kernels, respectively. The improvements offered by the VOP-ELM method over the original ELM, BP, and OP-ELM are around 2.36%, 2.37%, and 1.35% with the linear kernel, 1.47%, 1.90%, and 1.77% with the sigmoid kernel, and 2.03%, 3.14%, and 1.13% with the Gaussian kernel, respectively. The ensemble based V-ELM performs slightly worse than the VOP-ELM. However, V-ELM is better than the rest methods in general when using the nonlinear kernels. Similar to the experiment in Section 4.1, for the randomly generated training and testing protein sequences, the classifier employing the nonlinear kernel functions (**S** and **G**) generally outperforms using the linear kernel function (**L**).

As shown in boldface in Table 4, ELM is the fastest method and the training phase of the protein sequence dataset takes less than 1 second. However, the conventional BP method takes more than 1 thousand seconds when employing the linear and Gaussian kernels and more than 950 seconds when using the sigmoid kernel. The training time cost by SVM is more than 270 seconds for all the three kernels. Although the original ELM algorithm is employed by the proposed VOP-ELM and OP-ELM, the cost training times jump to dozens of seconds due to utilizing the framework of searching the optimal hidden neurons and multiple ensembles. With only using 7 ensembles in V-ELM, its training time is at the comparable level as the original ELM (as underlined in Table 4). The enhancement of classification rates obtained by V-ELM over the original ELM is 1.94%, 1.03%, and 1.09% for the linear, sigmoid, and Gaussian kernels, respectively.

To further study the performance of V-ELM w.s.t. the number of ensembles, three different numbers of ensembles $\mathbf{K} = 5$, $\mathbf{K} = 7$, and $\mathbf{K} = 15$ are tested on the randomly generated protein sequence datasets. The classification rates and their corresponding training time are compared with the original ELM by using the sigmoid kernel and the results are depicted in Figure 5 and Figure 6, respectively. As illustrated in these two figures, the classification rates obtained by V-ELM gradually increase when the used ensembles are increased from 5 to 15. All the results obtained by V-ELM are better than the original ELM. Although the training times also increase linearly along with the numbers of ensembles, the training phase for V-ELM still can be finished less than 2 seconds, as shown in Figure 6. Similar to Section 4.1, the performance of the proposed VOP-ELM w.s.t. the number of ensembles is not studied in this section because when more than 7 ensembles are used, the training time cost by VOP-ELM will jump to hundreds or thousands seconds. For the large protein sequence dataset, it may not be feasible.

From Tables 1 and 3, it is interesting to see that although the number of samples used from training dataset is the same, the classification rate obtained by using randomly generated protein sequence training dataset is higher than the one by fixing the PIR1 as the training dataset. The reason behind this may be explained as follows. When fixing the PIR1 as the training dataset, some families of the protein sequences become an imbalance problem, such as the families of Cytochrome c, Cytochrome c6, Triosephosphate isomerase, and Globin as shown in Section 2. In such case, the testing samples have a high risk to be misclassified. However, when generating the training and testing dataset from the whole pool, the samples have equal probability to be partitioned into the training dataset and the testing dataset. Hence, the classification model would be trained with a balance dataset, which may result in a high testing rate.

4.3. Performance on Different Numbers of Ensembles. In this section, we show the relationship between the number of ensembles used in the proposed algorithms and the

Table 4: Training time comparisons of different classifiers: randomly generated training and testing datasets from the mixed protein sequences, where **L**, **S**, and **G** stand for the linear kernel, the sigmoid kernel, and the Gaussian kernel, respectively.

Methods	L		S		G	
	Training time (s)	Speedup	Training time (s)	Speedup	Training time (s)	Speedup
VOP-ELM	5.2799	266.44	83.55	11.38	92.00	12.43
V-ELM	*0.6326*	2223.83	*0.6917*	1374.51	*0.9282*	1232.17
OP-ELM	1.8627	755.24	30.93	30.74	32.59	35.09
SVM	278.35	5.05	284.30	3.34	272.41	4.2
BP	1406.8	1	950.75	1	1143.70	1
ELM	**0.0746**	18858	**0.0861**	11042	**0.0983**	11635

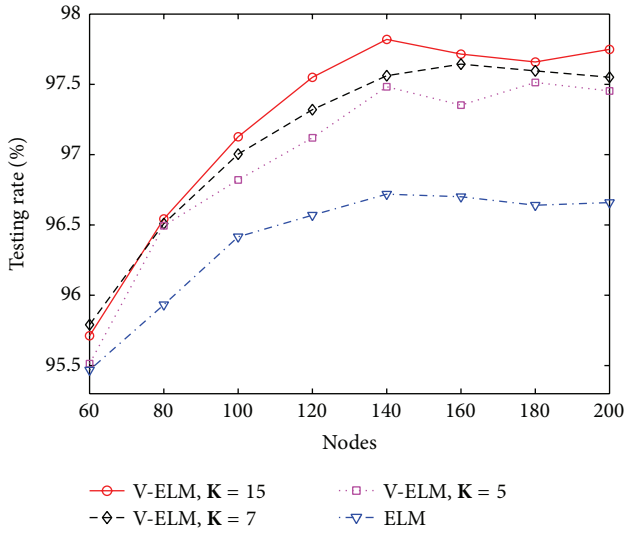


Figure 5: The testing rates of ELM and V-ELM w.s.t. different nodes on randomly generated training and testing protein sequence datasets.

classification performance on the protein sequences. The proper number of ensembles which should be adopted in the proposed algorithms is also discussed in the section through the simulations. In the experiments, the number of ensembles is increased from 1 to 51 with an interval 2. The sigmoid kernel is utilized for the experiments as the performance is similar to the one using the other two kernels. The average testing rate for each number of ensembles with 50 independent trials is recorded for both the V-ELM algorithm and the VOP-ELM method. The experiments are running using the PIR1 as the training dataset only. For randomly generated protein sequence dataset, we can find the similar trend of the testing rate and training time on different numbers of ensembles as the one using fixed training dataset. Therefore, the performance on randomly generated training dataset is not repeated here.

Figure 7 depicts the testing rate on different numbers of ensembles for the fixed protein sequence training dataset while Figure 8 shows its corresponding training time. As we can find from Figure 7, the testing rate generally increases when the number of ensembles is increasing for both two algorithms. However, the increment is very small when the number of ensemble is larger than 15. It is readily to see that the improvements from using 1 ensemble to using 15 ensembles are around 2.5% and 3.5% for V-ELM and VOP-ELM, respectively. However, when further increasing the number of ensembles from 15 to 51, the improvements of 51 ensembles over 15 ensembles are only around 0.4% and 0.5%, respectively. But the training increases dramatically, especially for the VOP-ELM method. As shown in Figure 8, the training cost by VOP-ELM with 51 ensembles is longer than 1500 seconds. In addition, the training time used for both the two methods is proportional to the training time by the original ELM and OP-ELM, respectively. The proportional rate is related to the number of ensembles. Likewise, we can obtain the similar performance when using randomly generated training protein sequence dataset. Hence, considering both the enhancement of the classification rate and the increment of the training time, we suggest that the best choice of the number ensembles for both V-ELM and VOP-ELM should be less than 15.

5. Discussions and Conclusions

As demonstrated by the experimental results in Section 4, the proposed ensemble based VOP-ELM has the highest classification rate and outperforms V-ELM, OP-ELM, SVM, BP, and the original ELM algorithms for the protein sequence recognition. However, employing OP-ELM for each ensemble also increases the training time of the VOP-ELM method, as expected. Even though, VOP-ELM still learns much faster than the conventional BP and the popular SVM.

The original ELM has the fastest training speed of the SLFN model for the protein sequence classification problem among all classifiers. However, the classification performance on protein sequence by ELM is only comparable to the BP method and worse than the proposed ensemble based VOP-ELM and V-ELM, the OP-ELM, and SVM, in general. In addition, using the linear kernel can reduce the training time for some algorithms, such as the OP-ELM and the VOP-ELM algorithms. But the classification rate also reduces when employing the linear kernel as activation function. For the rest of algorithms, the training time used by the linear function is close to the one with the nonlinear kernels, as shown in Tables 2 and 4, respectively. Hence, how to choose the kernel function depends on the requirement of the protein sequence applications.

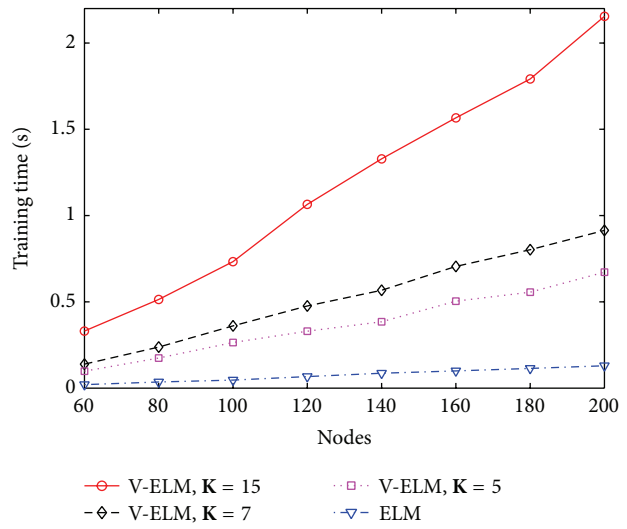


Figure 6: The training time of ELM and V-ELM w.s.t. different nodes on randomly generated training and testing protein sequence datasets.

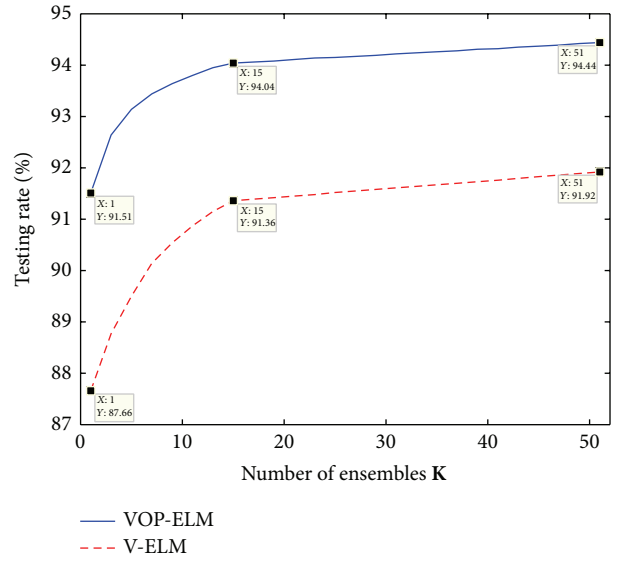


Figure 7: The testing rates of VOP-ELM and V-ELM w.s.t. different numbers of ensembles.

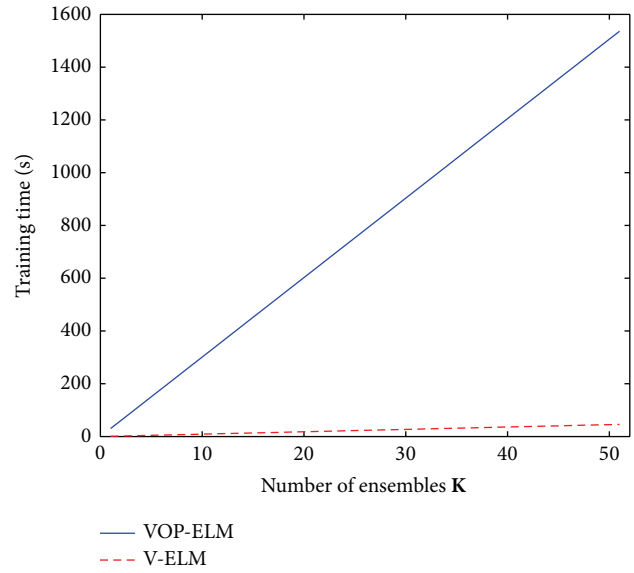


Figure 8: The training time of VOP-ELM and V-ELM w.s.t. different numbers of ensembles.

Employing ELM as the training algorithm for each ensemble, the developed V-ELM outperforms the original ELM in general. The increment of V-ELM compared to ELM is more than 1% using the fixed protein sequence training dataset and is more than 2% using the randomly generated protein sequence training dataset for all the 3 kernels, respectively. Although the training time by V-ELM is longer than the one by ELM, the training phase for the protein sequence dataset still can be finished less than 1 second. Beside ELM, V-ELM can learn many times faster than the rest of compared algorithms. For certain applications, such as developing the online training model for protein sequence analysis, the original ELM and the V-ELM algorithm would be the best choices as researchers normally do a batch model retraining by using the past dataset combined with the new coming samples.

Therefore, as stated above, how to choose the proper classifier for the protein sequence classification depends on the requirements of the application. If only requiring the high recognition rate, the proposed VOP-ELM is the best choice. If the learning speed of the model for the protein sequence is the only concern, the original ELM is the proper classifier. However, if both the high classification rate and the fast training speed are required, the developed V-ELM and the existed OP-ELM should be used.

In conclusion, we have studied the protein sequence classification problem based on SLFNs in this paper. The existed OP-ELM has been first employed as the classifier. Then, the ensemble based structure of SLFNs has been proposed to enhance the performance of protein sequence classification. Two algorithms, namely, V-ELM and VOP-ELM, have been developed for protein sequence classifications by employing the original ELM and the OP-ELM to train each ensemble, respectively. Experimental results on the protein sequence datasets from the Protein Information Resource center demonstrated that VOP-ELM has the highest recognition rate among all compared state-of-art methods while the V-ELM outperforms the original ELM and BP but maintains the training speed as comparable as the original ELM. Moreover, to obtain a reasonable testing rate with an acceptable training time for the ensemble based algorithms, we have shown by simulations that the proper number of ensemble should be chosen less than 15.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China under Grant 61333009 and in part by the National Basic Research Program of China under Grant 2012CB821200.

References

[1] A. Barve, S. Ghaskadbi, and S. Ghaskadbi, "Structural and sequence similarities of hydra xeroderma pigmentosum a protein to human homolog suggest early evolution and conservation," *BioMed Research International*, vol. 2013, Article ID 854745, 9 pages, 2013.

[2] C. Chen, P. B. McGarvey, H. Huang, and C. H. Wu, "Protein bioinformatics infrastructure for the integration and analysis of multiple high-throughput omics data," *Advances in Bioinformatics*, vol. 2010, Article ID 423589, 2010.

[3] H. Cong, M. Zhang, Q. Zhang et al., "Analysis of structures and epitopes of surface antigen glycoproteins expressed in bradyzoites of *Toxoplasma gondii*," *BioMed Research International*, vol. 2013, Article ID 165342, 9 pages, 2013.

[4] J. Machado, A. C. Costa, and M. Quelhas, "Can power laws help us understand gene and proteome information?" *Advances in Mathematical Physics*, vol. 2013, Article ID 917153, 10 pages, 2013.

[5] V. Carregari, R. Floriano, L. Rodrigues-Simioni et al., "Biochemical, pharmacological, and structural characterization of new basic PLA_2 bbil-tx from *Bothriopsis bilineata* snake venom," *BioMed Research International*, vol. 2013, Article ID 612649, 12 pages, 2013.

[6] L. Liu, J. Cui, X. Zhang, T. Wei, P. Jiang, and Z. Wang, "Analysis of structures, functions, and epitopes of cysteine protease from *Spirometra erinaceieuropaei* spargana," *BioMed Research International*, vol. 2013, Article ID 198250, 7 pages, 2013.

[7] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*, The MIT Press, Cambridge, Mass, USA, 2001.

[8] Y. Yang, B.-L. Lu, and W.-Y. Yang, "Classification of protein sequences based on word segmentation methods," in *Proceedings of the 6th Asia-Pacific Bioinformatics Conference (APBC '08)*, vol. 6, pp. 177–186, Imperial College Press, 2008.

[9] C. Caragea, A. Silvescu, and P. Mitra, "Protein sequence classification using feature hashing," *Proteome Science*, vol. 10, no. 1, pp. 1–8, 2012.

[10] D. Wang, N. K. Lee, T. S. Dillon, and N. J. Hoogenraad, "Protein sequences classification using radial basis function (RBF) neural networks," in *Proceedings of the 9th International Conference on Neural Information Processing*, vol. 2, pp. 764–768, 2002.

[11] D. Wang, N. K. Lee, and T. S. Dillon, "Extraction and optimization of fuzzy protein sequences classification rules using GRBF neural networks," *Information Processing Letters and Reviews*, vol. 1, no. 1, pp. 53–559, 2003.

[12] D. Wang and G.-B. Huang, "Protein sequence classification using extreme learning machine," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '05)*, vol. 3, pp. 1406–1411, Montreal, Canada, 2005.

[13] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, pp. 985–990, 2004.

[14] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.

[15] G.-B. Huang, Q.-Y. Zhu, K. Z. Mao, C.-K. Siew, P. Saratchandran, and N. Sundararajan, "Can threshold networks be trained directly?" *IEEE Transactions on Circuits and Systems II*, vol. 53, no. 3, pp. 187–191, 2006.

[16] K. Levenberg, "A method for the solution of certain problems in least squares," *The Quarterly of Applied Mathematics*, vol. 2, pp. 164–168, 1944.

[17] D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM Journal on Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[18] J. Cao, Z. Lin, and G.-B. Huang, "Self-adaptive evolutionary extreme learning machine," *Neural Processing Letters*, vol. 36, no. 3, pp. 285–305, 2012.

[19] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse, "OP-ELM: optimally pruned extreme learning machine," *IEEE Transactions on Neural Networks*, vol. 21, no. 1, pp. 158–162, 2010.

[20] T. Simila and J. Tikka, "Multiresponse sparse regression with application to multidimensional scaling," in *Proceedings of 15th International Conference Artificial Neural Networks: Formal Models and Their Applications*, pp. 97–102, 2005.

[21] 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.

[22] S. Haykin, *Neural Networks, A Comprehensive Foundation*, Pearson Education, Upper Saddle River, NJ, USA, 2001.

[23] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 42, no. 2, pp. 513–529, 2012.

[24] D. Serre, *Matrices: Theory and Applications*, Springer, New York, NY, USA, 2002.

[25] G.-B. Huang, M.-B. Li, L. Chen, and C.-K. Siew, "Incremental extreme learning machine with fully complex hidden nodes," *Neurocomputing*, vol. 71, no. 4–6, pp. 576–583, 2008.

[26] N.-Y. Liang, P. Saratchandran, G.-B. Huang, and N. Sundararajan, "Classification of mental tasks from EEG signals using extreme learning machine," *International Journal of Neural Systems*, vol. 16, no. 1, pp. 29–38, 2006.

[27] R. Zhang, G.-B. Huang, N. Sundararajan, and P. Saratchandran, "Multi-category classification using an extreme learning machine for microarray gene expression cancer diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 3, pp. 485–494, 2007.

[28] J. Cao, Z. Lin, and G.-B. Huang, "Composite function wavelet neural networks with extreme learning machine," *Neurocomputing*, vol. 73, no. 7–9, pp. 1405–1416, 2010.

[29] Q.-Y. Zhu, A.-K. Qin, P.-N. Suganthan, and G.-B. Huang, "Evolutionary extreme learning machine," *Pattern Recognition*, vol. 38, no. 10, pp. 1759–1763, 2005.

[30] J. Cao, Z. Lin, and G.-B. Huang, "Composite function wavelet neural networks with differential evolution and extreme learning machine," *Neural Processing Letters*, vol. 33, no. 3, pp. 251–265, 2011.

[31] G. Bontempi, M. Birattari, and H. Bersini, "Recursive lazy learning for modeling and control," in *Proceedings of the 10th European Conference on Machine Learning*, pp. 292–303, 1998.

[32] R. Myers, *Classical and Modern Regression with Applications*, Duxbury Press, 2nd edition, 2000.

[33] J. Cao, Z. Lin, G.-B. Huang, and N. Liu, "Voting based extreme learning machine," *Information Sciences*, vol. 185, no. 1, pp. 66–77, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 371397, 9 pages
http://dx.doi.org/10.1155/2014/371397

Research Article

Identifying Gastric Cancer Related Genes Using the Shortest Path Algorithm and Protein-Protein Interaction Network

Yang Jiang,[1] Yang Shu,[2] Ying Shi,[3] Li-Peng Li,[1] Fei Yuan,[2] and Hui Ren[4]

[1] *Colorectal Surgery Department, China-Japan Union Hospital of Jilin University, Changchun 130033, China*
[2] *State Key Laboratory of Medical Genomics, Institute of Health Sciences, Chinese Academy of Sciences, Shanghai Jiao Tong University School of Medicine and Shanghai Institutes for Biological Sciences, Shanghai 200025, China*
[3] *Breast and Thyroid Surgery Department, The Second Hospital of Jilin University, Changchun 130041, China*
[4] *Colorectal Surgery Department, The Second Hospital of Jilin University, Changchun 130041, China*

Correspondence should be addressed to Hui Ren; hren@jlu.edu.cn

Received 29 December 2013; Accepted 3 February 2014; Published 5 March 2014

Academic Editor: Tao Huang

Gastric cancer, as one of the leading causes of cancer related deaths worldwide, causes about 800,000 deaths per year. Up to now, the mechanism underlying this disease is still not totally uncovered. Identification of related genes of this disease is an important step which can help to understand the mechanism underlying this disease, thereby designing effective treatments. In this study, some novel gastric cancer related genes were discovered based on the knowledge of known gastric cancer related ones. These genes were searched by applying the shortest path algorithm in protein-protein interaction network. The analysis results suggest that some of them are indeed involved in the biological process of gastric cancer, which indicates that they are the actual gastric cancer related genes with high probability. It is hopeful that the findings in this study may help promote the study of this disease and the methods can provide new insights to study various diseases.

1. Introduction

Gastric carcinogenesis is a multistep process involving genetic and epigenetic alteration of protein-coding protooncogenes and tumor-suppressor genes. Gastric cancer (GC) is the fourth most commonly diagnosed cancer and is estimated to be the second most common cause of cancer related death and causes about 800,000 deaths worldwide per year [1, 2]. Because of the improvement of the dietary structure, the mortality rate shows a declining trend worldwide [3]. However, the incidences of gastric cancer are still remarkable in areas where infection by *Helicobacter pylori* is prevalent [4]. Besides *H. pylori*, smoking and alcohol consumption also increase the risk of developing gastric cancer significantly [5, 6]. Compared with women, men have a higher incidence, while estrogen may protect women against the gastric cancer [7].

In the previous cases, over 90% gastric cancers are adenocarcinomas, which could be divided into two major types in terms of the histopathology [8]. Intestinal type gastric cancer is often related to environmental factors such as *H. pylori*, while diffuse type gastric cancer is more often associated with genetic abnormalities. Caldas et al. reviewed that the diffuse type gastric cancer tended to occur in female and young individuals [9]. Besides adenocarcinomas, other types of gastric cancers like lymphomas occurred in a very low incidence [10]. Since the gastric cancer leads to high mortality, the early diagnose especially the molecular diagnose is particularly important for the therapy.

So far, numerous genes have been found involved in gastric tumorigenesis. Among the reported gastric cancer related genes, most of them could have also been found in other types of carcinomas. p53, famous for its tumor-suppressing role, has a mutated rate ranging from 0 to 21% in diffuse type GC and 36–43% in intestinal type GC [11]; E-cadherin, which plays a pivotal role in EMT (Epithelial Mesenchymal Transition), is predisposed to mutagenesis in sporadic diffuse type GC (33–50%) [12]; another star gene

harboring high correlation with gastric cancer is RNUX3, which manifests to be a tumor-suppressor gene of GC [13]. Although dozens of genes have been found related to gastric cancer, they are insufficient to elucidate the tumorigenesis of GC unless more relevant genes being uncovered.

It is time-consuming to discover novel gastric cancer related genes by experiment alone, because the search space is very large. Computational approach is an alternative way which can help investigators screen out some related genes. On the other hand, lots of computational approaches have been developed to settle various biological problems, such as drug design [14–19] and analysis of complicated biological network [20–24]. In this study, a computational method was built to discover novel gastric cancer related genes based on some known related ones retrieved from Gastric Cancer Database, UniProtKB, and TSGene Database. After applying the shortest path algorithm in protein-protein interaction network to search the shortest path connecting any pair of known gastric cancer related genes, the candidate genes were found. Further analysis suggests that some of them are related to the formation and development of gastric cancer. We hope that this contribution may give help to uncover the mechanism of this disease, thereby designing effective treatments.

2. Materials and Methods

2.1. Materials. Gastric cancers related genes are collected from the following three datasets: (1) 102 genes are picked up from Gastric Cancer Database (http://www.gastric-cancer.site40.net/); (2) 128 reviewed gastric cancer related genes were found in the UniProtKB (Protein Knowledgebase, http://www.uniprot.org/uniprot/) by setting the keyword as human gastric cancer oncogene/suppressor gene, where 86 are oncogenes and 42 are suppressor genes; (3) 9 genes were obtained from TSGene Database (Tumor Suppressor Gene Database, http://bioinfo.mc.vanderbilt.edu/TSGene/) by searching the human gastric cancer in the Literature Search box. After combining these genes, we obtained 150 gastric cancer related genes, which were available in Supplemenary Material I (available online at http://dx.doi.org/10.1155/2014/371397).

2.2. Protein-Protein Interaction (PPI) Network. It is known that interactions of proteins are important for the majority of biological functions. Many studies have shown that proteins in one interaction always share similar functions [25–29]. Since gastric cancers related genes may have some common features, it is feasible to discover novel gastric cancers related genes based on known related ones and PPI network. In this study, the PPI network was constructed based on the protein interaction information retrieved from STRING (Search Tool for the Retrieval of Interacting Genes/Proteins, http://string.embl.de/) (version 9.0) [30], a well-known database integrating known and predicted protein interactions. In the obtained file, each interaction consists of two protein IDs and a score measuring the likelihood of the interaction's occurrence. For later formulation, the score of the interaction between proteins p_1 and p_2 was denoted by $I(p_1, p_2)$. To construct the weighted network, proteins in the STRING were taken as nodes and two nodes were adjacent if and only if the score of the interaction between the corresponding proteins was greater than zero. In addition, the score of the interaction was used to label the weight of the corresponding edge as follows:

$$w(v_1, v_2) = 1000 - I(p_1, p_2), \quad (1)$$

where p_i $(i = 1, 2)$ was the corresponding protein of node v_i.

2.3. Shortest Path Genes. As described in Section 2.1, 150 gastric cancer related genes were collected, which must have some common features related to gastric cancer. On the other hand, according to Section 2.2, two proteins in one interaction, that is, they are adjacent in the constructed PPI network, always share common features. It can be further deduced that proteins in the shortest path connecting two known gastric cancer related genes may share some common features that the two known gastric cancer related genes have. Therefore, we searched the shortest path between any pair of known gastric cancer related genes by Dijkstra's algorithm, the most famous shortest path algorithm proposed by Dijkstra in 1956 [31].

After collecting the shortest paths connecting any pair of known gastric cancer related genes, we found that some nodes/genes occurred in many paths, while the majority of nodes/genes in PPI network were not in any path. To distinguish these nodes/genes, the betweenness of each node/gene was calculated, which is defined as the number of the shortest paths containing the node/gene as an inner node. Since the concept of betweenness accounts for direct and indirect influences of proteins at distant network [32], it has been employed in the study of various natural and man-made networks [33–38].

It is easy to see that genes with high betweenness may share more features related to gastric cancer than those with low betweenness, while the likelihood of gene with betweenness equal to 0 to be the novel gastric cancer related gene is zero. Accordingly, we picked out genes with betweenness greater than 0 and termed them as the shortest path genes. Since the main purpose of this study is to discover novel gastric cancer related genes, the known gastric cancer related genes were not included in the set of shortest path genes.

2.4. Further Filtering Based on Permutation Test. As described in Section 2.3, some of the shortest path genes can be obtained based on their betweenness. However, the betweenness of some nodes may be strongly influenced by the essential structure of the network. For example, the cut-vertex of the network may always receive high betweenness easier than other vertices. To control this false discovery, a permutation test was conducted to further filter these shortest path genes as follows.

(i) Randomly select 1,000 gene sets $G_1, G_2, \ldots, G_{1000}$ in PPI network with the same size of known gastric cancer related gene set.

(ii) Calculate the betweenness of each shortest path gene on each gene set G_i $(1 \le i \le 1000)$.

(iii) The permutation FDR of the shortest path gene p was computed by

$$\mathrm{FDR}(p) = \frac{\sum_{i=1}^{1000} \delta_i}{1000}, \quad (2)$$

where δ_i was defined to be 1 if the betweenness of p on G_i was greater than that of p on the known gastric cancer related gene set.

It is obvious that smaller permutation FDR of one shortest path gene indicates that it is the actual gastric cancer related gene with high possibility.

2.5. Gene Set Enrichment Analysis. DAVID [39] is a functional annotation tool, which has been widely used to analyze gene lists derived from different biological problems [40–45]. Here, it was also employed for KEGG pathway and GO enrichment analysis of the obtained gene set. The enrichment P value was corrected to control family-wise false discovery rate under certain rate (e.g., ≤0.05) with Benjamin multiple testing correction method [46]. During the enrichment analysis, all genes in the human genome were considered. 13 items in the output of DAVID and their meanings are listed in Table 1. For detailed description, please see Huang et al.'s study [39].

Table 1: Items in the output of DAVID and their meanings.

Item	Meaning
Category	DAVID category, that is, KEGG or GO
Term	Gene set name
Count	The number of genes associated with this gene set
Percentage	Calculated by "gene associated with this gene set"/"total number of query genes"
P value	Modified Fisher Exact P value
Genes	The list of genes from your query set that are annotated to this gene set
List total	The number of genes in your query list mapped to any gene set in this ontology
Pop hits	The number of genes annotated to this gene set on the background list
Pop total	The number of genes on the background list mapped to any gene set in this ontology
Fold enrichment	The ratio of the proportions on query genes and the background information which are associated with the gene set
Bonferroni	Bonferroni adjusted P value
Benjamini	Benjamini adjusted P value
FDR	FDR adjusted P value

3. Results and Discussion

3.1. Candidate Genes. Of the 150 known gastric cancer related genes, the shortest path connecting any pair of them was searched in PPI network constructed in Section 2.2. After counting the betweenness of each gene in PPI network, 466 shortest path genes with betweenness greater than zero were retrieved. These 466 genes and their betweenness can be found in Supplementary Material II. To exclude the false discovery, the permutation test was conducted. The permutation FDRs of 466 shortest path genes were calculated by (2) and also listed in Supplementary Material II. It can be observed that 144 genes were with permutation FDRs no more than 0.1. These genes were considered to have a strong relationship with gastric cancer.

3.2. Results of Gene Set Enrichment Analysis. DAVID, as a functional annotation tool, was employed to analyze the 144 shortest path genes. The analysis results included two categories: GO and KEGG. These results were available in Supplementary Materials III and IV, respectively. The detailed discussion based on these results was as follows.

From Supplementary Material III, 294 GO terms were enriched by the 144 genes. We investigated the first 10 GO terms in the list, which were shown in Figure 1. The "Count" items in the output of DAVID for these 10 GO terms were also shown in Figure 1. Among these 10 GO terms, 5 out of the 10 GO terms are cellular component (CC) GO terms including (1) GO:0005654: nucleoplasm ("count" = 30); (2) GO:0031981: nuclear lumen ("count" = 34); (3) GO:0043233: organelle lumen ("count" = 37); (4) GO:0031974: membrane-enclosed lumen ("count" = 37); (5) GO:0005829: cytosol ("count" = 30). As we know, tumorigenesis is a very complicated biological process which means the transform processes could take place everywhere in the cells [47]. In our analysis results, the related proteins distribute both in nuclear and cytosol which is in accordance with the characters of the gastric cancer. The remaining 5 GO terms are biological process (BP) GO terms: (1) GO:0032268: regulation of cellular protein metabolic process ("count" = 20); (2) GO:0009725: response to hormone stimulus ("count" = 17); (3) GO:0031399: regulation of protein modification process ("count" = 15); (4) GO:0009719: response to endogenous stimulus ("count" = 17); (5) GO:0010604: positive regulation of macromolecule metabolic process ("count" = 25). Liu et al. reported that the cancer cells usually harbor abnormal metabolic status [48]. In our results 80% (4/5) BP are relative to the metabolic stress response by means of direct regulation of the metabolic process or indirect regulation by altering the stimulus-related pathways. Besides, protein modification, which is also enriched in our results, plays an important role in the carcinogenesis by altering the pivotal proteins [49]. Although these genes may not be the indispensable factors in gastric cancer, the common points among them would give us the hints about the tumorigenesis of the gastric cancer.

From Supplementary Material IV, 8 KEGG pathways were enriched by 144 genes, which were shown in Figure 2. It can be observed that 6 out of 8 KEGG pathways were with P value less than 0.05, which were investigated as follows. The first pathway was hsa04110: cell cycle pathway ("count" = 10). 10 genes including PCNA, MYC, and CCND1 are enriched

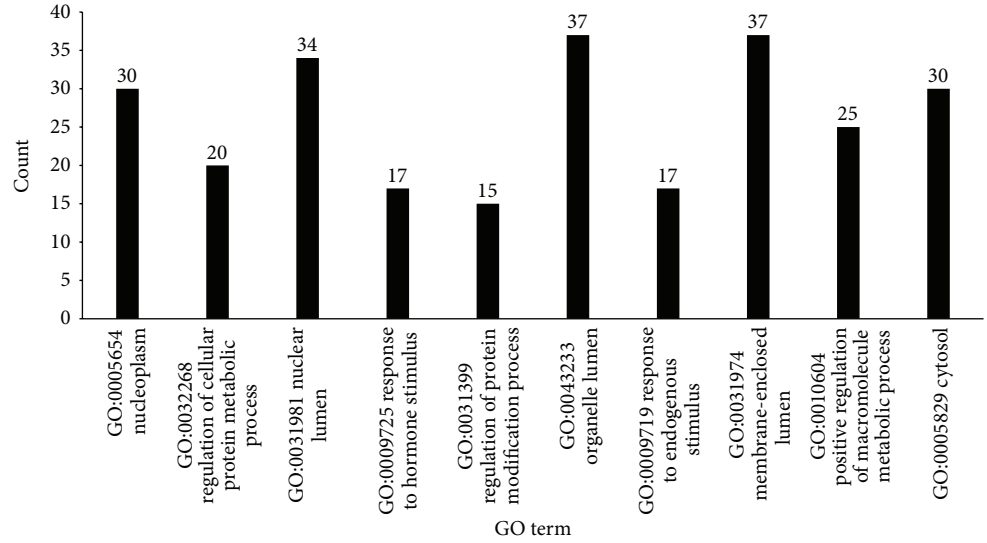


Figure 1: The top 10 GO terms enriched by 144 genes. The x-axis lists GO's ID and name, while the y-axis represents the number of genes that shared the GO term among the 144 genes.

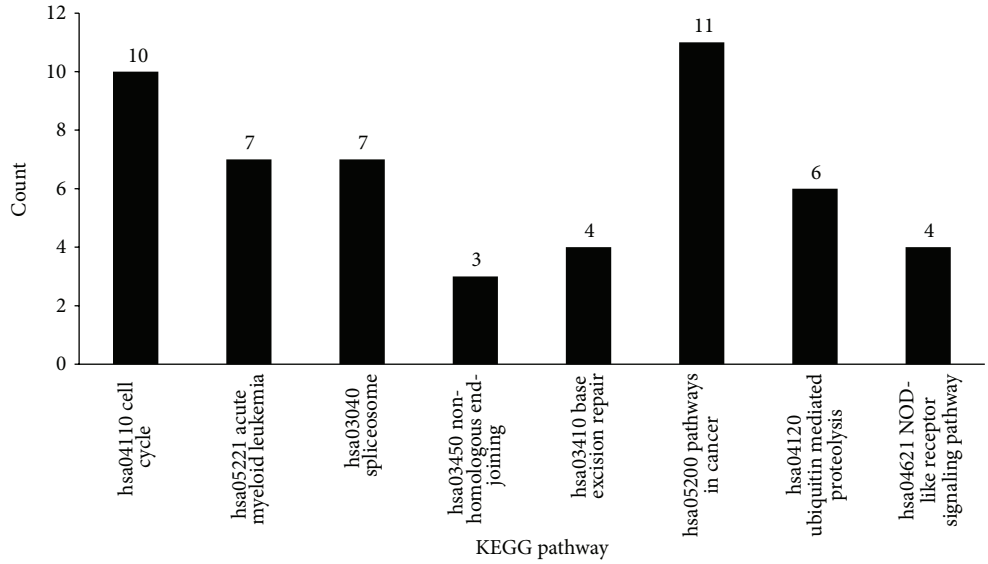


Figure 2: The 8 KEGG pathways enriched by 144 genes. The x-axis lists pathway's ID and name, while the y-axis represents the number of genes that shared the pathway among the 144 genes.

in this pathway. One of the significant characters of gastric cancer is the abnormal activated cell cycle [50]. Among these genes, PCNA is responsible for the DNA synthesis and CCND1 could alter cell cycle by regulating the CDK kinases [51, 52]. Other 2 pathways found in our study are related to the DNA repair which are also very critical for the carcinogenesis. Hsa03450: nonhomologous end joining (NHEJ) ("count" = 3) is a pathway that repairs double-strand breaks in DNA and base excision repair (BER) is a cellular mechanism that repairs damaged DNA throughout the cell cycle [53, 54]. Another intriguing pathway is hsa03040: spliceosome pathway ("count" = 7) which was always abnormal in cancer cells [55]. We speculate that the spliceosome could modify the expression of the oncogenes or tumor-suppress genes which eventually lead to the tumorigenesis. Finally, we also find the hsa05221: acute myeloid leukemia (AML) pathway ("count" = 7) and cancer related pathways in our list. The results imply that the gastric cancer has the common mechanism as well as other cancers especially the AML. Look has reviewed that RUNX1 is the key factor in the hematopoietic development and highly correlated with AML [56]. However, its homologous protein RUNX3 that shares 70% similarity has been reported playing pivotal role in gastric cancer [57]. The finding unravels that cancer normally has the common molecular mechanism as well as the specific pathway with type-dependent pattern. Although several reported pathways

are included in our study, the novel pathways with gastric cancer would expand our views of mechanisms about the tumorigenesis of gastric cancer. On the other hand, we have observed that some genes in these pathways could play a very important role in the carcinogenesis of gastric cancer.

3.3. Analysis of the Relationship of Some Candidate Genes and Gastric Cancer. As described in Section 3.1, 144 genes were discovered by our method. Some of them may have strong relationship with gastric cancer and were discussed as follows. Table 2 listed these genes and their betweenness and permutation FDRs.

Proliferating Cell Nuclear Antigen (PCNA) (see row 2 of Table 2), also known as cyclin, is an auxiliary protein of DNA polymerase-δ that plays important roles both in DNA synthesis and DNA repair [51, 58]. PCNA could act as a homotrimer and helps increase the processivity of leading strand synthesis during DNA replication [59, 60]. In response to DNA damage, this protein is ubiquitinated and is involved in the RAD6-dependent DNA repair pathway [61, 62]. As we know, DNA repair is the main way to remove the carcinogenic lesions caused by UV or other common mutagens [63]. Pascucci et al. have reviewed that the NER (nucleotide excision repair) was highly correlated with skin cancer and intestinal cancer [64]. Intriguingly, numerous works have considered PCNA labeling rate as the prognostic indicator of gastric cancer because its expression was consistent with malignant potential of gastric cancer [65–67]. Ji et al. have found the abnormal increase of PCNAexpression in 58 gastric carcinoma tissues [68]. Similar conclusion was also achieved by Takamura et al. who have performed immunohistochemical study on 164 patients with gastric carcinomas [69]. Although the strong correlation is observed between PCNAand gastric cancer, the detailed mechanism of how PCNA promotes the gastric cancer needs further elucidation.

Besides PCNA, another protein in highly conserved cyclin family was also found in our study. CCND1 (see row 3 of Table 2), with official full name of cyclin D1, was firstly described by Motokura et al. in 1991 [70]. In the following decades, the importance of CCND1 in cell cycle and tumorigenesis was underlined by different labs. Because of the amplification of the 11q13 region where CCND1 locates, CCND1 is frequently overexpressed in human cancers accompanied with abnormalities that are driven by multiple mechanisms including genomic alternations, posttranscriptional regulation and posttranslational protein stabilization [71–73]. On one hand, cyclin D1 could increase CDK activity and consequently result in continuous proliferation which is necessary for tumorigenesis [74, 75]. On the other hand, cyclin D1 may induce the tumorigenesis in certain types of cancers by means of its nuclear receptor-agonistic activity in the CDK-independent way [52, 76].

MYC (see row 4 of Table 2) is a regulator gene that codes for a transcription factor, and it is frequently mutated in many cancers. In Myc-related cancers, Myc is constitutively expressed and leads to the abnormal expression of many genes which may be involved in cell proliferation, differentiation and apoptosis, and these uncontrolled biological processes finally underlie the cancer. Myc is believed to regulate expression of 15% of all genes [77]. Similar with CCND1, Myc expression could be regulated transcriptionally, posttranscriptionally, or posttranslationally [78]. Chung and Levens have reviewed that the deregulated expression of Myc is sufficient to lead to cellular transformation *in vitro* and tumorigenesis *in vivo* [79]. Besides the transforming role, Myc could also promote chromosomal instability by means of its function as a transcriptional regulator [80]. In the previous reports, Myc overexpression has been described in over 40% of gastric cancer [81]. Among nearly half the gastric cancer, copy number gains are frequently detected along chromosome 8 where Myc locates [82, 83]. As the key factor of tumorigenesis, Myc could provide potential target for therapy for gastric cancer [84].

FOS (see row 5 of Table 2), well known as c-fos, encodes a 62 kDa protein, which forms heterodimer with c-jun and subsequently results in the formation of AP-1 complex. FOS has been found to be overexpressed in a variety of cancers. Bakin and Curran have found that c-fos could change DNA methylation pattern by regulating DNMT1 and thereby cause the downregulation of tumor suppressor genes [85]. In addition, c-fos could lead to the loss of cell polarity and EMT which is critical for the metastatic and invasive growth of cancer cells [86]. Hu et al. also found that c-fos is required for the expression of matrix metalloproteinases that are indispensable for invasive growth of cancer cells [87]. However, some recent studies have unraveled the tumor suppressor activity of c-fos, including prohibition of the cell cycle progression, promotion of cell death, or repressing the anchorage-independent growth [88]. In coincidence with the negative role of c-fos in tumorigenesis, Jin et al. analyzed 625 consecutive gastric cancers; 388 cases (62.1%) showed loss of nuclear c-fos expression [89]. Consistent results were concluded by Zhou et al. in 58 patients with gastric cancer [90]. However, Mazurenko et al. reported that high level of c-fos expression was observed in stomach carcinomas [91]. The discordance may be caused by the different stages of progression in different studies. In conclusion, c-fos is a double-edged sword, which could promote or suppress tumorigenesis of gastric cancer.

RUNX1 (see row 6 of Table 2), better known as AML1, plays a critical role in hematopoietic development [92]. RUNX1 belongs to the RUNX family whose 3 members (RUNX1, RUNX2, and RUNX3) share 70% resemblance. Unlike its familial protein RUNX3 that is a strong candidate as a gastric cancer tumor suppressor. RUNX1 is always considered as a tumor suppressor for acute lymphoblastic leukemia (AML) [56]. Usui et al. have examined mRNA expression of all three RUNX genes in the gastric mucosa, and they found that RUNX1 was coexpressed with RUNX3 in pit cells [93]. Sakakura et al. observed remarkable downregulation of RUNX1 and RUNX3 in 9 gastric cancer cell lines and 56 primary gastric cancer specimens [94]. Although RUNX1 is famous for its involving in AML, more lines of evidence shed light to its anticarcinogenesis activity in other carcinoma including gastric cancer.

Other genes found in our study have also been reported relating with gastric cancer. Specific SNPs (Single

Table 2: Important candidate shortest path genes and their betweenness and permutation FDRs.

Ensemble ID of shortest path genes	Gene name	Betweenness	Permutation FDR
ENSP00000368438	PCNA	454	0.083
ENSP00000227507	CCND1	594	0.02
ENSP00000367207	MYC	779	0.01
ENSP00000306245	FOS	318	0.035
ENSP00000300305	RUNX1	224	0.002
ENSP00000262887	XRCC1	152	0.094
ENSP00000352516	DNMT1	169	0.093
ENSP00000379110	CXCL1	107	0.033

Nucleotide Polymorphism) in XRCC1 (X-ray repair cross-complementing 1) (see row 7 of Table 2) are highly associated with gastric cancer [95]. DNMT1 (DNA methyltransferase 1) (see row 8 of Table 2), which is overexpressed in gastric cancer, is associated with increased risks of gastric atrophy with its abnormal polymorphisms [96]. The expression of CXCL1 (chemokine (C-X-C motif) ligand 1) (see row 9 of Table 2) is higher in gastric cancer tissues and endows the cancer cells with more powerful migration and invasion ability [97]. Beyond these genes, more genes associated with gastric tumorigenesis require more evidences for validation or further exploration.

4. Conclusion

Identification of disease genes is one of the most important problems in biomedicine and genomics. For gastric cancer, as one of the leading causes of cancer related deaths worldwide, it is eager to discover its related genes, which can help to uncover its mechanism and design effective treatments. This contribution presented a computational method to identify novel gastric cancer related genes based on known related ones by shortest path algorithm and PPI network. The analysis implies that some genes discovered in this study have direct or indirect relationship with gastric cancer. It is hopeful that this contribution would give a new insight to study this disease and other diseases.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Author's Contribution

Yang Jiang and Yang Shu contributed equally to this work.

Acknowledgments

This paper is supported by Natural Science Fund Projects of Jilin province (201215059), Development of Science and Technology Plan Projects of Jilin province (20100733, 201101074), 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), and SRF for ROCS (Jilin Department of Human Resource & Social Security, 2012–2014).

References

[1] D. M. Parkin, F. Bray, J. Ferlay, and P. Pisani, "Estimating the world cancer burden: globocan 2000," *International Journal of Cancer*, vol. 94, no. 2, pp. 153–156, 2001.

[2] J. Ferlay, H.-R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin, "Estimates of worldwide burden of cancer in 2008: globocan 2008," *International Journal of Cancer*, vol. 127, no. 12, pp. 2893–2917, 2010.

[3] D. Palli, "Epidemiology of gastric cancer: an evaluation of available evidence," *Journal of Gastroenterology*, vol. 35, supplement 12, pp. 84–89, 2000.

[4] E. M. El-Omar, M. Carrington, W.-H. Chow et al., "Interleukin-1 polymorphisms associated with increased risk of gastric cancer," *Nature*, vol. 404, no. 6776, pp. 398–402, 2000.

[5] A. Nomura, J. S. Grove, G. N. Stemmermann, and R. K. Severson, "Cigarette smoking and stomach cancer," *Cancer Research*, vol. 50, article 7084, 1990.

[6] N. Y. Sung, K. S. Choi, E. C. Park et al., "Smoking, alcohol and gastric cancer risk in Korean men: the National Health Insurance Corporation Study," *British Journal of Cancer*, vol. 97, no. 5, pp. 700–704, 2007.

[7] E. Chandanos and J. Lagergren, "Oestrogen and the enigmatic male predominance of gastric cancer," *European Journal of Cancer*, vol. 44, no. 16, pp. 2397–2403, 2008.

[8] P. Lauren, "The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. An attempt at a histo-clinical classification," *Acta Pathologica et Microbiologica Scandinavica*, vol. 64, pp. 31–49, 1965.

[9] C. Caldas, F. Carneiro, H. T. Lynch et al., "Familial gastric cancer: overview and guidelines for management," *Journal of Medical Genetics*, vol. 36, no. 12, pp. 873–880, 1999.

[10] V. Kumar, A. K. Abbas, N. Fausto, and J. C. Aster, *Robbins & Cotran Pathologic Basis of Disease*, Elsevier Health Sciences, Philadelphia, Pa, USA, 2009.

[11] C. Maesawa, G. Tamura, Y. Suzuki et al., "The sequential accumulation of genetic alterations characteristic of the colorectal adenoma-carcinoma sequence does not occur between gastric

adenoma and adenocarcinoma," *The Journal of Pathology*, vol. 176, no. 3, pp. 249–258, 1995.

[12] K.-F. Becker, M. J. Atkinson, U. Reich et al., "E-cadherin gene mutations provide clues to diffuse type gastric carcinomas," *Cancer Research*, vol. 54, no. 14, pp. 3845–3852, 1994.

[13] Q.-L. Li, K. Ito, C. Sakakura et al., "Causal relationship between the loss of *RUNX3* expression and gastric cancer," *Cell*, vol. 109, no. 1, pp. 113–124, 2002.

[14] L. Chen, W.-M. Zeng, Y.-D. Cai, K.-Y. 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.

[15] 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.

[16] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. 12, pp. i246–i254, 2010.

[17] B. M. Padhy and Y. K. Gupta, "Drug repositioning: re-investigating existing drugs for new therapeutic indications," *Journal of Postgraduate Medicine*, vol. 57, no. 2, pp. 153–160, 2011.

[18] L. Chen, J. Lu, X. Luo, and K.-Y. Feng, "Prediction of drug target groups based on chemical-chemical similarities and chemical-chemical/protein connections," *Biochimica et Biophysica Acta*, vol. 1844, no. 1, part B, pp. 207–213, 2014.

[19] L. Chen, J. Lu, N. Zhang, T. Huang, and Y. Cai -D, "A hybrid method for prediction and repositioning of drug Anatomical Therapeutic Chemical classes," *Molecular BioSystems*, 2014.

[20] H.-W. Ma and A.-P. Zeng, "The connectivity structure, giant strong component and centrality of metabolic networks," *Bioinformatics*, vol. 19, no. 11, pp. 1423–1430, 2003.

[21] J. M. Dale, L. Popescu, and P. D. Karp, "Machine learning methods for metabolic pathway prediction," *BMC Bioinformatics*, vol. 11, article 15, 2010.

[22] L. Chen, T. Huang, X.-H. Shi, Y.-D. Cai, and K.-C. Chou, "Analysis of protein pathway networks using hybrid properties," *Molecules*, vol. 15, no. 11, pp. 8177–8192, 2010.

[23] 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.

[24] 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, no. 2, pp. 200–207, 2013.

[25] Y. F. Gao, L. Chen, Y. D. Cai, K. Y. Feng, T. Huang, and Y. Jiang, "Predicting metabolic pathways of small molecules and enzymes based on interaction information of chemicals and proteins," *PLoS ONE*, vol. 7, no. 9, Article ID e45944, 2012.

[26] L. Hu, T. Huang, X. Shi, W.-C. Lu, Y.-D. 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.

[27] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular Systems Biology*, vol. 3, no. 1, article 88, 2007.

[28] 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.

[29] 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.

[30] 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, supplement 1, pp. D412–D416, 2009.

[31] T. H. Gormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Eds., *Introduction to Algorithms*, The MIT Press, Cambridge, Mass, USA, 1990.

[32] J. B. M. Craven, *Markov Networks for Detecting Overlapping Elements in Sequence Data*, The MIT Press, Cambridge, Mass, USA, 2005.

[33] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 233–240, June 2006.

[34] R. Bunescu, R. Ge, R. J. Kate et al., "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial Intelligence in Medicine*, vol. 33, no. 2, pp. 139–155, 2005.

[35] D. E. Johnson and G. H. I. Wolfgang, "Predicting human safety: screening and computational approaches," *Drug Discovery Today*, vol. 5, no. 10, pp. 445–454, 2000.

[36] B.-Q. Li, B. Niu, L. Chen et al., "Identifying chemicals with potential therapy of HIV based on protein-protein and protein-chemical interaction network," *PLoS ONE*, vol. 8, no. 6, Article ID e65207, 2013.

[37] L. Chen, B.-Q. Li, M.-Y. Zheng, J. Zhang, K.-Y. Feng, and Y.-D. Cai, "Prediction of effective drug combinations by chemical interaction, protein interaction and target enrichment of KEGG pathways," *BioMed Research International*, vol. 2013, Article ID 723780, 10 pages, 2013.

[38] J. Zhang, M. Jiang, F. Yuan et al., "Identification of age-related macular degeneration related genes by applying shortest path algorithm in protein-protein interaction network," *BioMed Research International*, vol. 2013, Article ID 523415, 8 pages, 2013.

[39] 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.

[40] M. J. Clark, N. Homer, B. D. O'Connor et al., "U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line," *PLoS Genetics*, vol. 6, no. 1, Article ID e1000832, 2010.

[41] P. Teekakirikul, S. Eminaga, O. Toka et al., "Cardiac fibrosis in mice with hypertrophic cardiomyopathy is mediated by non-myocyte proliferation and requires Tgf-β," *The Journal of Clinical Investigation*, vol. 120, no. 10, pp. 3520–3529, 2010.

[42] C. Chu, K. Qu, F. Zhong, S. Artandi, and H. Chang, "Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions," *Molecular Cell*, vol. 44, no. 4, pp. 667–678, 2011.

[43] B. Kalverda, H. Pickersgill, V. V. Shloma, and M. Fornerod, "Nucleoporins directly stimulate expression of developmental and cell-cycle genes inside the nucleoplasm," *Cell*, vol. 140, no. 3, pp. 360–371, 2010.

[44] Y. Mayshar, U. Ben-David, N. Lavon et al., "Identification and classification of chromosomal aberrations in human induced

pluripotent stem cells," *Cell Stem Cell*, vol. 7, no. 4, pp. 521–531, 2010.

[45] S. J. Sanders, A. G. Ercan-Sencicek, V. Hus et al., "Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism," *Neuron*, vol. 70, no. 5, pp. 863–885, 2011.

[46] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, vol. 29, no. 4, pp. 919–1188, 2001.

[47] P. Hohenberger and S. Gretschel, "Gastric cancer," *The Lancet*, vol. 362, no. 9380, pp. 305–315, 2003.

[48] R. Liu, Z. Li, S. Bai et al., "Mechanism of cancer cell adaptation to metabolic stress: proteomics identification of a novel thyroid hormone-mediated gastric carcinogenenic signaling pathway," *Molecular and Cellular Proteomics*, vol. 8, no. 1, pp. 70–85, 2009.

[49] D. F. Sun, Y. J. Zhang, X. Q. Tian, Y. X. Chen, and J. Y. Fang, "Inhibition of mTOR signalling potentiates the effects of trichostatin A in human gastric cancer cell lines by promoting histone acetylation," *Cell Biology International*, vol. 38, no. 1, pp. 50–63, 2014.

[50] E. Tahara, "Genetic pathways of two types of gastric cancer," *IARC Scientific Publications*, no. 157, pp. 327–349, 2004.

[51] J. Essers, A. F. Theil, C. Baldeyron et al., "Nuclear dynamics of PCNA in DNA replication and repair," *Molecular and Cellular Biology*, vol. 25, no. 21, pp. 9350–9359, 2005.

[52] O. Coqueret, "Linking cyclins to transcriptional control," *Gene*, vol. 299, no. 1-2, pp. 35–55, 2002.

[53] J. K. Moore and J. E. Haber, "Cell cycle and genetic requirements of two pathways of nonhomologous end-Joining repair of double-strand breaks in *Saccharomyces cerevisiae*," *Molecular and Cellular Biology*, vol. 16, no. 5, pp. 2164–2173, 1996.

[54] Y. Liu, R. Prasad, W. A. Beard et al., "Coordination of steps in single-nucleotide base excision repair mediated by apurinic/apyrimidinic endonuclease 1 and DNA polymerase β," *The Journal of Biological Chemistry*, vol. 282, no. 18, pp. 13532–13541, 2007.

[55] C. Naro and C. Sette, "Phosphorylation-mediated regulation of alternative splicing in cancer," *International Journal of Cell Biology*, vol. 2013, Article ID 151839, 15 pages, 2013.

[56] A. T. Look, "Oncogenic transcription factors in the human acute leukemias," *Science*, vol. 278, no. 5340, pp. 1059–1064, 1997.

[57] Q.-L. Li, K. Ito, C. Sakakura et al., "Causal relationship between the loss of *RUNX3* expression and gastric cancer," *Cell*, vol. 109, no. 1, pp. 113–124, 2002.

[58] M. K. K. Shivji, M. K. Kenny, and R. D. Wood, "Proliferating cell nuclear antigen is required for DNA excision repair," *Cell*, vol. 69, no. 2, pp. 367–374, 1992.

[59] G.-L. Moldovan, B. Pfander, and S. Jentsch, "PCNA, the maestro of the replication fork," *Cell*, vol. 129, no. 4, pp. 665–679, 2007.

[60] G. Prelich, C.-K. Tan, and M. Kostura, "Functional identity of proliferating cell nuclear antigen and a DNA polymerase-δ auxiliary protein," *Nature*, vol. 326, no. 6112, pp. 517–520, 1987.

[61] C. Hoege, B. Pfander, G.-L. Moldovan, G. Pyrowolakis, and S. Jentsch, "RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO," *Nature*, vol. 419, no. 6903, pp. 135–141, 2002.

[62] M. K. K. Shivji, "Nucleotide excision repair DNA synthesis by DNA polymerase ε in the presence of PCNA, RFC, and RPA," *Biochemistry*, vol. 34, no. 15, pp. 5011–5017, 1995.

[63] J. de Boer and J. H. J. Hoeijmakers, "Nucleotide excision repair and human syndromes," *Carcinogenesis*, vol. 21, no. 3, pp. 453–460, 2000.

[64] B. Pascucci, M. D'Errico, E. Parlanti, S. Giovannini, and E. Dogliotti, "Role of nucleotide excision repair proteins in oxidative DNA damage repair: an updating," *Biochemistry*, vol. 76, no. 1, pp. 4–15, 2011.

[65] S. Ohyama, Y. Yonemura, and I. Miyazaki, "Proliferative activity and malignancy in human gastric cancers: significance of the proliferation rate and its clinical application," *Cancer*, vol. 69, no. 2, pp. 314–321, 1992.

[66] R. G. Kuang, H. X. Wu, G. X. Hao, J. W. Wang, and C. J. Zhou, "Expression and significance of IGF-2, PCNA, MMP-7, and α-actin in gastric carcinoma with Lauren classification," *The Turkish Journal of Gastroenterology*, vol. 24, no. 2, pp. 99–108, 2013.

[67] S. Jain, M. I. Filipe, P. A. Hall, N. Waseem, D. P. Lane, and D. A. Levison, "Prognostic value of proliferating cell nuclear antigen in gastric carcinoma," *Journal of Clinical Pathology*, vol. 44, no. 8, pp. 655–659, 1991.

[68] J. Ji, P. Zhao, and B. Huang, "Study of gastric carcinoma and PCNA and c-met gene abnormality," *Wei Sheng Yan Jiu*, vol. 37, no. 4, pp. 479–482, 2008.

[69] H. Takamura, Y. Yonemura, L. Fonseca et al., "Correlation of DNA ploidy, c-erB-2 protein tissue status, level of PCNA expression and clinical outcome in gastric carcinomas," *Nihon Geka Gakkai Zasshi*, vol. 96, no. 4, pp. 213–222, 1995.

[70] T. Motokura, T. Bloom, H. G. Kim et al., "A novel cyclin encoded by a bcl1-linked candidate oncogene," *Nature*, vol. 350, no. 6318, pp. 512–515, 1991.

[71] W. Jiang, S. M. Kahn, N. Tomita, Y.-J. Zhang, S.-H. Lu, and I. B. Weinstein, "Amplification and expression of the human cyclin D gene in esophageal cancer," *Cancer Research*, vol. 52, no. 10, pp. 2980–2983, 1992.

[72] M. F. Buckley, K. J. E. Sweeney, J. A. Hamilton et al., "Expression and amplification of cyclin genes in human breast cancer," *Oncogene*, vol. 8, no. 8, pp. 2127–2133, 1993.

[73] J. K. Kim and J. A. Diehl, "Nuclear cyclin D1: an oncogenic driver in human cancer," *Journal of Cellular Physiology*, vol. 220, no. 2, pp. 292–296, 2009.

[74] M. Malumbres, I. P. de Castro, M. I. Hernández, M. Jiménez, T. Corral, and A. Pellicer, "Cellular response to oncogenic ras involves induction of the Cdk4 and Cdk6 inhibitor p15(INK4b)," *Molecular and Cellular Biology*, vol. 20, no. 8, pp. 2915–2925, 2000.

[75] M. Serrano, E. Gomez-Lahoz, R. A. DePinho, D. Beach, and D. Bar-Sagi, "Inhibition of ras-induced proliferation and cellular transformation by p16(INK4)," *Science*, vol. 267, no. 5195, pp. 249–252, 1995.

[76] R. M. L. Zwijsen, E. Wientjens, R. Klompmaker, J. van der Sman, R. Bernards, and R. J. A. M. Michalides, "CDK-independent activation of estrogen receptor by cyclin D1," *Cell*, vol. 88, no. 3, pp. 405–415, 1997.

[77] P. C. Fernandez, S. R. Frank, L. Wang et al., "Genomic targets of the human c-Myc protein," *Genes and Development*, vol. 17, no. 9, pp. 1115–1129, 2003.

[78] R. C. Sears, "The life cycle of c-Myc: from synthesis to degradation," *Cell Cycle*, vol. 3, no. 9, pp. 1133–1137, 2004.

[79] H.-J. Chung and D. Levens, "c-myc expression: keep the noise down!," *Molecules and Cells*, vol. 20, no. 2, pp. 157–166, 2005.

[80] E. V. Prochownik and Y. Li, "The ever expanding role for c-Myc in promoting genomic instability," *Cell Cycle*, vol. 6, no. 9, pp. 1024–1029, 2007.

[81] A. N. Milne, R. Sitarz, R. Carvalho, F. Carneiro, and G. J. A. Offerhaus, "Early onset gastric cancer: on the road to unraveling

gastric carcinogenesis," *Current Molecular Medicine*, vol. 7, no. 1, pp. 15–28, 2007.

[82] C. Sakakura, T. Mori, T. Sakabe et al., "Gains, losses, and amplifications of genomic materials in primary gastric cancers analyzed by comparative genomic hybridization," *Genes Chromosomes Cancer*, vol. 24, no. 4, pp. 299–305, 1999.

[83] S. Yang, H.-C. Jeung, H. J. Jeong et al., "Identification of genes with correlated patterns of variations in DNA copy number and gene expression level in gastric cancer," *Genomics*, vol. 89, no. 4, pp. 451–459, 2007.

[84] M. Vita and M. Henriksson, "The Myc oncoprotein as a therapeutic target for human cancer," *Seminars in Cancer Biology*, vol. 16, no. 4, pp. 318–330, 2006.

[85] A. V. Bakin and T. Curran, "Role of DNA 5-methylcytosine transferase in cell transformation by fos," *Science*, vol. 283, no. 5400, pp. 387–390, 1999.

[86] I. Fialka, H. Schwarz, E. Reichmann, M. Oft, M. Busslinger, and H. Beug, "The estrogen-dependent c-JunER protein causes a reversible loss of mammary epithelial cell polarity involving a destabilization of adherens junctions," *Journal of Cell Biology*, vol. 132, no. 6, pp. 1115–1132, 1996.

[87] E. Hu, E. Mueller, S. Oliviero, V. P. Papaioannou, R. Johnson, and B. M. Spiegelman, "Targeted disruption of the c-fos gene demonstrates c-fos-dependent and -independent pathways for gene expression stimulated by growth factors or oncogenes," *The EMBO Journal*, vol. 13, no. 13, pp. 3094–3103, 1994.

[88] M. Mikula, J. Gotzmann, A. N. M. Fischer et al., "The proto-oncoprotein c-Fos negatively regulates hepatocellular tumorigenesis," *Oncogene*, vol. 22, no. 42, pp. 6725–6738, 2003.

[89] S. P. Jin, J. H. Kim, M. A. Kim et al., "Prognostic significance of loss of c-fos protein in gastric carcinoma," *Pathology and Oncology Research*, vol. 13, no. 4, pp. 284–289, 2007.

[90] L. Zhou, J.-S. Zhang, J.-C. Yu et al., "Negative association of c-fos expression as a favorable prognostic indicator in gastric cancer," *Archives of Medical Research*, vol. 41, no. 3, pp. 201–206, 2010.

[91] N. N. Mazurenko, E. A. Kogan, N. M. Sukhova, and I. B. Zborovskaia, "Synthesis and distribution of oncoproteins in tumor tissue," *Voprosy Meditsinskoj Khimii*, vol. 37, no. 6, pp. 53–59, 1991.

[92] W.-J. Song, M. G. Sullivan, R. D. Legare et al., "Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia," *Nature Genetics*, vol. 23, no. 2, pp. 166–175, 1999.

[93] T. Usui, K. Aoyagi, N. Saeki et al., "Expression status of *RUNX1/AML1* in normal gastric epithelium and its mutational analysis in microdissected gastric cancer cells," *International Journal of Oncology*, vol. 29, no. 4, pp. 779–784, 2006.

[94] C. Sakakura, A. Hagiwara, K. Miyagawa et al., "Frequent down-regulation of the runt domain transcription factors *RUNX1*, *RUNX3* and their cofactor *CBFB* in gastric cancer," *International Journal of Cancer*, vol. 113, no. 2, pp. 221–228, 2005.

[95] H. Xue, P. Ni, B. Lin, H. Xu, and G. Huang, "X-ray repair cross-complementing group 1 (*XRCC1*) genetic polymorphisms and gastric cancer risk: a HuGe review and meta-analysis," *The American Journal of Epidemiology*, vol. 173, no. 4, pp. 363–375, 2011.

[96] J. Jiang, Z. Jia, D. Cao et al., "Polymorphisms of the DNA methyltransferase 1 associated with reduced risks of *Helicobacter pylori* infection and increased risks of gastric atrophy," *PLoS ONE*, vol. 7, no. 9, Article ID e46058, 2012.

[97] W.-L. Cheng, C.-S. Wang, Y.-H. Huang, M.-M. Tsai, Y. Liang, and K.-H. Lin, "Overexpression of CXCL1 and its receptor CXCR2 promote tumor invasion in gastric cancer," *Annals of Oncology*, vol. 22, no. 10, pp. 2267–2276, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 416323, 10 pages
http://dx.doi.org/10.1155/2014/416323

Review Article

Approaches for Recognizing Disease Genes Based on Network

Quan Zou,[1] Jinjin Li,[1] Chunyu Wang,[2] and Xiangxiang Zeng[1]

[1] *School of Information Science and Technology, Xiamen University, Xiamen 361005, China*
[2] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China*

Correspondence should be addressed to Xiangxiang Zeng; xzeng@xmu.edu.cn

Received 6 December 2013; Revised 6 January 2014; Accepted 9 January 2014; Published 24 February 2014

Academic Editor: Tao Huang

Diseases are closely related to genes, thus indicating that genetic abnormalities may lead to certain diseases. The recognition of disease genes has long been a goal in biology, which may contribute to the improvement of health care and understanding gene functions, pathways, and interactions. However, few large-scale gene-gene association datasets, disease-disease association datasets, and gene-disease association datasets are available. A number of machine learning methods have been used to recognize disease genes based on networks. This paper states the relationship between disease and gene, summarizes the approaches used to recognize disease genes based on network, analyzes the core problems and challenges of the methods, and outlooks future research direction.

1. Introduction

Although the human genome project has been accomplished and has achieved great success, and new methods that verify gene function with high-throughput have been applied, studying genetic problems that induce diseases is still one of the major challenges facing humanity [1]. The traditional gene mapping method is based on family genetic disease. First, genes inducing diseases are located in a chain interval. Most recent studies at recognizing disease gene that involves linkage analysis or association studies have resulted in a genomic interval of 0.5 cm to 10 cm, which contains 300 genes [2, 3]. Second, using the biological experiment method to identify each gene located in a chain interval requires a large number of human resources and capital support [4]. In addition, recognizing disease gene by checking the genes set in the interval is often not possible [5]. However, the study of candidate association works well when using a set of known functional candidate genes, which have a clear biological relationship to the disease [6]. Selecting known functional candidate genes is not easy and is often limited by a good deal of factors. The selection of functional candidate genes and prioritization candidate genes has been one of the keys in recognizing disease genes because several reorganization approaches are based on the functions of these genes.

In recent years, a number of recognizing disease gene approaches and computer tools have been developed through building mathematic models based on functional annotation, sequence-based features, protein interaction, and disease phenotype [7–16], such as sequence features [15], functional annotation [7, 8, 10, 13], and physical interactions [12, 13, 17]. Based on the above features, an approach to rank candidate disease genes by computing a correlation score that stands for the correlation between genes and diseases has been introduced. However, various factors may affect the association between genes and diseases.

System biology has indicated that diseases with overlapping clinical manifestations are induced by one or more mutations from the same function module [18–21]. Researches in biological experiments of human disease and patterns have found that genes causing similar disease phenotype often interact with each other directly or indirectly [22–24]. These discoveries have shown that positive correlation exists between disease phenotype and disease gene. Many researchers have proposed disease gene prediction methods based on gene interaction and disease phenotype similarity [7, 25–29]. Recently, many approaches making full use of gene interaction and disease similarity have established the gene interaction and disease phenotype similarity network

to predict disease genes. Some typical methods based on networks will be introduced in detail in this paper.

2. Datasets

In the field of biological information, construction dataset is the data foundation of all subsequent work. The validity of datasets directly affects the validity and reliability of the learning algorithm and test. Thus, building a dataset is a basic and important preparatory work.

The recognition of disease gene datasets is mainly obtained from two databases: Online Mendelian Inheritance in Man (OMIM), which is a synthesis database [30–32], and Human Protein Reference Database (HPRD) [33, 34]. Although none of the datasets from OMIM or HPRD are currently complete, they are comprehensive enough [6].

OMIM has the most abundant information, most extensive resources, most comprehensive, authoritative, and timely human genes and genetic disorders based on knowledge composed to support human genetics research and education and the clinical genetics research. OMIM is daily updated and has free access and acquisition at http://www.omim.org/. In OMIM, each item has a short text summary of a generally determined phenotype or gene and a large number of links to other genetic databases [30]. Datasets of disease phenotype and gene-disease phenotype can be obtained from OMIM. However, the data from OMIM need to be disposed to recognize the disease genes [35].

HPRD is a database which curetted proteomic information suited to human proteins. Even though HPRD is updated relatively slow, it is a full-scale resource for studying the relationship between human diseases and genes [36] and is linked to an outline of human signaling paths. HPRD is also available for free at http://www.hprd.org/ [34]. The dataset of gene interaction can be obtained from HPRD [35].

3. Networks

Most research on recognizing disease genes use networks, including the disease phenotype network, protein-protein interaction network, and gene-disease phenotype network, among others. In this study, we introduce only the most commonly used networks. G_{PPI} represents the gene (proteins) interaction network, G_{DP} represents the disease phenotype network, and $G_{P\text{-}DP}$ represents the gene-disease phenotype network [6, 16].

(1) $G_{PPI} = (G, E_G)$ is an undirected graph and denotes the gene-gene interaction. $G = \{g_1, g_2, \ldots, g_n\}$ is the subset of the gene set, and $E_G \subset G \times G$ expresses the interaction of genes with weight. Figure 1(a) shows G_{PPI}. In the gene-gene interaction network, the relationship between genes is obtained from the gene-gene relationship database, which is one of the most important databases in the biological information field.

(2) $G_{DP} = (D, E_D)$ is also an undirected graph and denotes the disease phenotype network. $D = \{d_1, d_2, \ldots, d_n\}$ is the subset of the disease phenotype set, and $E_D \subset D \times D$ represents the similarity of the disease phenotype with weight. Figure 1(b) shows G_{DP}. In the disease phenotype network, the relationship between disease phenotypes is obtained from the phenotype relationship database, which can also be replaced by the disease-disease relationship database.

(3) $G_{P\text{-}DP} = (G, D, E_{P\text{-}DP})$, which is an undirected biograph, is a gene-disease phenotype network. G is the subset of the gene set, and D is the subset of the disease phenotype set. $E_{P\text{-}DP} \subset G \times D$ expresses the link between the known gene and the disease phenotype. Figure 1(c) stands for $G_{P\text{-}DP}$. The association between disease gene and disease phenotype can be obtained from the gene-disease relationship database.

4. Methods

In previous research, various methods, such as CIPHER, RWRH, Prince, Meta-path, Katz, Catapult, Diffusion Kernel [5], and ProDiGe, were used to recognize disease genes. In the current paper, we introduce several types of typical recognition disease gene methods.

4.1. CIPHER. CIPHER [6] is a tool for predicting and prioritizing disease genes. Furthermore, CIPHER is applied to general genetic phenotypes, which do better in the genome-wide scan of disease genes; furthermore, they are extendable for exploring gene cooperatives in complex diseases. CIPHER is based on an assumption that if two genes have the closest connection in the gene interaction, then the two genes can lead to more similar phenotypes. A regression model can be formulated according to this assumption. A score assessing how likely a gene is associated with a specific phenotype is obtained from the regression model. To construct the regression model, the similarity between phenotypes, interaction between proteins and genes, and list of associations between known disease gene and phenotype must be prepared. The next paragraph expresses the procedures of prioritizing disease genes.

For a given query phenotype and candidate genes, CIPHER first combines the gene interaction network, disease phenotype network, and gene-disease phenotype network into a single network. The similarity scores of the query phenotype with all known phenotypes in the disease phenotype network are derived directly from the phenotype network and the topological distances between the candidate genes. All known disease genes in the gene interaction network are counted and grouped on the basis of their phenotypes. The correlation between phenotypes and disease genes is obtained and acts as the concordance score for each candidate gene by using the regression model. Finally, all candidate genes for the query phenotype are ranked in line with the concordance scores. On account of different neighborhood systems, two ways are available to define the topological distance: direct neighbor and shortest path. Thus, there are two versions of CIPHER which are CIPHER-SP and CIPHER-DN [6].

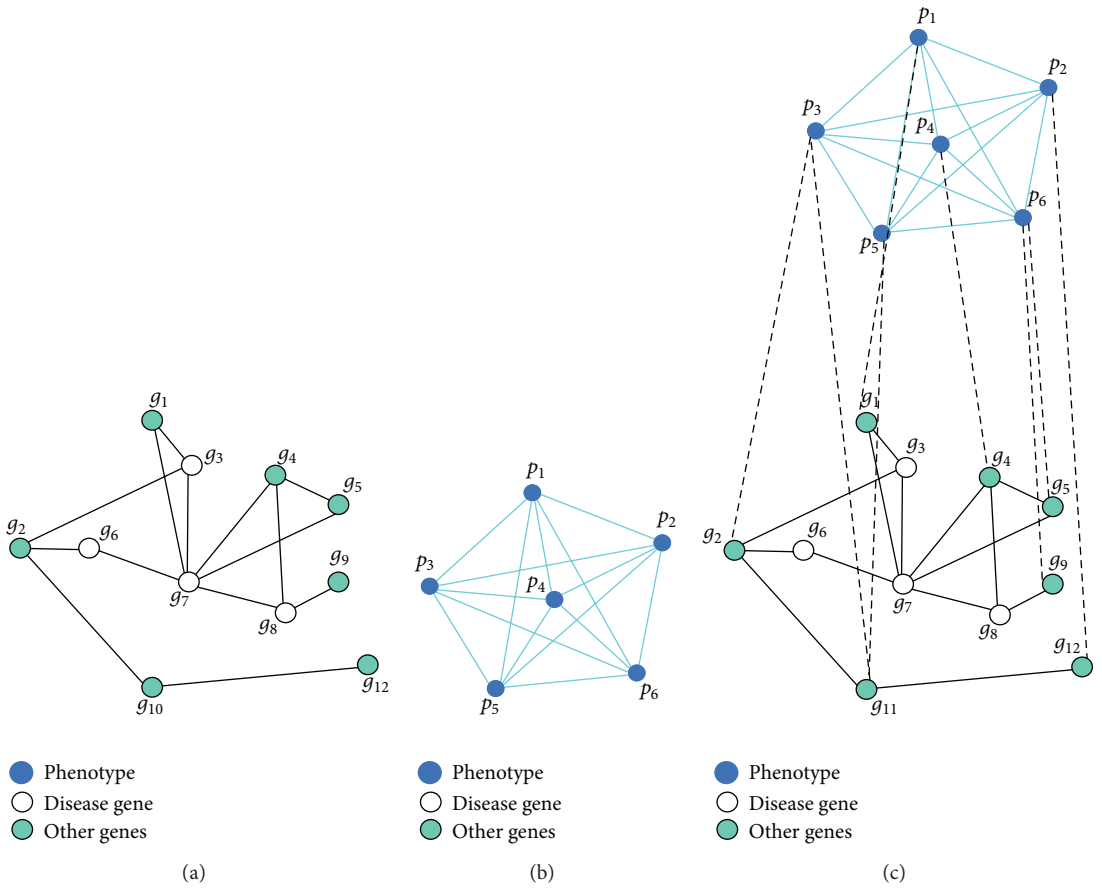


Figure 1: Illustration of the network by using a specific example: (a) is a gene-gene network, (b) is a disease phenotype network, and (c) is a gene-disease phenotype network [6].

The similarity scores of the query phenotype and all phenotypes in the disease phenotype network are calculated by the following formulation:

$$S_{pp'} = C_p + \sum_{g \in G(p)} \sum_{g' \in G(p')} \beta_{pg} e^{-L_{gg'}^2}. \quad (1)$$

In the above formulation, $S_{pp'}$ is the similarity score of the query phenotype and another phenotype in the disease phenotype network. $L_{gg'}$ is the topological distance between the candidate genes and g' in the gene interaction network. There exit two ways to define the topological distance, $L_{gg'}$, on the basis of how to consider indirect the interaction. One way to define the topological distance is shortest path; $L_{gg'}$ is the graph theory shortest path length between genes g and g' in the gene interaction network. The other way to define the topological distance is direct neighbor; $L_{gg'}$ is infinity when g and g' are indirect neighbors. $G(p)$ indicates all disease genes belonging to phenotype p. C_p is a constant and can act as the basal similarity between p and other phenotypes whose causative genes are not connected to those of p in the gene interaction network. β_{pg} is the coefficient of the regression model and stands for the level of gene g contributing to the similarity of the phenotype p to any other phenotype p'. To denote the association between a phenotype and a gene, the following formulation (2) is defined:

$$\Phi_{gp'} = \sum_{g' \in G(p')} e^{-L_{gg'}^2}. \quad (2)$$

The following vector is used to denote the similarities between the query phenotype and all phenotypes in the disease phenotype network: $S_p = (S_{pp_1}, S_{pp_2}, S_{pp_3}, \ldots, S_{pp_n})$.

In the same way, the following vector is used to denote the closeness between the genes and the phenotypes in the disease phenotype network: $\Phi_g = (\Phi_{gp_1}, \Phi_{gp_2}, \Phi_{gp_3}, \ldots, \Phi_{gp_n})$. Synthesizing Formulas (1) and (2) and two vectors extends Formula (1) to the following form:

$$S_p = C_p + \sum_{g \in G(p)} \beta_{pg} \Phi_g. \quad (3)$$

In this regression model, the concordance score is defined by Formula (4):

$$CS_{pg} = \frac{\operatorname{cov}(S_p, \Phi_g)}{\sigma(S_p)\sigma(\Phi_g)}, \quad (4)$$

where cov and σ are the covariance and standard deviation, respectively. The candidate genes for the query phenotype are

ranked according to the values obtained from Formula (4). If a gene that does not connect to any disease genes exists, then Formula (4) cannot be used and the gene will rank at the tail.

4.2. PRINCE. PRINCE is another approach based on networks for ranking candidate disease genes for a given disease and inferring the complex associations between genes. PRINCE is on account of formulating constraints on the ranking function that involved usage of prior information and its smoothness over the network.

Before using PRINCE, the gene disease composed of the phenotype network (set of gene-disease associations), gene-gene interaction network (set of gene-gene association), and at least a query disease phenotype is prepared. $G = (V, E, w)$ denotes the gene-gene interaction network, where V is the set of genes, E is the set of interactions, and w is a weight function denoting the reliability of each interaction. Given a query disease phenotype, PRINCE ranks all the genes in V.

Suppose a gene $v \in V$, the direct neighborhood of gene v is denoted by $N(v)$. The prioritization candidate disease gene function is denoted by $F : V \rightarrow \mathfrak{R}$, and $F(v) = q$ reflects the relevance of v to q. Another function is defined as prior knowledge function denoted by $Y : V \rightarrow \{0, 1\}$. In the prior knowledge function, gene v is related to $q, V(v) = 1$; otherwise, $V(v) = 0$. PRINCE computes function F that is smooth over the network. Thus, function F is a combination of two conditions:

$$F(v) = \alpha \left[\sum_{u \in N(v)} F(u)\, w'(v, u) \right] + (1 - \alpha)\, Y(v), \quad (5)$$

where the parameter $\alpha \in (0, 1)$ weighs the relative importance of gene v to gene u. w' is a normalized form of w. Formally, a diagonal matrix D is defined, and $D(i, i)$ is the sum of row i of W. W is normalized by $W' = D^{-1/2} W D^{-1/2}$, which obtains a symmetric matrix. Here, $W'_{ij} = W_{ij} / \sqrt{D(i, i) D(j, j)}$. Formula (5) can be expressed in linear form as follows:

$$F = \alpha W' F + (1 - \alpha)\, Y \Longleftrightarrow F = \left(I - \alpha W'\right)^{-1} (1 - \alpha)\, Y, \quad (6)$$

where F and V are viewed as vectors of size $|V|$. W' is a matrix whose values are given by w'. Given that the eigenvalues of W' are set in $[-1, 1]$, $\alpha \in (0, 1)$, and the eigenvalues of $(I - \alpha W')$ are positive. In addition, $(I - \alpha W')^{-1}$ exists.

The above linear system can be solved accurately because an iterative propagation-based algorithm works fast and is guaranteed to converge to the system solution for larger networks. Formula (6) is transferred to an iterative algorithm and is denoted as follows:

$$F^t = \alpha W' F^{t-1} + (1 - \alpha)\, Y, \quad (7)$$

where $F^1 = Y$. Every node propagates the information received in the previous iteration to its neighbors. Finally, the values obtained from Formula (7) rank all the candidate disease genes for a query disease phenotype.

4.3. RWRH. Random walk with restart on heterogeneous network (RWRH) is extended from the random walk with restart algorithm to the heterogeneous network. The heterogeneous network is constructed by connecting the gene-gene interaction network and disease phenotype network by using the gene-disease phenotype relationship information. In brief, the gene-disease phenotype network is the heterogeneous network. RWRH prioritizes the genes and the phenotypes simultaneously, which is inspired by the coranking framework [37]. Given a query disease, seed nodes as genes and phenotypes are associated with the disease, and the top ranked phenotype is the most similar to the query disease.

Random walk is defined as an iterative walker transition from its current node to a randomly selected neighbor, starting at a given source node v in the network. However, RWRH allows the restart of the walk in every time step at node v with probability r. P_0 is the probability vector at step 0, indicating that it is the initial probability vector with the sum of probabilities equal to 1. Similarly, P_s is the probability vector at step s, in which the ith element holds the probability of finding the random walker at node i at step s. The probability vector at step $s + 1$ is denoted as follows:

$$P_{s+1} = (1 - \gamma)\, M^T P_s + \gamma P_0, \quad (8)$$

where M is the transition matrix of the heterogeneous network; M_{ij} is the transition probability from node i to node j; $\gamma \in (0, 1)$ is the restart probability in every time step. After several iterations, P_∞ reaches a steady-state that is obtained by performing the iteration until the change between P_s and P_{s+1} falls below 10^{-10}. P_∞ is the measure of closing to seed nodes. In vector P_∞, when $P_\infty(i) > P_\infty(j)$, node i is more likely to be the seed node than node j.

M is the transition matrix of the heterogeneous network. In addition, M consists of four subnetwork transition networks and is denoted as follows:

$$M = \begin{bmatrix} M_G & M_{GP} \\ M_{PG} & M_p \end{bmatrix}, \quad (9)$$

where M_G is the transition matrix of the gene-gene interaction network, which is the intrasubnetwork of the heterogeneous network. M_p is the transition matrix of the disease phenotype network, which is also the intrasubnetwork of the heterogeneous network. M_{PG} and M_{GP} are the intersubnetwork transition matrixes. Supposing the probability of jumping from gene-gene interaction network to the disease phenotype network is λ, the reverse is the same. In the gene-gene interaction network, $\lambda = 0$ if a node is not connected to the phenotype. If a node is directly linked to the disease phenotype network, then the node will jump to the disease phenotype network with probability λ. The node will jump to other nodes in the gene-gene interaction network with probability $1 - \lambda$. Thus, the transition probability from g_i to p_j can be denoted as follows:

$$\left(M_{GP}\right)_{i,j} = P\left(p_j \mid g_i\right) = \begin{cases} \dfrac{\lambda B_{ij}}{\sum_j B_{ij}}, & \text{if } \sum_j B_{ij} \neq 0 \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

In the same way, the transition probability from p_i to g_j can be denoted as follows:

$$(M_{PG})_{i,j} = P(g_j \mid p_i) = \begin{cases} \dfrac{\lambda B_{ji}}{\sum_j B_{ji}}, & \text{if } \sum_j B_{ji} \neq 0 \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

The transition probability from g_i to g_j, which is the element of M_G at the ith row and jth column, can be denoted as follows:

$$(M_G)_{i,j} = \begin{cases} \dfrac{(A_G)_{i,j}}{\sum_j (A_G)_{i,j}}, & \text{if } \sum_j B_{ij} = 0 \\ \dfrac{(1-\lambda)(A_G)_{i,j}}{\sum_j (A_G)_{i,j}}, & \text{otherwise.} \end{cases} \quad (12)$$

The transition probability from p_i to p_j, which is the element of M_P at the ith row and the jth column, can be denoted as follows:

$$(M_P)_{i,j} = \begin{cases} \dfrac{(A_P)_{i,j}}{\sum_j (A_P)_{i,j}}, & \text{if } \sum_j B_{ij} = 0 \\ \dfrac{(1-\lambda)(A_P)_{i,j}}{\sum_j (A_P)_{i,j}}, & \text{otherwise.} \end{cases} \quad (13)$$

In the above four formulations, $A_{G(n\times n)}$, $A_{P(m\times m)}$, and $B_{(n\times m)}$ are the adjacency matrixes for the gene-gene interaction network, disease phenotype network, and gene-disease phenotype network, respectively. The adjacency matrix of the heterogeneous network can be denoted as follows:

$$A = \begin{bmatrix} A_G & B \\ B^T & A_P \end{bmatrix}. \quad (14)$$

The initial probability of the gene-gene interaction network and phenotype network is denoted by u_0 and v_0, respectively. The initial probability of the gene network u_0 makes the equal probabilities to all the seed nodes in the gene network, with the sum of the probabilities equal to 1. The initial probability of the phenotype network v_0 is the same as the gene-gene interaction network. Thus, the initial probability vector of the heterogeneous network is denoted as $P_0 = [(1-\eta)u_\infty \ \ \eta v_\infty]^T$. In the initial probability vector of the heterogeneous network, the parameter $\eta \in (0, 1)$ acts as the judge to weight the importance of each subnetwork. When $\eta = 0.5$, the importance of the gene-gene interaction network and the disease phenotype network are equal. If $\eta > 0.5$, then the importance of the gene-gene interaction network is greater than the disease phenotype network. When $\eta < 0.5$, the gene-gene interaction network is more important than the disease phenotype network is. P_0 and the transition matrix M are substituted into Formula (8). After many iterations, steady-state P_∞ is denoted as $P_\infty = [(1-\eta)u_\infty \ \ \eta v_\infty]^T$. In this way, the steady probabilities u_∞ and v_∞ are used to rank the genes and disease phenotypes. A web server named GeneWanderer, which is a computational method that prioritizes a set of candidate genes according to their probability to become involved in a particular disease or phenotype using HWRH or diffusion kernel, is used.

4.4. Katz. The Katz method is successfully applied to social network link prediction. Predicting the social network link is close to the problem of predicting disease genes. The Katz approach, which is based on a graph, finds the similar nodes for the query nodes in the network [38].

An adjacency matrix A is available in an undirected unweighted graph. The Katz approach counts the number of walks of different lengths that connects i and j. These walks act as the similarity of the two nodes i and j. $(A^l)_{ij}$ is the number of walks of length l that connect i and j. $(A^l)_{ij}$ gives a measure of similarity between i and j. A single similarity measure based on the different walk lengths is necessary. The measure is given below, in which β is a constant that restrains contributions of longer walks:

$$S_{ij} = \sum_{l=1}^{k} \beta_l \left(A^l\right)_{ij}. \quad (15)$$

The above measure is denoted as follows:

$$S = \sum_{l=1}^{k} \beta_l A^l. \quad (16)$$

If $l \to \infty$, $\beta_l \to 0$. In this study, setting $\beta_l = \beta^l$ leads to the well-known Katz method:

$$S^{\text{katz}} = \sum_{l\geq 1} \beta^l A^l = (I - \beta A)^{-1} - I, \quad (17)$$

where β is chosen, such that $\beta < 1/\|A\|^2$. In the case of the Katz method, the connections in the graph are weighed so that A_{ij} is the strength of the connection between nodes i and j. For the choice of k, the sum over infinitely many path lengths is not necessarily considered. According to the experimental results, small values of k ($k = 3$ or $k = 4$) obtain good performance in the task of recommending similar nodes.

The adjacency matrix of the heterogeneous network is denoted as follows:

$$A = \begin{bmatrix} A_G & B \\ B^T & A_P \end{bmatrix}. \quad (18)$$

One of the advantages of Katz is A, which can represent the other species if we want to study human disease phenotypes and other species disease phenotypes.

$$B = \begin{bmatrix} P_{HS} & P_S \end{bmatrix}, \qquad A_P = \begin{bmatrix} A_{PHS} & 0 \\ 0 & A_{PS} \end{bmatrix}. \quad (19)$$

Here, A_{PHS} and A_{PS} represent human phenotypes and the other species phenotypes, respectively. P_{HS} and P_S indicate gene-disease phenotype association of human and other species, respectively. When an experiment on human is conducted, set $P_S = 0$ and $A_{PS} = 0$. By synthesizing expressions (18) and (19), we substitute matrix A into Formula (17) and obtain the similarity of genes and phenotypes from the similarity matrix.

Initialize $\theta = 0$, $s(x) = 0$, $\forall x \in U$, and $n(x) = 1$, $\forall x \in U$
For t $= \mathbf{1}, \mathbf{2}, \mathbf{3}, \ldots, \mathbf{T}$**:**
Step 1. Draw a bootstrap sample $U_t \subseteq U$ of size $n+$
Step 2. Train a linear classifier θ_t using the positive training examples A and U_t as negative examples by solving:

$$\min_{\theta' \in R^d} \frac{1}{2}\left\|\theta'\right\|^2 + C_- \sum_{i \in U_t} \xi_i + C_+ \sum_{i \in A} \xi_i$$

Subject to $\xi_i \geq 0, \quad \forall i \in A \cup U_t$

$$\left\langle \Phi(x_i), \theta' \right\rangle \geq 1 - \xi_i, \quad \forall i \in A$$

$$-\left\langle \Phi(x_i), \theta' \right\rangle \geq 1 - \xi_i, \quad \forall i \in U_t$$

Step 3. For any $x \in U \setminus U_t$ update:
(i) $n(x) \leftarrow n(x) + 1$
(ii) $s(x) \leftarrow s(x) + \left\langle \theta_t, \Phi(x) \right\rangle$
Return $s(x) \leftarrow s(x)/n(x), \quad \forall x \in U$.

Algorithm 1: CATAPULT algorithm description.

4.5. CATAPULT. Combining data across species by using positive-unlabeled learning techniques is abbreviated to CATAPULT. And CATAPULT is a supervised machine learning method which uses a biased support vector machine (SVM), where the features are derived from walks in a heterogeneous gene-trait network.

Given a query disease phenotype, a gene is not associated with the query phenotype. Scholars report positive association between genes and phenotypes; however, the negative associations are rarely reported. In the CATAPULT approach, the unlabeled gene-disease phenotype pairs act as negative associations. The characteristics of the dataset are that only the positive associations are known, and the negative associations and a large number of unlabeled gene-disease phenotype pairs as negative associations are unknown. The general idea of CATAPULT is that the examples are not known to be negative. The false positives are not penalized heavily, but the false negatives are penalized heavily.

CATAPULT uses a biased SVM to classify the gene-phenotype pairs of humans with a single training phase. This approach draws a random bootstrap sample of a few unlabeled examples from the set of all unlabeled examples and trains a classifier to classify the bootstrap samples as negatives along with the positive samples. CATAPULT also uses the bagging technique to obtain an aggregate classifier by using positive and unlabeled examples. The algorithm description is shown in Algorithm 1. T denotes the number of bootstraps, A is the set of positive, n_+ denotes the number of examples in A, U denotes the set of unlabeled gene-disease phenotype pairs, C_- is a penalty for false positives, and C_+ is a relatively larger penalty for false negatives. The source code can be downloaded from http://marcottelab.org/index.php/Catapult.

Before applying any supervised machine learning approach, extracting the features for gene-disease phenotypes is essential. The features are derived from the paths in the heterogeneous network. For a given gene-disease phenotype pair, different walks of the same length and walks of different lengths can be used as features for the gene-disease phenotype pair.

4.6. Meta-Path. The meta-path approach mainly uses the technology of multilabel classification. The multilabel classification method is useful for recognizing disease genes. A gene may exhibit many diseases caused by the gene. In the above example, the gene is an instance, and various diseases are different labels. Given an instance, a large space of all possible label sets may exist, which may be exponential to the number of candidate labels. The frequently used approach to solve the above problem is exploiting correlations among different labels. In the network, exploiting the correlations among different labels denoted by nodes is an advantage.

Meta-path is defined as a sequence of relations in the network. The objects in the network are linked through multiple-type associations. Multiple-type associations help exploit the correlations among different labels for multilabel classification. In recognizing the disease genes, the labels of the genes are diseases, and the labels of the diseases are genes. The explanation of the correlations among genes is summed up in this study.

Given a set of meta-paths among the gene nodes acting as labels, $S_l = \{P_1, P_2, \ldots, P_{cl}\}$, the meta-path-based label correlations can be used as follows:

$$\forall i, \quad P\left(Y_i \mid x_i\right) = \prod_{k=1}^{q} P\left(Y_i^k \mid x_i, Y_i^{P_1(k)}, Y_i^{P_2(k)}, \ldots, Y_i^{P_{cl}(k)}\right), \tag{20}$$

where $P_j(k)$ denotes the index set of the genes linked to the kth gene through the meta-path $P_j \in S_l$. x_i denotes the feature vector of node i in the input space. Y_i denotes the association between a gene and a gene set. The set of all candidate genes is denoted as $V_l = \{l_1, l_2, \ldots, l_q\}$, and Y_i is denoted as $Y_i = (Y_i^1, Y_i^2, \ldots, Y_i^q)^T \in \{0, 1\}^q$. In the same way, given a set of meta-paths among disease phenotype nodes acting as instances, $S_I = \{P_1', \ldots, P_{cI}'\}$, the meta-path-based label correlations can be used as follows:

$$P\left(Y \mid X\right) \approx \prod_i P\left(Y_i \mid x_i, Y_{p_1'(i)}, \ldots, Y_{p_{cI}'(i)}\right), \tag{21}$$

(i) $x_l = LabelPathFeature(l_k, Y_i)$
For each meta-path $P_j' \in S_I$:
(1) Get related labels for node l_k through meta-path P_j.
(2) $x_i = Aggregation(\{Y_i \mid i \in C\})$
Return $(\ldots, x_j^T, \ldots)^T$

(ii) $x_I = LabelPathFeature(i, Y)$
For each meta-path $P_j \in S_I$:
(1) Get related labels for node I_i through meta-path P_j'.
(2) $x_i = Aggregation(\{Y_i \mid i \in C\})$
Return $(\ldots, x_j^T, \ldots)^T$

Algorithm 2: The function of construction relational features for meta-path-based label correlations and meta-path-based instance correlations.

where $P_j'(i)$ denotes the index set of disease phenotypes linked to the ith disease phenotype through meta-path $P_j' \in S_I$.

To perform multilabel collective classification more effectively in heterogeneous information network, both meta-path-based label correlations and meta-path-based instance correlations are performed simultaneously:

$$P(Y \mid X) \approx \prod_i \prod_{k=1}^{q} P\left(Y_i^k \mid x_i, Y_i^{p_j(k)}, \ldots, Y_{P_j'(i)}^k\right), \quad (22)$$

where the gene is set as the label set, and the disease phenotype is set as the instance set. On the contrary, the disease phenotype is set as the label set, and the gene set is as the instance set.

Some research has proposed algorithms based on multilabel collective classification. We briefly introduce the multilabel collective classification algorithm called PIPL. The algorithm roughly includes the following steps.

(1) Meta-path constructions: extracting all nonredundant meta-paths for label correlations and instance correlations.

(2) Training initialization: construction of q extended training sets for all $1 \le k \le q$, $D_k = \{(x_i^k, y_i^k)\}$ by converting each instance x_i to x_i^k by using the functions in Algorithm 2. Training one classifier on each label by using the extended training sets.

(3) Iterative inference: the inference step is an iterative classification algorithm. It updates the testing instance label set predictions and the relational features of label and instance correlations.

5. Evaluation Methods

A comprehensive comparison should be conducted among these methods. In the next several paragraphs, we will introduce some of the key comparisons for recognizing the disease genes reported so far.

Cross-validation is the most frequently used approach in evaluating these methods. However, this method is similar to that used in a previous work, which performs leave-one-out. Each of the known gene-disease phenotype associations is taken as a test case, and a set of genes is assigned as the negative control for each test case. In each round of cross-validation, the disease phenotype is held out, a link between the disease phenotype and one of the associated genes is removed, and the link removed gene is added into the test genes. The rank of the test genes is obtained according to the recognizing methods. Several processing approaches are available for the rank, such as the enrichment score, setting a threshold, precision, recall, and receiver-operating characteristic (ROC).

5.1. Enrichment Score. Suppose the number of test genes is 100. If a recognition disease gene method ranks the actual disease gene as the highest and is sequenced first, then an enrichment of 50-fold exists. The formula of the enrichment score is Enrichment = 50/rank.

5.2. ROC Analysis. ROC analysis denotes the true-positive rate (TPR) versus the false-positive rate (FPR) subject to the threshold dividing the prediction classes. The TPR/FPR is the rate of correctly/incorrectly classified samples of all samples classified to the positive class. To evaluate the scores of disease gene predictions, ROC is explained as a plot of the number of the disease genes above the threshold versus the number of the disease genes below the threshold. The area under the ROC curve for each curve is calculated to compare the different curves obtained by the ROC analysis.

5.3. Setting a Threshold. Concordance score is calculated for each test gene. If the true disease gene ranks first based on the concordance score, then the prediction is successful, and precision is used as the proportion of the successful predictions among all predictions. Another evaluation approach is setting a threshold, in which the highest score of all test genes in this case is not less than the threshold. Thus, *recall* is the fraction of true disease genes predicted among all disease genes.

6. Materials and Results

In the above section, several recognition disease gene methods and evaluation methods have been mentioned. This part introduces the data used and the comparison results.

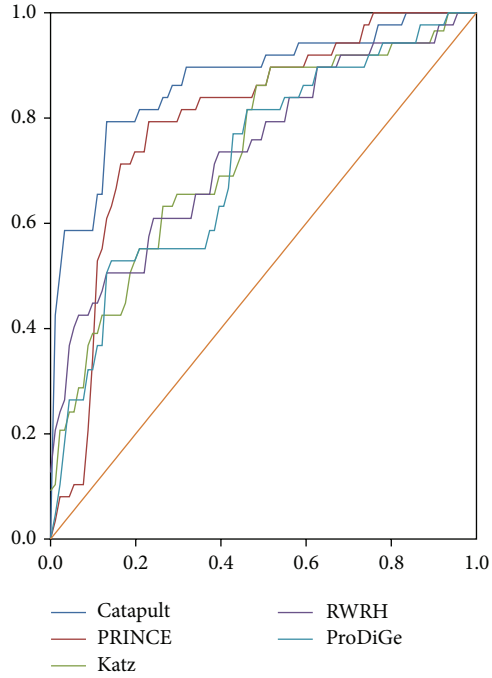


Figure 2: Setting a threshold to compare recognition disease gene methods.

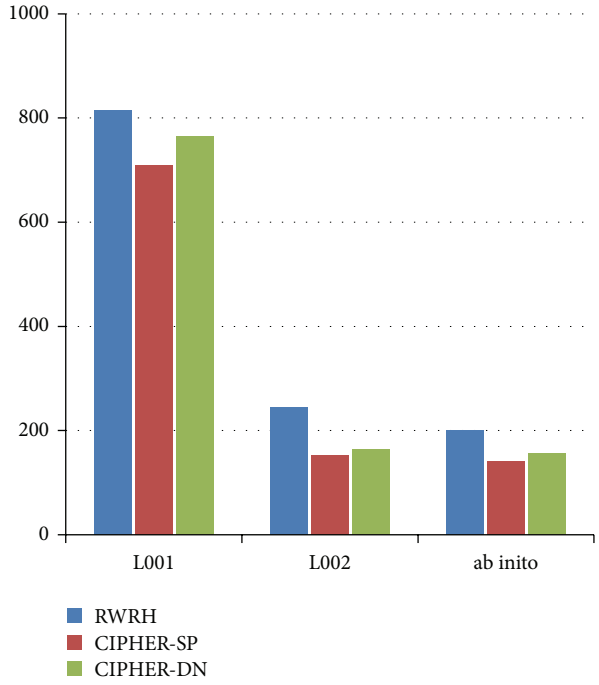


Figure 3: The comparison of different methods.

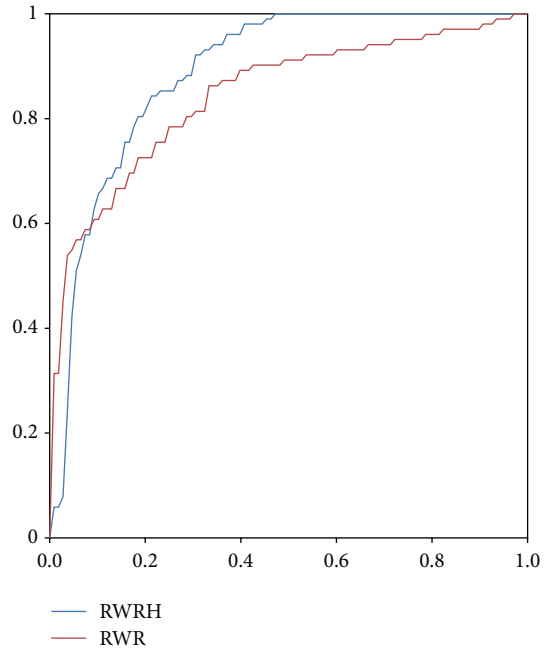


Figure 4: ROC curve of RWR and RWRH.

Figure 2 [17] denotes the comparison of different recognition methods by setting a threshold. In Figure 2, six recognizing methods are shown, including Catapult, Katz, ProDiGe, RWRH, PRINCE, and Degree. HumanNet gene network, which is a part of the OMIM dataset, is the dataset used to compare the different recognition methods. Figure 2 shows that Katz and Catapult do better than the others with the HumanNet gene network and the evaluation method.

RWRH is compared with CIPHER-DN and CIPHER-SP. The evaluation experiment is based on gene network containing 34,364 interactions between 8919 genes, the phenotypic similarity matrix between 5080 phenotype entities calculated by using MimMiner, and 1428 gene-phenotype links between 937 genes and 1216 phenotype entities. The comparison result is denoted by Figure 3 When $\gamma = 0.7$, $\lambda = \eta = 0.5$, RWRH successfully ranks 814 known disease genes as top 1. The result is denoted by L001 in Figure 3. The column of L002 is the result of removing a known gene-phenotype link and using the phenotype and the rest of the disease genes associated with this phenotype as seed nodes. The identification of disease genes for phenotype from the genome is called *ab initio* prediction. The *ab initio* method removes all the links from a phenotype to disease genes and uses the phenotype entity as seed node to run RWRH. If one of the disease genes associated to the phenotype ranks top 1, then the prediction is successful. The result of *ab initio* is shown in Figure 3. From the L001, L002, and *ab initio*, RWRH is better than CIPHER-SP and CIPHER-DN. Figure 4 denotes the result of the comparison between RWR and RWRH. Leave-one-out cross-validation is conducted for each disorder. In each cross-validation, a disease gene is selected, and the links between the phenotype entries and the disease gene are removed. The rest of the disease genes and the phenotype entry are used as seed nodes. The selected disease gene and all disease genes in the artificial linkage are ranked by RWRH and RWR. ROC analysis is used to evaluate the two recognizing approaches.

7. Conclusion

Identifying disease genes is one of the fundaments of medical care and has been a goal in biology. Although traditional linkage analysis and modern high-throughput techniques often provide hundreds of disease gene candidates, identifying disease genes in the candidate genes by using the biological experiment method time-consuming and expensive. To deal with the above issues, the methods based on networks have been proposed. Many methods based on network have been created to recognize disease genes. In this paper, five typical algorithms based on networks, namely, CIPHER, PRINCE, RWRH, Katz, and CATAPULT, are introduced in detail.

Some novel methods have been put forward to recognize and prioritize disease genes. For instance, BRIDGE [39] takes advantage of multiple regression models with penalty to automatically weight different data sources. A researcher employed the ensemble boosting learning technique to combine variant computational approaches for gene prioritization to improve overall performance [40].

Biological relationships are showed by networks, which brings forth new ideas. A network can be used to denote the association between genes and disease to recognize the gene-disease phenotype and to obtain a more complete understanding of the biological system. Networks have been successfully used in biology. However, combining experiments with networks results in the challenge of defining node similarities. Different ways to define node similarity may lead to different effects.

With the development of biology and the emergence of a large number of relevant data, disease gene research based on networks constantly matures. New machine learning methods and technologies will be used to predict disease genes. Research on disease gene recognition will achieve new breakthroughs. The disease gene research will open a new era of medical treatment.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The work was supported by the Natural Science Foundation of China (no. 61370010, no. 61202011, no. 61271346, no. 61172098, and no. 60932008) and the Ph.D. Programs Foundation of Ministry of Education of China (20120121120039).

References

[1] M. D. Adams, J. M. Kelley, J. D. Gocayne et al., "Complementary DNA sequencing: expressed sequence tags and human genome project," *Science*, vol. 252, no. 5013, pp. 1651–1656, 1991.

[2] A. M. Glazier, J. H. Nadeau, and T. J. Aitman, "Genetics: finding genes that underline complex traits," *Science*, vol. 298, no. 5602, pp. 2345–2349, 2002.

[3] D. Botstein and N. Risch, "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease," *Nature Genetics*, vol. 33, pp. 228–237, 2003.

[4] M. Lin and S. Gottschalk, "Collision detection between geometric models: a survey," in *Proceedings of the IMA Conference on the Mathematics of Surfaces*, 1998.

[5] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

[6] 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.

[7] C. Perez-Iratxeta, P. Bork, and M. A. Andrade, "Association of genes to genetically inherited diseases using data mining," *Nature Genetics*, vol. 31, no. 3, pp. 316–319, 2002.

[8] J. Freudenberg and P. Propping, "A similarity-based method for genome-wide prediction of disease-relevant human genes," *Bioinformatics*, vol. 18, supplement 2, pp. S110–S115, 2002.

[9] M. A. van Driel, K. Cuelenaere, P. P. C. W. Kemmeren, J. A. M. Leunissen, and H. G. Brunner, "A new web-based data mining tool for the identification of candidate genes for human genetic disorders," *European Journal of Human Genetics*, vol. 11, no. 1, pp. 57–63, 2003.

[10] F. S. Turner, D. R. Clutterbuck, and C. A. M. Semple, "POCUS: mining genomic sequence annotation to predict disease genes," *Genome Biology*, vol. 4, no. 11, article R75, 2003.

[11] N. Tiffin, J. F. Kelso, A. R. Powell, H. Pan, V. B. Bajic, and W. A. Hide, "Integration of text- and data-mining using ontologies successfully selects disease gene candidates," *Nucleic Acids Research*, vol. 33, no. 5, pp. 1544–1552, 2005.

[12] S. Aerts, D. Lambrechts, S. Maity et al., "Gene prioritization through genomic data fusion," *Nature Biotechnology*, vol. 24, no. 5, pp. 537–544, 2006.

[13] L. Franke, H. Van Bakel, L. Fokkens, E. D. De Jong, M. Egmont-Petersen, and C. Wijmenga, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes," *American Journal of Human Genetics*, vol. 78, no. 6, pp. 1011–1025, 2006.

[14] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard, "SUSPECTS: enabling fast and effective prioritization of positional candidates," *Bioinformatics*, vol. 22, no. 6, pp. 773–774, 2006.

[15] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard, "Speeding disease gene discovery by sequence based candidate prioritization," *BMC Bioinformatics*, vol. 6, no. 1, article 55, 2005.

[16] N. López-Bigas and C. A. Ouzounis, "Genome-wide identification of genes likely to be involved in human genetic disease," *Nucleic Acids Research*, vol. 32, no. 10, pp. 3108–3114, 2004.

[17] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner, "Predicting disease genes using protein-protein interactions," *Journal of Medical Genetics*, vol. 43, no. 8, pp. 691–698, 2006.

[18] P. Jiménez, F. Thomas, and C. Torras, "3D collision detection: a survey," *Computers and Graphics*, vol. 25, no. 2, pp. 269–285, 2001.

[19] B. Grisart, W. Coppieters, F. Farnir et al., "Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition," *Genome Research*, vol. 12, no. 2, pp. 222–231, 2002.

[20] G. Thaller, C. Kühn, A. Winter et al., "DGAT1, a new positional and functional candidate gene for intramuscular fat deposition in cattle," *Animal Genetics*, vol. 34, no. 5, pp. 354–357, 2003.

[21] A. Clop, F. Marcq, H. Takeda et al., "A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep," *Nature Genetics*, vol. 38, no. 7, pp. 813–818, 2006.

[22] M. Oti and H. G. Brunner, "The modular nature of genetic diseases," *Clinical Genetics*, vol. 71, no. 1, pp. 1–11, 2007.

[23] L. D. Wood, D. W. Parsons, S. Jones et al., "The genomic landscapes of human breast and colorectal cancers," *Science Signaling*, vol. 318, no. 5853, pp. 1108–1113, 2007.

[24] J. Lim, T. Hao, C. Shaw et al., "A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration," *Cell*, vol. 125, no. 4, pp. 801–814, 2006.

[25] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. M. Leunissen, "A text-mining analysis of the human phenome," *European Journal of Human Genetics*, vol. 14, no. 5, pp. 535–542, 2006.

[26] S. Li, L. Wu, and Z. Zhang, "Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach," *Bioinformatics*, vol. 22, no. 17, pp. 2143–2150, 2006.

[27] K. J. Gaulton, K. L. Mohlke, and T. J. Vision, "A computational system to select candidate genes for complex human traits," *Bioinformatics*, vol. 23, no. 9, pp. 1132–1140, 2007.

[28] V. van Heyningen and P. L. Yeyati, "Mechanisms of non-Mendelian inheritance in genetic disease," *Human Molecular Genetics*, vol. 13, supplement 2, pp. R225–R233, 2004.

[29] 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.

[30] 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, supplement 1, pp. D514–D517, 2005.

[31] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, "Onlined Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 30, no. 1, pp. 52–55, 2002.

[32] A. Hamosh, A. F. Scott, J. Amberger, D. Valle, and V. A. McKusick, "Online Mendelian inheritance in man (OMIM)," *Human Mutation*, vol. 15, no. 1, pp. 57–61, 2000.

[33] L. Baolin and H. Bo, "HPRD: a high performance RDF database," in *Network and Parallel Computing*, vol. 4672 of *Lecture Notes in Computer Science*, pp. 364–374, Springer, Berlin, Germany, 2007.

[34] T. S. Keshava Prasad, R. Goel, K. Kandasamy et al., "Human protein reference database—2009 update," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D767–D772, 2009.

[35] Y. Li and J. C. Patra, "Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network," *Bioinformatics*, vol. 26, no. 9, Article ID btq108, pp. 1219–1224, 2010.

[36] S. Peri, J. D. Navarro, T. Z. Kristiansen et al., "Human protein reference database as a discovery resource for proteomics," *Nucleic Acids Research*, vol. 32, supplement 1, pp. D497–D501, 2004.

[37] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles, "Co-ranking authors and documents in a heterogeneous network," in *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM '07)*, pp. 739–744, Omaha, Neb, USA, October 2007.

[38] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[39] Y. Chen, X. Wu, and R. Jiang, "Integrating human omics data to prioritize candidate genes," *BMC Medical Genomics*, vol. 6, no. 1, article 57, 2013.

[40] P. F. Lee and V. W. Soo, "An ensemble rank learning approach for gene prioritization," in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '13)*, pp. 3507–3510, Osaka, Japan, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 818670, 8 pages
http://dx.doi.org/10.1155/2014/818670

Research Article

Predicting Glycerophosphoinositol Identities in Lipidomic Datasets Using VaLID (Visualization and Phospholipid Identification)—An Online Bioinformatic Search Engine

Graeme S. V. McDowell,[1,2,3] Alexandre P. Blanchard,[1,2,3] Graeme P. Taylor,[1,2] Daniel Figeys,[2,3] Stephen Fai,[3,4] and Steffany A. L. Bennett[1,2,3]

[1] *Neural Regeneration Laboratory, Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, ON, Canada K1H 8M5*
[2] *Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, ON, Canada K1H 8M5*
[3] *CIHR Training Program in Neurodegenerative Lipidomics, Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, ON, Canada K1H 8M5*
[4] *Carleton Immersive Media Studio, Azrieli School of Architecture and Urbanism, Carleton University, ON, Canada K1S 5B6*

Correspondence should be addressed to Steffany A. L. Bennett; sbennet@uottawa.ca

Received 6 November 2013; Accepted 23 December 2013; Published 20 February 2014

Academic Editor: Tao Huang

The capacity to predict and visualize all theoretically possible glycerophospholipid molecular identities present in lipidomic datasets is currently limited. To address this issue, we expanded the search-engine and compositional databases of the online Visualization and Phospholipid Identification (VaLID) bioinformatic tool to include the glycerophosphoinositol superfamily. VaLID v1.0.0 originally allowed exact and average mass libraries of 736,584 individual species from eight phospholipid classes: glycerophosphates, glyceropyrophosphates, glycerophosphocholines, glycerophosphoethanolamines, glycerophosphoglycerols, glycerophosphoglycerophosphates, glycerophosphoserines, and cytidine 5′-diphosphate 1,2-diacyl-sn-glycerols to be searched for any mass to charge value (with adjustable tolerance levels) under a variety of mass spectrometry conditions. Here, we describe an update that now includes all possible glycerophosphoinositols, glycerophosphoinositol monophosphates, glycerophosphoinositol bisphosphates, and glycerophosphoinositol trisphosphates. This update expands the total number of lipid species represented in the VaLID v2.0.0 database to 1,473,168 phospholipids. Each phospholipid can be generated in skeletal representation. A subset of species curated by the Canadian Institutes of Health Research Training Program in Neurodegenerative Lipidomics (CTPNL) team is provided as an array of high-resolution structures. VaLID is freely available and responds to all users through the CTPNL resources web site.

1. Introduction

The emerging field of lipidomics seeks to answer two seemingly simple questions: How many lipid species are there? What effect does lipid diversity have on cellular function? To address these questions, lipidomics requires a comprehensive assessment of cellular, regional, and systemic lipid homeostasis. This assessment expands beyond lipid profiling to include the transcriptomes and proteomes of lipid metabolic enzymes and transporters, as well as that of the protein targets that affect downstream lipid signalling [1]. Lipidomic analyses also encompass an unbiased mechanistic assessment of lipid function ranging from the physicochemical basis of lipid behaviour to lipid-protein and lipid-lipid interactions triggered by intrinsic and extrinsic stimuli [1]. The first step, however, lies in identifying the molecular identities of the lipid constituents in different membrane compartments.

Recent technological advances in electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) mass spectrometry (MS), coupled to high performance liquid chromatography (LC), allow lipid diversity and membrane composition to be quantified at the molecular level [4–7]. Thousands of unique lipid species across the six major lipid structural categories in mammalian cells (fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, and prenol lipids) and two lipid categories synthesized by other organisms (saccharolipids and polyketides) can now be identified using LC-ESI-MS and, in some cases, MALDI-MS imaging [1, 4, 8]. Yet, with these successes come new challenges. Turning raw MS spectral data into annotated lipidomic datasets is a time-consuming, labour-intensive, and highly inefficient process. Predicting identities of "new" species, not previously curated, is exceedingly difficult. Lipidomic investigations lack essential bioinformatic tools capable of enabling automated data processing and exploiting the rich compositional data present in MS lipid spectra.

The critical first step is to unambiguously assign molecular identities from the MS structural information present in large lipidomic datasets [9]. Where genomics and proteomics capitalize on sequence-based signatures, lipids lack such easily definable molecular fingerprints. Identities must be reconstructed by analysis of (a) lipid mass to charge (m/z) ratios following "soft" ionization ESI and MALDI techniques and (b) defining fragmentation patterns obtained after collision-induced dissociation in various MS modes [7]. Once these molecular identities are predicted, further information about stereospecificity of critical species can then be assessed (e.g., by tandem MS, analysis of *lyso*-form fragment ions, and product ion spectral evaluation) [10–13]. For example, membrane phospholipids are derivatives of *sn*-glycero-3-phosphate with (a) an acyl, an alkyl (ether-linked plasmanyl), or an alkenyl (alkyl-1′-enyl, vinyl ether-linked plasmenyl) carbon chain at the *sn*-1 position; (b) a long-chain fatty acid that is usually esterified to the *sn*-2 position; and (c) a polar headgroup composed of a nitrogenous base, a glycerol, or an inositol unit modifying the phosphate group at the *sn*-3 position. The polar head group defines membership in one of 20 different phospholipid classes (e.g., glycerophosphoserines (PS), glycerophosphoethanolamines (PE), glycerophosphocholines (PC), glycerophosphoinositols (PI), etc.) [14]. Molecular species are further distinguished by individual combinations of carbon residues (chain length and degree of unsaturation) and the nature of each *sn*-1 or *sn*-2 chemical linkage (acyl, alkyl, or alkenyl) to the glycerol backbone. PI(18:0/22:6), for example, defines a lipid with a phosphoinositol polar head group (PI), a fully saturated 18 carbon chain (referred to as :0) ester-linked at the *sn*-1 position, and a 22 carbon chain which is characterized by six unsaturations (indicated by :6) ester-linked at the *sn*-2 position (Figure 1). Immediate PI metabolites (PIP_x) are then produced by carbon-specific phosphorylation of the PI headgroup with unique fatty acyl, alkyl, and/or alkenyl *sn*-1 and *sn*-2 chains (Figures 1 and 2). The tight regulation of PI metabolism and its critical impact on cellular function clearly underlines the importance of these compositional changes (Figure 1). Yet, to date, biological significance of the astonishing number of potentially unique PIs and PIP_xs is unknown. This is primarily due to the challenges associated with unambiguous compositional identification of PIs and PIP_xs in biological membranes [1, 15–19].

Key advances in lipidomic bioinformatics have been led by the LIPID MAPS consortium both in the development of online spectral databases and the reorganization of lipid class ontologies [14, 20]. These toolsets and classification systems have recently been complemented by the *in silico* generation of a searchable library of all theoretically possible MS/MS lipid spectra in different ionization modes (LipidBlast) [21]. Such fundamental toolkits are supported by a growing compendium of targeted spectral tools, reviewed in [6, 7, 20, 22]. Few existing bioinformatic resources, however, provide necessary information on all potential acyl chain inversions (e.g., *sn*-1 versus *sn*-2), critical phospholipid linkages that define lipid function, or theoretically possible double bond positions for every possible species. To address this need, we have developed Visualization and Phospholipid Identification (VaLID)—a web-based application linking a user-friendly online search engine, structural composition database, and multiple visualization features—that is capable of providing users with all theoretically possible phospholipids calculated from any m/z under a variety of MS conditions. VaLID version 1.0.0 was initially restricted to 736,584 unique PS, PE, PC, glycerophosphate (PA), glyceropyrophosphate (PPA), glycerophosphoglycerol (PG), glycerophosphoglycerophosphate (PGP), and cytidine 5′-diphosphate 1,2-diacyl-*sn*-glycerol (CDP-DG) identities (Table 1) [22]. At first release, we did not include the PI family or their bioactive PIP_x metabolites given the significant challenges associated with automating the visualization of all theoretically possible combinations of *sn*-1 and *sn*-2 carbon chain lengths, linkages, and variations in phosphorylation of the phosphoinositol head group. Here, we address this deficit through the development of VaLID version 2.0.0, now coded with an exhaustive PI and PIP_x database, capable of computing and visualizing a total of 1,473,168 theoretically possible phospholipids predicted from any user-inputted m/z value and MS condition. VaLID version 2.0.0 is freely available for commercial and noncommercial use at http://neurolipidomics.ca and http://neurolipidomics.com/resources.html.

2. Materials and Methods

2.1. Programming Language and Packages. VaLID version 2.0.0. was developed using Oracle's Java programming language version 6 and external Java libraries from JExcelApi and structures are displayed within the program by ChemAxon's Marvin View 5.5.1.0. software. The code was written using the IDE Eclipse Kepler, and packaged using the Fat Jar Eclipse version 0.0.31 plugin. VaLID is a web-based Java applet, and thus it requires that Java be both installed and enabled on a user's web browser. The most recent Java security update is recommended, and can be downloaded from http://www.oracle.com/technetwork/java/index.html.

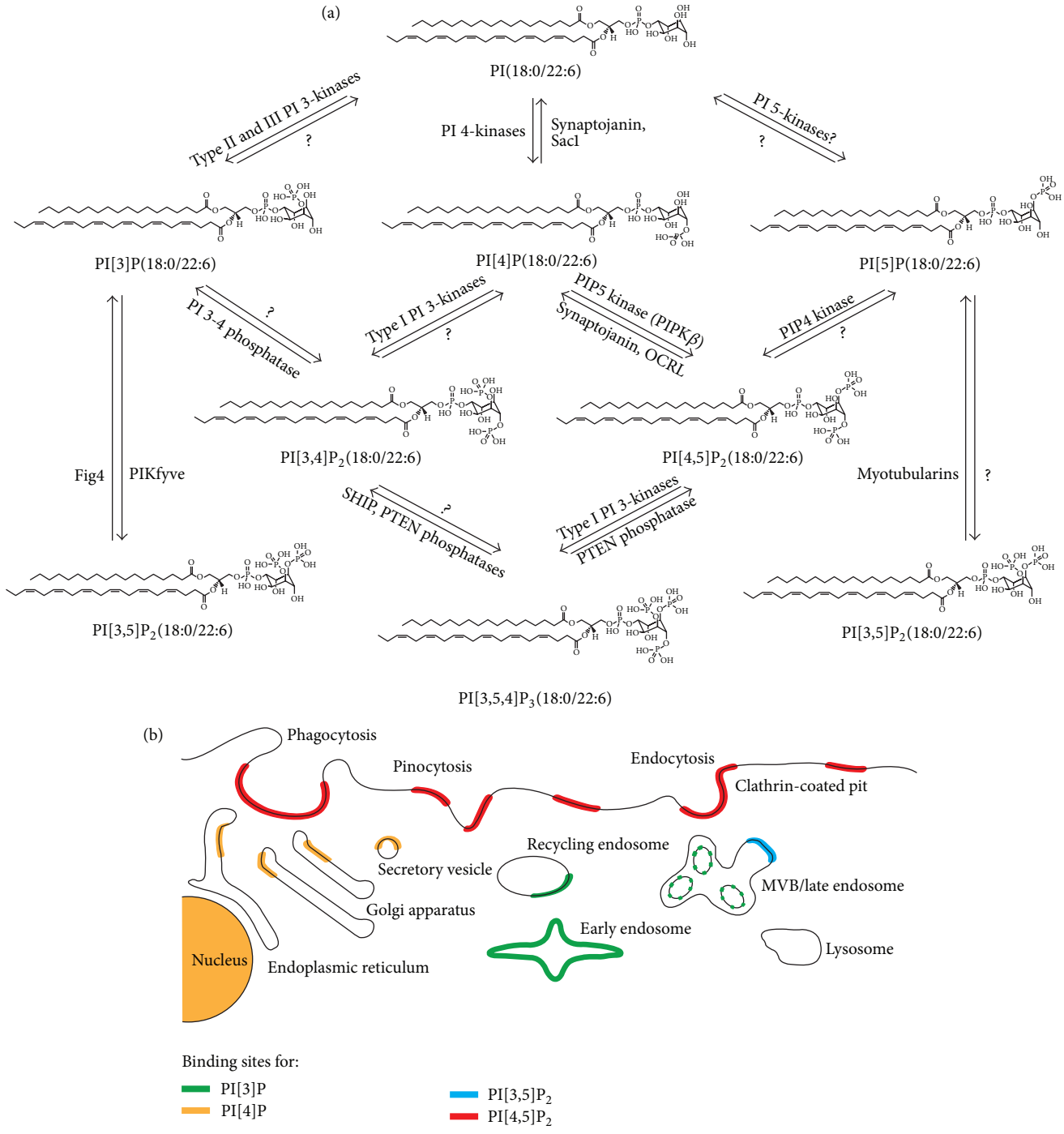


Figure 1: Glycerophosphoinositol (PI) metabolism to PI phosphates (PIP_x). (a) Metabolism of membrane PIs to bioactive PIP_x second messengers. The molecular identity of each species, defined by carbon chain length and linkage to the glycerophospholipid backbone, is predicted to affect signalling specificity in addition to known effects of PI headgroup phosphorylation. (b) Phosphorylation of PIP_x species regulates the localization of different PI-binding proteins and targets them to specific organelles (i.e., lipid-protein interaction). Phosphorylation status and carbon chain length dictate localization and likely restrict functions. Together, structural PIs and their PIP_x second messengers regulate vesicular fusion, exocytosis, and endocytosis as reviewed in (and adapted from) [2, 3].

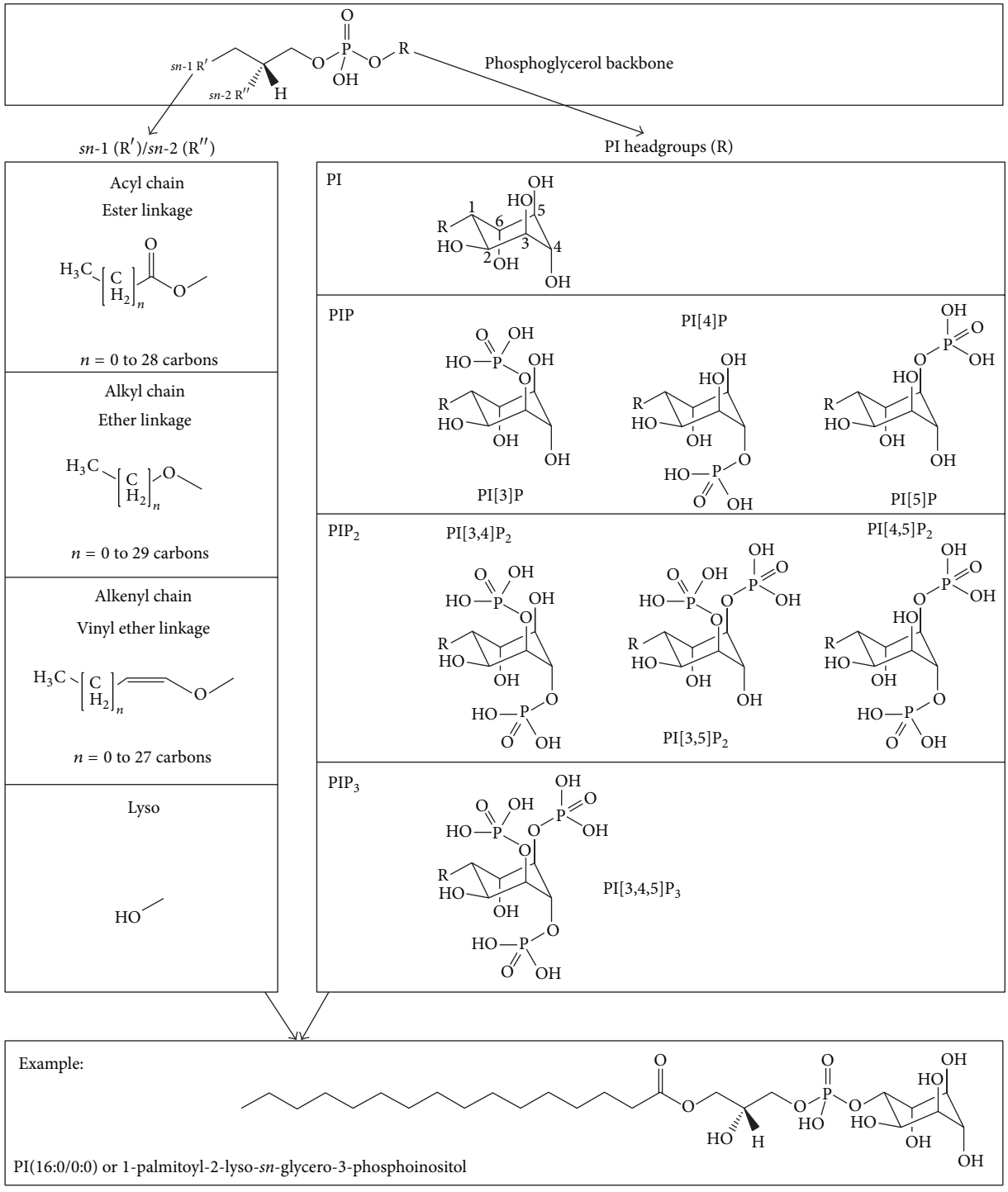


Figure 2: Component and composite structural PI and PIP_x features used to calculate masses. Exact and average masses for all theoretically possible PI and PIP_x species were calculated from the masses of every component possibility: (top panel) the phosphoglycerol backbone, (left panel) *sn*-1 and *sn*-2 hydroxyl residues (lyso-lipids) and *sn-1* and *sn-2* fatty chains ranging from 0 to 30 carbons with up to six unsaturations, considering ester, ether, or vinyl ether linkages to the phosphoglycerol backbone, and (right panel) PI polar headgroups and all biologically relevant phosphorylation possibilities. The bottom panel provides one composite PI example.

Table 1: Total number of species from each subclass that is included in VaLID.

Phospholipid subclass	LIPIDMAPS classification	Abbreviation	Number of species*
Glycerophosphates	GP10	PA	92073
Glyceropyrophosphates	GP11	PPA	92073
Glycerophosphocholines	GP01	PC	92073
Glycerophosphoethanolamines	GP02	PE	92073
Glycerophosphoglycerols	GP04	PG	92073
Glycerophosphoglycerophosphates	GP05	PGP	92073
Glycerophosphoinositols	GP06-09	PI, PIP_x	736584
Glycerophosphoserines	GP03	PS	92073
Cytidine 5′-diphosphate glycerols	GP13	CDP-DG	92073
		Total	1473168

*The calculated number of species does not include lipids formed by changing the position of the double bond beyond those represented in VaLID's structural models. Each lipid *m/z* has been calculated for exact and average masses and can be searched using even and odd carbon chains with mass tolerance ranging from ±0.0001 to ±2 and MS ion modes $[M + H]^+$, $[M + K]^+$, $[M + Li]^+$, $[M + Na]^+$, $[M - H]^-$, or [M (Neutral)].

2.2. The PI and PIP_x Compositional Database. Briefly, the underlying database contains masses of all theoretically possible PI and PIP_x species calculated from both exact and average atomic masses [23]. Component structural masses were first established for: (a) the glycerol backbone, (b) PI polar headgroups with all phosphorylation possibilities, (c) *sn*-1 and *sn*-2 hydroxyl residues (lyso-lipids), (d) *sn*-1 and *sn*-2 fatty chains ranging from 0 to 30 carbons with up to six unsaturations, considering (e) ester, ether, or vinyl ether linkages to the phosphoglyceride backbone (Figure 2). Composite masses were then calculated for every theoretically possible combination. Thus, the underlying database includes all PIs, as well as every acyl, alkyl, and alkenyl variant, for every carbon chain and double bond position, of all mono- (PIP), bis- (PIP_2), and tris- (PIP_3) phosphorylated PI headgroups modified on the hydroxyl group of carbons 3, 4, and/or 5.

2.3. PI and PIP_x Structural Visualizations. We have updated the automated representation drawing feature of VaLID to be able to draw all theoretically possible PI and PIP_x molecular identities. Structures have been restricted to display only cis double bonds separated by a minimum of two carbons. To achieve this goal, the basic structure of the PI backbone was created manually and the atom placement corrected mathematically to match known structures. Slight adjustments to atom placement were further made to improve visibility. The locations of each atom in the headgroup were then established on a Cartesian plane and coded into the software. The automated drawing feature update was integrated into the database and search functions, allowing all PI and PIP_x to be visualized on demand. Chemical structures are displayed using ChemAxon's MarvinView software (Marvin 5.5.1.0, 2011, http://www.chemaxon.com).

3. Results and Discussion

PI and PIP_x are derivatives of *sn*-glycero-3-phosphate with (a) an acyl, an alkyl (ether-linked plasmanyl), or an alkenyl (alkyl-1′-enyl, vinyl ether-linked plasmenyl) carbon chain; (b) a fatty acid commonly esterified but also with possible alkyl or alkenyl linkages to the *sn*-2 position; and (c) a polar headgroup composed of an inositol unit modifying the phosphate group at the *sn*-3 position. Individual species are distinguished by their particular combination of carbon chains (chain length and degree of unsaturation) and by the nature of their *sn*-1 or *sn*-2 chemical linkages (acyl, alkyl, or alkenyl). PI[3, 4, 5]P_3(*O*-16:0/20:4), for example, defines a lipid species with a phosphoinositol polar head group (PI) phosphorylated at the 3rd, 4th, and 5th carbon positions, an ether linkage at the *sn*-1 position (*O*-), 16 carbons at the *sn*-1, and 20 carbons at the *sn*-2 positions, of which the *sn*-1 chain is fully saturated. The number of possible structural and biochemical combinations results in colossal structural diversity; however, PIP_x lipids account for less than 15 percent of the total phospholipid composition in eukaryote cells [24]. The molecular identities of these critical species have yet to be determined in different lipidomes despite emerging evidence that differences in carbon chain length, linkage, and phosphorylation status fundamentally alter biological activity [1, 15–19] (Figure 1).

Here, we enhanced VaLID's capacity to (a) predict identities of glycerophosphoinositol species present in MS spectra from *m/z* under user-defined MS conditions and (b) automatically visualize every theoretically possible PI molecular species at given *m/z*. The updated VaLID interface, showing all of the available search terms, is presented in Figure 3. Since its inception, VaLID was designed to be a comprehensive glycerophospholipid database linking a convenient search engine with visualization features for identification and dissemination of large-scale lipidomic datasets. The intent of this tool was to aid in lipid discovery obtained through multiple MS methodologies and significantly reduce the time required to validate critical phospholipid identities present in target lipidomes. The program initially contained eight phospholipid subclasses, excluding the PI subfamily. In VaLID version 2.0.0, this capacity is now expanded to all theoretically possible PI and PIP_x glycerophospholipids and comprises a total of 1,473,168 unique structures.

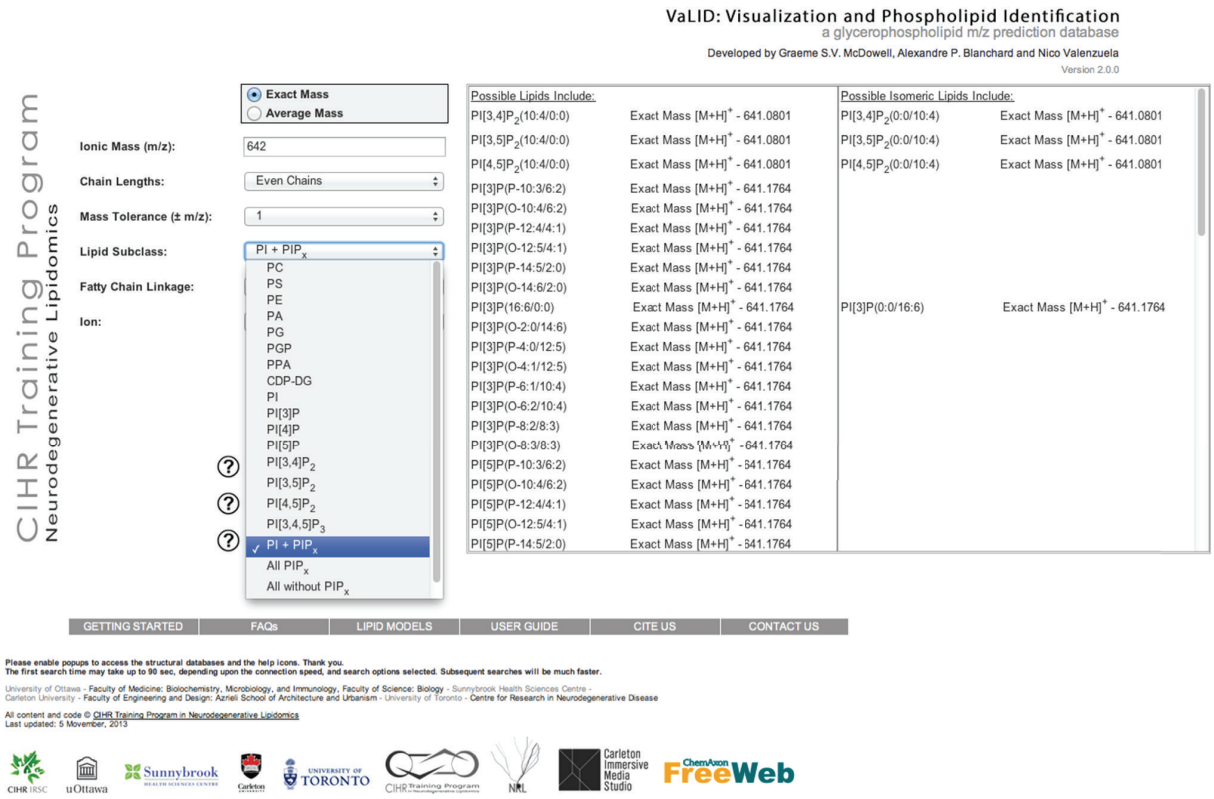


Figure 3: VaLID 2.0.0 interface. The new PI and PIP_x search options are shown for the VaLID interface's drop down menu. Each member of the PI family can be searched individually, as well as in various combinations. PIP_x refers to phosphoinositol mono-, bis-, or tris-phosphate, and can be searched with, or without, PIs.

These additions are meant to provide lipidomic researchers with the additional tools necessary to mine their lipidomes for PI and PIP_x species with specific *m/z* under their particular MS experimental conditions including the ion mode and the lipid subclass. Due to the complexity of the PI superfamily, and to accelerate searching, users can restrict searches to subclasses (PI, PIP_x) or sub-subclasses (PI, PI[3]P, PI[4]P, PI[5]P, PI[3, 4]P_2, PI[3, 5]P_2, PI[4, 5]P_2, PI[3, 4, 5]P_3). For example, if the option PI[3, 4]P_2 is chosen, all molecular species with an inositol backbone phosphorylated only at the 3rd and 4th carbon positions will be provided and VaLID will not return any related PI[3, 5]P_2 or PI[4, 5]P_2 species. The PI +PIP_x option restricts searches to the entire PI superfamily excluding other phospholipid families. The "All without the PIP_x" option returns all of the phospholipids in the database including PI structural precursors with the exception of PIP_x metabolites. Finally, the "All" option returns results from every headgroup. When more than one headgroup is being searched, the program will let the user know how many headgroups have been loaded, and how many are remaining to be loaded.

With respect to the visualization features for PIP or PIP_2, the program will draw the phosphate groups on the inositol ring in the locations that the user specified from the dropdown menu for lipid species selected. As with the other subclasses, choosing the "Display All" button will draw all the theoretically possible structures associated with the selected lipid name. Potential variants in degrees of unsaturation are drawn sequentially in every location along the fatty acid chain, separated by at least two carbons, and in cis configuration. An example of this can be seen in Figure 4. If the selected lipid meets criteria for the "Best Prediction," selecting this option will return only the lipids in VaLID's "Predicted to be Common" database. These species are categorized based on the relative abundance of prevalent fatty acid chains in mammalian cells [25].

4. Conclusions

VaLID is, to our knowledge, the first search engine that has an exhaustive *m/z* and visualization database of all the theoretically possible glycerophospholipids updated here from eight to twelve of the twenty phospholipid subclasses defined by the LIPID MAPS Consortium [26]. The purpose of this update is to facilitate prediction and visualization of the identities of all unknown species, now including all PIs and their metabolites, with given *m/z* and MS condition that may be present in users' lipidomes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

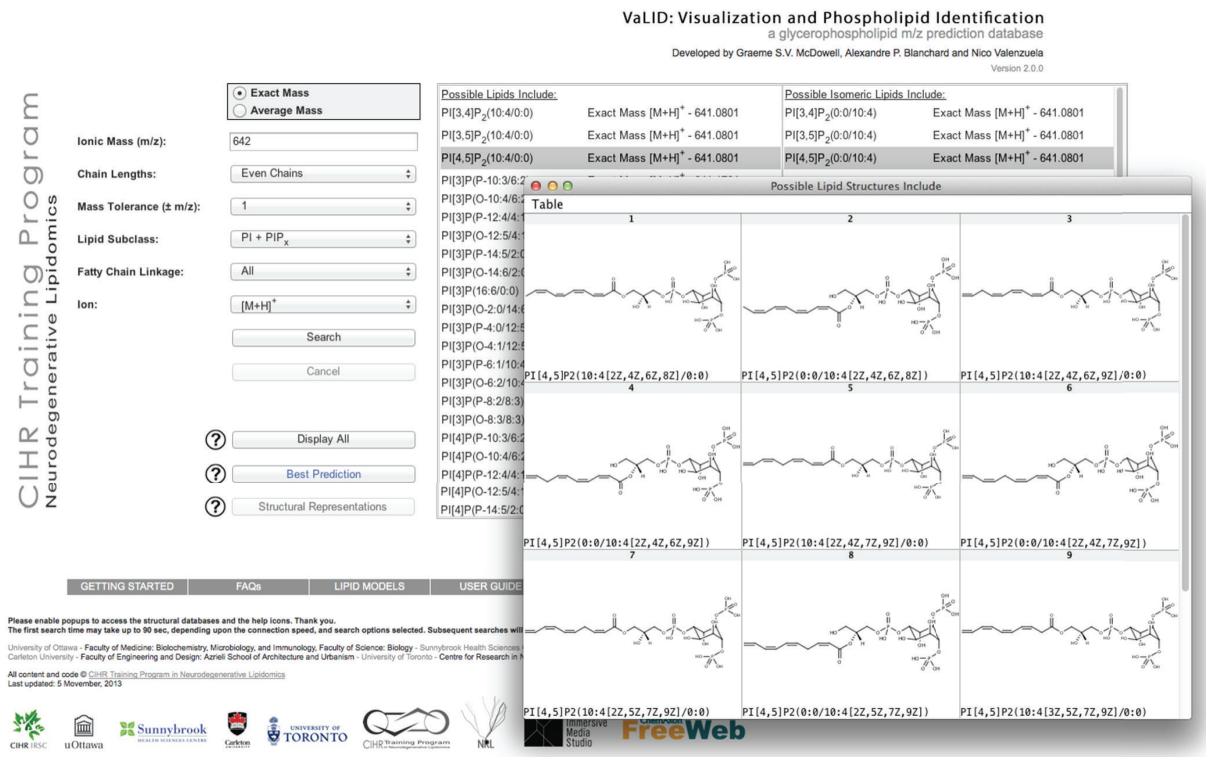


FIGURE 4: Automated drawing feature of VaLID 2.0.0. An example of a search button, returning all possible PI and PIP_x lipids with *m/z* of 642 (exact mass with a user-defined tolerance of 1 amu), restricted to displaying even carbon chains only, and selecting $[M+H]^+$ ion mode in MS (back panel). The user then selected PI[4, 5]P_2(10:4/0:0) and its *sn*-1/*sn*-2 chain inversion species and pressed "Display All" button. The window labelled "Possible Lipid Structures Include" displays a table containing the possible structures for this lipid, with the restrictions as laid out in the user manual (inset). These drawings can be easily exported for use in publication figures as described in the user manual.

Authors' Contribution

Graeme S. V. McDowell and Alexandre P. Blanchard contributed equally to this work.

Acknowledgments

This resource was funded by the Canadian Institutes of Health Research (CIHR) MOP 89999 to DF and SALB and a Strategic Training Initiative in Health Research (STIHR) CIHR/Training Program in Neurodegenerative Lipidomics (CTPNL) and the Institute of Aging TGF 96121 to DF, SF, and SALB. APB received a FRSQ and CTPNL graduate scholarship; GSVM received a CTPNL graduate scholarship.

References

[1] S. A. L. Bennett, N. Valenzuela, H. Xu et al., "Using neurolipidomics to identify phospholipid mediators of synaptic (dys)function in Alzheimer's Disease," *Frontiers in Physiology*, vol. 4, p. 168, 2013.

[2] C. Le Roy and J. L. Wrana, "Clathrin- and non-clathrin-mediated endocytic regulation of cell signalling," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 2, pp. 112–126, 2005.

[3] L. C. Skwarek and G. L. Boulianne, "Great Expectations for PIP: phosphoinositides as regulators of signaling during development and disease," *Developmental Cell*, vol. 16, no. 1, pp. 12–20, 2009.

[4] D. Piomelli, G. Astarita, and R. Rapaka, "A neuroscientist's guide to lipidomics," *Nature Reviews Neuroscience*, vol. 8, no. 10, pp. 743–754, 2007.

[5] H. A. Brown and R. C. Murphy, "Working towards an exegesis for lipids in biology," *Nature Chemical Biology*, vol. 5, no. 9, pp. 602–606, 2009.

[6] M. Bou Khalil, W. Hou, H. Zhou et al., "Lipidomics era: accomplishments and challenges," *Mass Spectrometry Reviews*, vol. 29, no. 6, pp. 877–929, 2010.

[7] H. Xu, N. Valenzuela, S. Fai et al., "Targeted lipidomics—advances in profiling lysophosphocholine and platelet-activating factor second messengers," *FEBS Journal*, vol. 280, pp. 5652–5667, 2013.

[8] X. Han, K. Yang, and R. W. Gross, "Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses," *Mass Spectrometry Reviews*, vol. 31, no. 1, pp. 134–178, 2012.

[9] P. S. Niemelä, S. Castillo, M. Sysi-Aho, and M. Orešič, "Bioinformatics and computational methods for lipidomics," *Journal of Chromatography B*, vol. 877, no. 26, pp. 2855–2862, 2009.

[10] W. Hou, H. Zhou, M. B. Khalil, D. Seebun, S. A. L. Bennett, and D. Figeys, "Lyso-form fragment ions facilitate the determination of stereospecificity of diacyl glycerophospholipids," *Rapid Communications in Mass Spectrometry*, vol. 25, no. 1, pp. 205–217, 2011.

[11] J. C. Smith, W. Hou, S. N. Whitehead, M. Ethier, S. A. L. Bennett, and D. Figeys, "Identification of lysophosphatidylcholine (LPC) and platelet activating factor (PAF) from PC12 cells and mouse cortex using liquid chromatography/multistage mass spectrometry (LC/MS3)," *Rapid Communications in Mass Spectrometry*, vol. 22, no. 22, pp. 3579–3587, 2008.

[12] S. N. Whitehead, W. Hou, M. Ethier et al., "Identification and quantitation of changes in the platelet activating factor family of glycerophospholipids over the course of neuronal differentiation by high-performance liquid chromatography electrospray ionization tandem mass spectrometry," *Analytical Chemistry*, vol. 79, no. 22, pp. 8539–8548, 2007.

[13] C.-H. Tang, P.-N. Tsao, C.-Y. Chen, M.-S. Shiao, W.-H. Wang, and C.-Y. Lin, "Glycerophosphocholine molecular species profiling in the biological tissue using UPLC/MS/MS," *Journal of Chromatography B*, vol. 879, no. 22, pp. 2095–2106, 2011.

[14] E. Fahy, D. Cotter, M. Sud, and S. Subramaniam, "Lipid classification, structures and tools," *Biochimica et Biophysica Acta*, vol. 1811, no. 11, pp. 637–647, 2011.

[15] U. Igbavboa, J. Hamilton, H.-Y. Kim, G. Y. Sun, and W. G. Wood, "A new role for apolipoprotein E: modulating transport of polyunsaturated phospholipid molecular species in synaptic plasma membranes," *Journal of Neurochemistry*, vol. 80, no. 2, pp. 255–261, 2002.

[16] M. J. Sharman, G. Shui, A. Z. Fernandis et al., "Profiling brain and plasma lipids in human apoe $\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$ knock-in mice using electrospray ionization mass spectrometry," *Journal of Alzheimer's Disease*, vol. 20, no. 1, pp. 105–111, 2010.

[17] R. B. Chan, T. G. Oliveira, E. P. Cortes et al., "Comparative lipidomic analysis of mouse and human brain with Alzheimer disease," *The Journal of Biological Chemistry*, vol. 287, no. 4, pp. 2678–2688, 2012.

[18] P. H. Axelsen and R. C. Murphy, "Quantitative analysis of phospholipids containing arachidonate and docosahexaenoate chains in microdissected regions of mouse brain," *Journal of Lipid Research*, vol. 51, no. 3, pp. 660–671, 2010.

[19] S. Osawa, S. Funamoto, M. Nobuhara et al., "Phosphoinositides suppress γ-secretase in both the detergent-soluble and -insoluble states," *The Journal of Biological Chemistry*, vol. 283, no. 28, pp. 19283–19292, 2008.

[20] E. Fahy, D. Cotter, R. Byrnes et al., "Bioinformatics for Lipidomics," *Methods in Enzymology*, vol. 432, pp. 247–273, 2007.

[21] T. Kind, K. H. Liu, Y. Lee do et al., "LipidBlast in silico tandem mass spectrometry database for lipid identification," *Nature Methods*, vol. 10, no. 8, pp. 755–758, 2013.

[22] A. P. Blanchard, G. S. McDowell, N. Valenzuela et al., "Visualization and Phospholipid Identification (VaLID): online integrated search engine capable of identifying and visualizing glycerophospholipids with given mass," *Bioinformatics*, vol. 29, no. 2, pp. 284–285, 2013.

[23] J. R. De Laeter, J. K. Böhlke, P. De Bièvre et al., "Atomic weights of the elements: review 2000," *Pure and Applied Chemistry*, vol. 75, no. 6, pp. 683–800, 2003.

[24] G. Di Paolo and P. De Camilli, "Phosphoinositides in cell regulation and membrane dynamics," *Nature*, vol. 443, no. 7112, pp. 651–657, 2006.

[25] M. Miyazaki and J. M. Ntambi, "Fatty acid desaturation and chain elongation in mammals," in *Biochemistry of Lipids, Lipoproteins and Membranes*, D. E. Vance and J. E. Vance, Eds., pp. 191–211, Elsevier, 2008.

[26] E. Fahy, S. Subramaniam, R. C. Murphy et al., "Update of the LIPID MAPS comprehensive classification system for lipids," *Journal of Lipid Research*, vol. 50, pp. S9–S14, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 281912, 10 pages
http://dx.doi.org/10.1155/2014/281912

Research Article

Microsatellites in the Genome of the Edible Mushroom, *Volvariella volvacea*

Ying Wang,[1] Mingjie Chen,[1] Hong Wang,[1] Jing-Fang Wang,[2] and Dapeng Bao[1]

[1] *National Engineering Research Center of Edible Fungi and Key Laboratory of Applied Mycological Resources and Utilization, Ministry of Agriculture and Shanghai Key Laboratory of Agricultural Genetics and Breeding and Institute of Edible Fungi, Shanghai Academy of Agriculture Science, Shanghai 201403, China*
[2] *Key Laboratory of Systems Biomedicine, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai 200240, China*

Correspondence should be addressed to Jing-Fang Wang; jfwang8113@gmail.com and Dapeng Bao; baodp@hotmail.com

Received 2 October 2013; Revised 23 October 2013; Accepted 23 October 2013; Published 19 January 2014

Academic Editor: Yudong Cai

Using bioinformatics software and database, we have characterized the microsatellite pattern in the *V. volvacea* genome and compared it with microsatellite patterns found in the genomes of four other edible fungi: *Coprinopsis cinerea*, *Schizophyllum commune*, *Agaricus bisporus,* and *Pleurotus ostreatus*. A total of 1346 microsatellites have been identified, with mono-nucleotides being the most frequent motif. The relative abundance of microsatellites was lower in coding regions with 21 No./Mb. However, the microsatellites in the *V. volvacea* gene models showed a greater tendency to be located in the CDS regions. There was also a higher preponderance of trinucleotide repeats, especially in the kinase genes, which implied a possible role in phenotypic variation. Among the five fungal genomes, microsatellite abundance appeared to be unrelated to genome size. Furthermore, the short motifs (mono- to tri-nucleotides) outnumbered other categories although these differed in proportion. Data analysis indicated a possible relationship between the most frequent microsatellite types and the genetic distance between the five fungal genomes.

1. Introduction

Volvariella volvacea, the Chinese straw mushroom, is an edible, straw-degrading, basidiomycetous fungus that has been cultivated for over 300 years. Currently ranked third in terms of production worldwide [1, 2], this mushroom has commercial importance that continues to increase due to its delicious flavor and texture, nutritional attributes, medicinal properties, and short cultivation cycle. *V. volvacea* is rich in protein, essential amino acids, vitamin C, and other bioactive components [3, 4]. According to traditional Chinese medicine, consuming the mushroom is good for the liver [5] and stomach, relieves summer heats, and enriches milk production in women following childbirth. Furthermore, antioxidants from *V. volvacea* are reported to enhance immunity, reduce cholesterol levels, and prevent atherosclerosis [6]. *V. volvacea* also plays an important ecological role by degrading the various agricultural wastes such as rice and wheat straw, cottonseed hulls, sugar cane bagasse, oil palm pericarp, banana leaves, and other carbonaceous materials used for cultivation [3, 7]. However, in spite of these nutritional, medicinal, and environmental benefits, relatively little is known about the molecular biology of this mushroom.

Microsatellites, also known as simple sequence repeats (SSRs) or short tandem repeats (STRs), are 1–6 base-pair nucleotide motifs, repeated in tandem at least five times. They are distributed in both coding and noncoding regions of eukaryotic and prokaryotic genomes [8, 9] and exhibit high levels of polymorphism. Since the late 1980s, many microsatellites have been used as genetic makers for species identification [10] and classification and in genome fingerprinting and mapping studies [11–14]. Subsequent research has revealed that microsatellites are involved in gene regulation, and organism development and evolution [15, 16]. In most cases, the effects of microsatellites are determined by their genomic location. For example, mutations to microsatellites located in coding and promoter regions lead to phenotype modification [17–20], while in

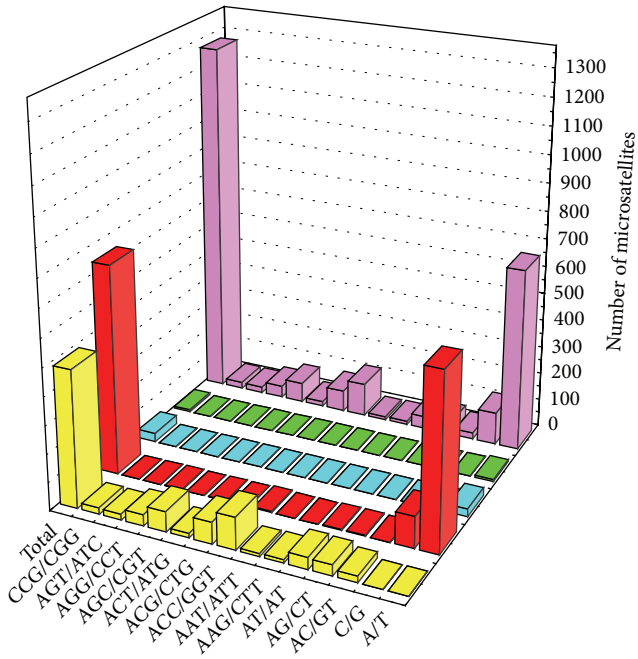


Figure 1

5′-untranslated regions (5′-UTRs) they affect gene transcription or regulation. In intron regions, microsatellite mutations impact gene transcription, regulation, mRNA splicing, and gene silencing, and in 3′-UTRs they are involved in gene silencing and transcription slippage [16, 21]. Moreover, since the frequency of these elements in genomes is proportional to the genome coverage, they can help explain how genomes are organized [11]. However, there are major drawbacks associated with microsatellite research including the time and costs associated with isolating microsatellites from the whole genome [22], a process that involves the screening of small insert genomic DNA libraries or the construction of SSR-enriched libraries [23]. However, our previous research into the degradation and sexual reproduction mechanisms adopted by *V. volvacea* has resulted in the sequencing of the whole genome [1], thereby facilitating genome-wide analyses of microsatellite distribution.

In this study, the entire genome of a monokaryotic *V. volvacea* strain (V23-1) has been screened to determine the distribution and density of microsatellites in different genomic regions. Particular emphasis has been given to microsatellites located in genes with molecular functions, and microsatellites in the *V. volvacea* genome have also been compared with those present in the genomes of four other edible fungi. For this purpose, we designed 100 primer pairs based on identified microsatellites loci used for genetic mapping. Our data will serve to establish the functional and evolutionary significance of these sequences and contribute to their future use as molecular markers.

2. Materials and Methods

The complete genome sequence of *V. volvacea* strain V23 was downloaded from the FTP site of GenBank. Information relating to the location of the gene models, introns, coding sequences (CDSs), 5′-untranslated-(5′-UTRs), 3′-untranslated-(3′-UTRs), and intergenic regions were obtained from the *V. volvacea* genome study group. 5′-UTRs are defined as the sequence located between a transcription start point and the beginning of the start codon of the transcript. 3′-UTRs are defined as the sequence between the stop codon and the last base of the transcript. Except introns, CDSs and UTRs, all the other regions in the genome are classified as intergenic regions. *Coprinus cinereus*, *Schizophyllum commune*, *Agaricus bisporus*, and *Pleurotus ostreatus* genome sequences used in this study (Table S1) (See Supplementary material available online at http://dx.doi.org/10.1155/2014/281912) were downloaded from the Joint Genome Institute (JGI) website.

MISA software (http://pgrc.ipk-gatersleben.de/misa/) was used to locate and identify both perfect microsatellites and compound microsatellites interrupted by a certain number of bases. Mono- to hexanucleotide microsatellite motifs were identified using the following default parameters: mono- with at least 10 repeats; di- with at least six repeats; tri-, penta-, and hexa- with at least five repeats; the maximum number of bases between two microsatellites was 100 bp. [18, 24]. Unit patterns of repeats with circular permutations were considered as one type for statistical analysis. The same conditions were used to identify microsatellites in all the genome assemblies [25, 26]. To more accurately compare all the repeat types existing in different genomic regions, the relative abundance (mean of the number of microsatellites per Mb of the sequence analyzed) and the relative density (mean of the microsatellite length in bp per Mb of the sequence analyzed) were calculated separately [27]. Since longer microsatellites may display higher levels of polymorphism, primers for these loci were designed using the Primer3 software (http://frodo.wi.mit.edu/primer3/).

Gene models having microsatellites in exons, introns, and UTRs were isolated and then scanned for InterPro domains and gene ontology (GO) annotation. The latter was used to assign each gene-encoded protein to one of the three defined categories (molecular function, biological process, or cellular component), and WEGO (Web Gene Ontology Annotation Plot) [28] was used to plot the GO annotation data.

3. Results

3.1. Identification and Location of Microsatellites in the V. volvacea Genome. The *V. volvacea* genome contained a total of 1346 microsatellites, with a relative abundance of 38 microsatellites per Mbp (Table 1). Microsatellites with periods ranging from 1 to 6 (i.e., mono- to hexa-) accounted for 57.4% (773), 8.2% (110), 29.7% (400), 1.4% (19), 1.0% (14), and 2.2% (30) of the total, respectively. From these, 100 microsatellite loci were selected and 100 primers were designed accordingly (Table S2).

The entire genome (35.72 Mb) was divided into five regional types consisting of 5′UTRs (0.69 Mb), CDSs (17.42 Mb), introns (5.71 Mb), 3′UTRs (1.6 Mb), and intergenic (10.31 Mb). Microsatellites were more numerous in the intergenic regions with a relative abundance of

Table 1: Number, percentage, and relative abundance of microsatellites in the different regions of the *V. volvacea* genome.

	5′ UTR	CDS	Introns	3′ UTR	Intergenic regions	Total
Genome size/Mb	0.69	17.42	5.71	1.61	10.31	35.72
Percentage of the genome	1.93	48.77	15.99	4.51	28.86	100
Mono No.	8	32	231	49	453	773
%	1.03	4.14	29.88	6.34	58.60	100
No./Mb	12	2	40	30	44	22
Di No.	1	16	21	2	70	110
%	0.91	14.55	19.09	1.82	63.64	100
No./Mb	1	1	4	1	7	3
Tri No.	6	299	29	15	51	400
%	1.5	74.75	7.25	3.75	12.75	100
No./Mb	9	17	5	9	5	11
Tetra No.	2	—	3	2	12	19
%	10.53	—	15.79	10.53	63.16	100
No./Mb	3	—	—	1	1	1
Penta No.	—	2	4	4	4	14
%	—	14.29	28.57	28.57	28.57	100
No./Mb	—	—	1	2	—	0
Hexa No.	—	16	4	2	8	30
%	—	53.33	13.33	6.67	26.67	100
No./Mb	—	1	1	1	1	1
All SSRs No.	17	365	292	74	598	1346
%	1.26	27.12	21.69	5.50	44.43	100
No./Mb	25	21	51	46	58	38

58 No./Mb (Table 1). Over 50% of the mono-, di-, and tetranucleotides were located in the intergenic regions, whereas tri- and hexanucleotides appeared more frequently in the CDSs. Pentanucleotide microsatellites were fairly evenly distributed within the CDSs, introns, 3′UTRs, and intergenic regions. No penta- and hexanucleotides were found in the 5′UTRs, and no tetranucleotides were found in CDSs.

The fourteen most abundant microsatellite types (≥10 motifs) detected in the *V. volvacea* genome constituted 94.7% of the total (Figure 1). A/T occurred at the highest frequency (51.6%), followed by ACC/GGT (9.34%), and C/G (9.11%). With the exception of the mononucleotide motifs, the majority of these most abundant microsatellites motifs were repeated less than 10 times. Only 13 (1%) microsatellite motifs exhibited a large repeat number. Trinucleotides were the primary types among the above fourteen most abundant microsatellites types, but their distribution was highly variable with a wide range (from 0.86% to 9.34%) of frequencies. In addition, there were fifteen tetranucleotide types, twelve pentanucleotide types and twenty-five hexanucleotide types identified within the whole genome, but none of these contained more than ten different motifs. Among the fourteen most abundant microsatellites types (Figure 1), the longest motifs were T (61 bp), CAC (42 bp), GAG (39 bp), TGA (30 bp), AAT (27 bp), AAG (24 bp), ACG (24 bp), TCA (24 bp), CTG (24 bp), AT (22 bp), C (21 bp), AG (20 bp), CCG (18 bp), and GT (14 bp), respectively.

3.2. Distribution of Microsatellites in the V. volvacea Gene Models and Functional Properties of the Genes Containing Microsatellites. Of the 11084 genes identified in the whole genome of *V. volvacea*, 649 (5.9%) contained 748 microsatellites, with 72 of these containing more than one microsatellite. The relative abundance of microsatellites in the gene models was 29 No./Mb. Altogether, 365 microsatellites were detected in the CDSs of 323 genes (2.9% of the total), contributing 27.1% of the total in the entire *V. volvacea* genome. Trinucleotides (299) were the most abundant category (81.9%) in all the gene models and, with the exception of the mononucleotide microsatellites, the trinucleotide CAC was the most frequent (with 35 motifs), followed by CAG (24 motifs), GAC (20 motifs), and CCA (16 motifs). All these motifs encoded aliphatic amino acids such as valine, leucine, and glycine. InterPro and KEGG database scanning revealed that, of the 649 gene models containing microsatellites, 365 contained at least one known domain, 190 participated in a biological pathway, and 147 had been annotated definitively. These genes encoded proteins including cytochrome P450 monooxygenases, carbohydrate-degrading enzymes, kinases, dehydrogenases, and transport proteins. Thorough analysis of the microsatellite distribution and motif type among the 147 annotated genes revealed that the microsatellites were more frequent within the CDS regions of the kinase genes and that trinucleotides were the most abundant motif (Table S3). However, these conditions did not apply to other gene categories. In terms of molecular function, the annotated genes included three with electron carrier activity, 23 with

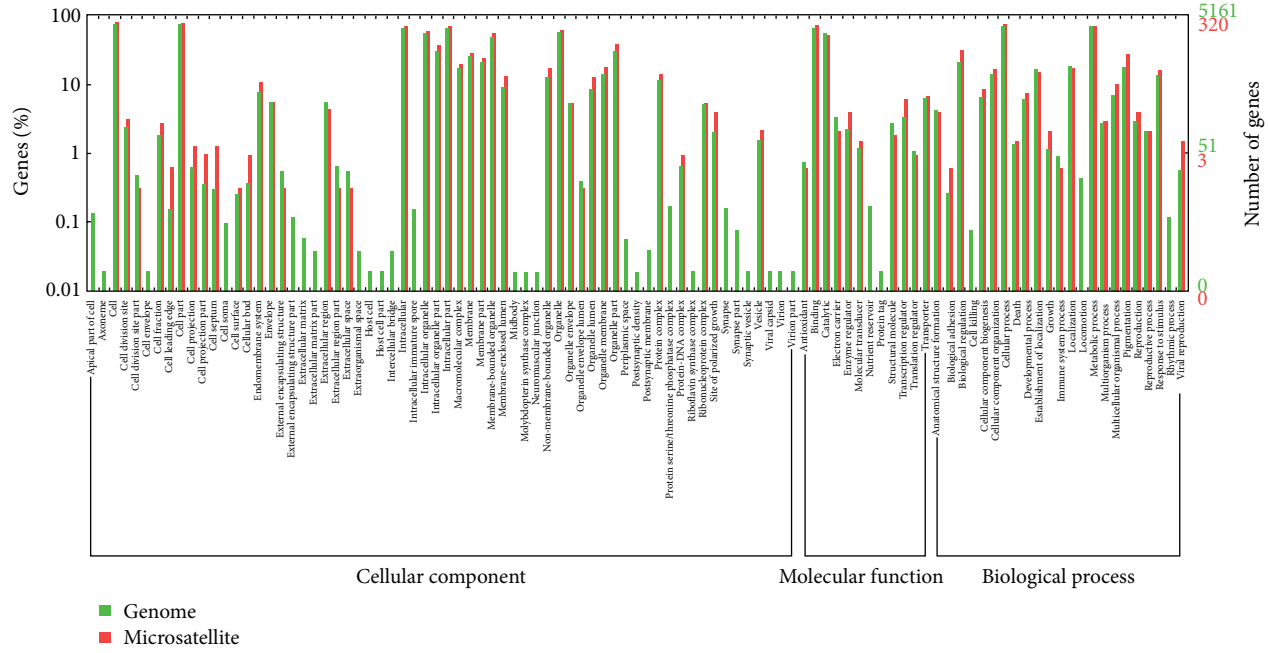


Figure 2

catalytic activity, five involved in binding, two with structural molecule activity, and two with transporter activity (Table 2). Some genes contained more than one microsatellite. For example, four different types of microsatellite motifs, (GAC)5, (GCT)5, (GAC)5, and (CCGCAC)6, were detected in the CDS of the CAMK/CAMK1 protein kinase gene. The Ca^{2+} transporting ATPase gene also contained four microsatellites, three in the CDS and one in the intron regions, while two or three microsatellites were identified in either the CDS or intron regions of other genes.

The whole genome of *V. volvacea* and the genes having microsatellites were categorized on the basis of their homologous gene function by the Gene Ontology Consortium. Gene ontology numbers for the best homologous hits were used to determine molecular function, cellular component, and biological process ontology for these sequences. An inferred putative gene ontology annotation was found for 5161 genes in the genome, of which only 320 (254 associated with cellular components, 243 with biological processes, and 242 with molecular function) contained microsatellite loci. The GO terms in the three ontologies of the whole genome totaled 105, with 73 in the genes having microsatellites. In the cellular component ontology, the cell (76.2% in whole genome and 79.4% in genes containing microsatellites) and the cell part (76.2% and 79.4%) were the two main types of GO term, followed by the organelle (56.8% and 60.9%) and organelle part (30.8% and 38.8%). In the biological process ontology, cellular process (69.9%, 75.9%) and metabolic process (70.1%, 68.8%) were the two main GO terms, and in the Molecular Function ontology, the functions of the major components were binding (65.5%, 75.6%) and catalytic activity (58.8%, 52.5%). Compared with the whole genome, some GO terms were not present in genes having microsatellites, including cell apex (GO:0045177), external encapsulating structure (GO:0044462), intracellular immature spore (GO:0042763), protein serine/threonine phosphatase complex (GO:0008287), synapse (0045202), nutrient reservoir activity (GO:0045735) and locomotion (GO:0040011), (Figure 2). Most of the absent GO terms were concentrated in the cellular component ontology.

3.3. Comparison of Microsatellite Distribution in the Genomes of V. Volvacea and Four Other Edible Fungi. The number and types of microsatellites in the *V. volvacea* genome and four other fully sequenced edible fungal genomes that varied in size from 30.2 Mb (*Agaricus bisporus*) to 38.6 Mb (*Schizophyllum commune*) were compared (Table 3). Microsatellite content in these species was not directly proportional to the size of the genome since *A. bisporus* contained the highest number (3134, relative abundance 103.8 per Mbp) compared with only 1206 in *S. commune* (relative abundance 31.2 per Mbp). The *C. cinereus* and *P. ostreatus* genomes contained 2050 and 1314 microsatellites, respectively. However, although *V. volvacea* and *P. ostreatus* showed the same relative abundance of microsatellites (38 per Mbp), the relative densities were different because the total length of microsatellites in *P. ostreatus* was longer. Yet, based on the comparison of microsatellites in the wholegenomes, the microsatellite content in the *V. volvacea* genome exhibited closer similarity to *P. ostreatus* and *S. commune* than to the *A. bisporus* and *C. cinereus*.

The five fungal genomes exhibited considerable differences with respect to the number, relative abundance, and relative density of mono-, di-, tri-, tetra-, penta-, and hexanucleotides (Table 4, Figure 3). For example, mononucleotide motifs outnumbered all other microsatellite classes

Table 2: Microsatellites in the gene models with some molecular functions in *V. volvacea*.

Molecular function	Gene product	Location
Electron carrier activity	CYP547B1	Intron
	CYP5080B3b	Intron
	CYP627A1	Intron
Catalytic activity	Nonribosomal peptide synthetase 12	Intron
	Protoporphyrinogen oxidase	Intron
	Glycoside hydrolase family 13 protein	CDS
	ATP citrate lyase isoform 2	3′ UTR
	Adenylate cyclase	CDS
	Xylanase	CDS
	Pyruvate dehydrogenase	CDS
	Modular protein with glycoside hydrolase family 13 and glycosyltransferase family 5 domains	Intron
	long-chain-fatty-acid-CoA ligase	3′ UTR
	Glycoside hydrolase family 18 protein	CDS, intron
	Glycoside hydrolase family 15 protein	Intron
	Glycoside hydrolase family 35 protein	Intron
	AMP dependent CoA ligase	Intron
	Serine palmitoyltransferase 2	CDS
	IMP dehydrogenase	CDS, intron
	Trehalase	Intron
	DNA helicase	CDS
	Exo-beta-1,3-glucanase	Intron
	Glycoside hydrolase family 10 and carbohydrate-binding module family 1 protein	Intron
	Potassium/sodium efflux P-type ATPase	CDS × 3, Intron
	Glycoside hydrolase family 38 protein	Intron
	Sodium transport ATPase	Intron
	Fructose-bisphosphate aldolase	CDS, intron
Binding	Sec7 guanine nucleotide exchange factor	3′ UTR
	COP8	CDS
	STE/STE11/cdc15 protein kinase	CDS
	Clathrin-coated vesicle protein	Intron
	Carnitine/acyl carnitine carrier	Intron
Structural molecule activity	Beta-tubulin 2 tubb2	Intron
	Iron sulfur assembly protein 1	Intron
Transporter activity	Urea transporter	3′ UTR
	Vacuole protein	Intron

Table 3: Overview of the five edible fungal genomes.

	V. volvacea	*C. cinereus*	*S. commune*	*A. bisporus*	*P. ostreatus*
Sequence analyzed (Mb)	35.7	36.2	38.6	30.2	34.3
GC contents (%)	48.8	51.67	57.5	46.48	50.94
No. of SSRs	1346	2050	1206	3134	1314
Relative abundance (No./Mb)	38	56	31	104	38
Total length of SSRs (bp)	19347	32601	21538	44690	25265
Relative density (bp/Mb)	541	898	558	1478	737
Genome content (%)	0.05	0.09	0.06	0.15	0.07

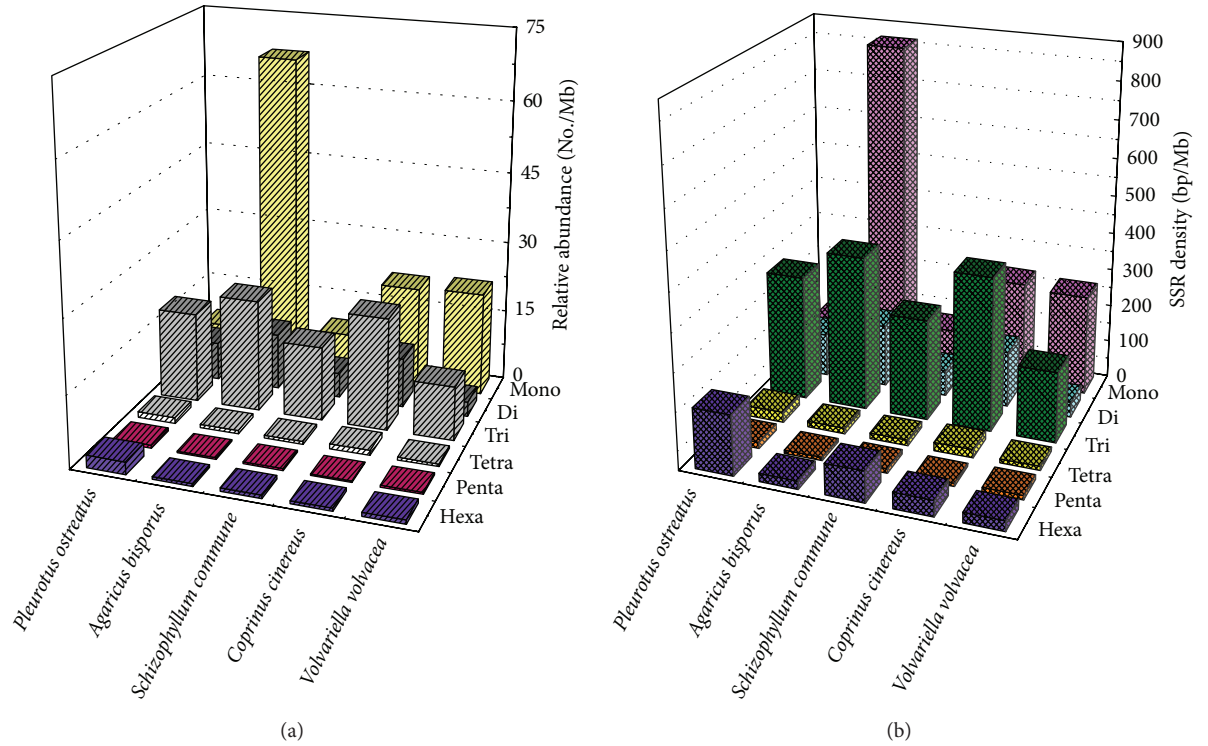


Figure 3

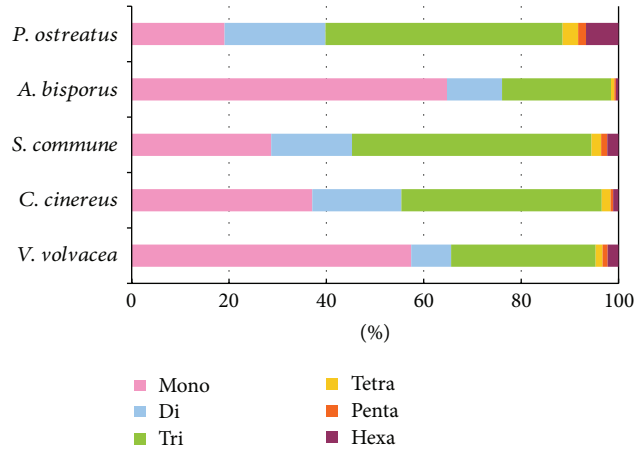


Figure 4

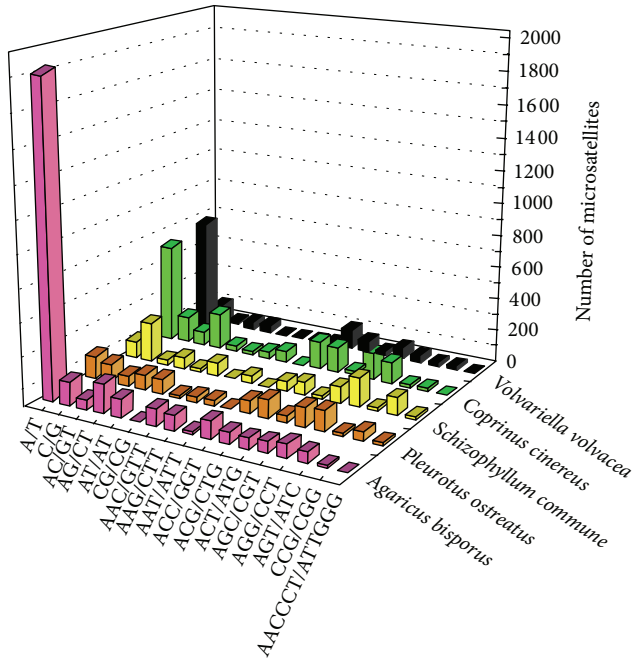


Figure 5

in the *V. volvacea* and *A. bisporus* genomes, with 2030 (64.8% of the total) detected in the latter. However, trinucleotide microsatellites (followed by mononucleotides) were the most common motifs in *S. commune* and *C. cinereus*, while trinucleotide microsatellites (followed by dinucleotides) were most frequent in *P. ostreatus*. Comparison of the relative abundance (Figure 3(a)) and the relative density (Figure 3(b)) of the six microsatellite categories revealed general agreement except in the case of the *P. ostreatus* genome where hexanucleotide microsatellites were infrequent but relatively dense. In contrast to the clear disparity in the total number of microsatellites in the *V. volvacea* and *A. bisporus* genomes, the proportion of the six microsatellite categories was very similar (Figure 4).

The number of motif types of 17 different microsatellites in each of the five mushroom species is shown in Figure 5. The mononucleotide A/T was the most frequent in the majority of species, and only the *S. commune* genome contained

Table 4: Occurrence, relative abundance, total length, and relative density of microsatellites in the five edible fungal genomes.

	V. volvacea	*C. cinereus*	*S. commune*	*A. bisporus*	*P. ostreatus*
Mono					
No.	773	761	346	2030	251
No./Mb	22	21	9	67	7
Length (bp)	9440	9950	4425	26065	3131
Bp/Mb	264	275	115	863	91
Di					
No.	110	375	200	354	272
No./Mb	3	10	5	12	8
Length (bp)	1464	5354	2664	4980	3904
Bp/Mb	41	148	69	165	114
Tri					
No.	400	843	593	701	640
No./Mb	11	23	15	23	19
Length (bp)	6660	14655	10299	12267	11295
Bp/Mb	187	405	267	406	329
Tetra					
No.	19	38	24	22	42
No./Mb	1	1	1	1	1
Length (bp)	392	812	568	460	940
Bp/Mb	11	22	15	15	27
Penta					
No.	14	10	15	10	21
No./Mb	0	0	0	0	1
Length (bp)	425	300	450	270	565
Bp/Mb	12	8	12	9	16
Hexa					
No.	30	23	28	17	88
No./Mb	1	1	1	1	3
Length (bp)	966	1530	3132	648	5430
Bp/Mb	27	42	81	21	158

more C/G than A/T. Of the dinucleotide motifs, CG was common in *C. cinereus*, *S. commune,* and *P. ostreatus* but least frequent in *V. volvacea* and *A. bisporus*. Conversely, the trinucleotide AAT/ATT was comparatively common in *V. volvacea* and *A. bisporus* but rare in *C. cinereus*, *S. commune,* and *P. ostreatus*. Furthermore, only eight AAC/GTT trinucleotide motifs were detected in *V. volvacea* and in *S. commune* compared with more than 35 in *C. cinereus*, *P. ostreatus*, and *A. bisporus*. Tetra-, penta-, and hexanucleotide SSR densities were very low and only the hexanucleotide AACCCT/ATTGGG was relatively common with at least 10 motifs identified in both *S. commune* and *P. ostreatus*. The longest motifs varied from 34 repeats of the dinucleotide AG/CT and the hexanucleotide AACCCT/ATTGGG in *A. bisporus* and *C. cinereus*, respectively, to 61 repeats of the mononucleotide A/T in *V. volvacea* (Table 5). AACCCT/ATTGGG was also the longest microsatellite identified in *S. commune* and *P. ostreatus* with 36 and 38 repeats, respectively.

4. Discussion

Microsatellites have contributed significantly to studies in population genetics [29, 30] and molecular ecology [11], have served to explain the phenomenon of genome expansion in certain species [31], influenced the expression of quantitative genetic traits [32], and have been used to analyze the human genome for human diseases [33, 34]. Considerable progress has been made in microsatellite development, including associated bioinformatics. For example, several studies have focused on the basic distribution patterns and diversity across whole genomes to better understand the role of microsatellites. However, since relatively little comprehensive analysis of microsatellites in the genomes of edible fungi has been undertaken, we have used computational analysis to characterize and compare the microsatellites in the entire genome of *V. volvacea* and the genomes of four other edible fungi (Table S1). Information on the relative abundance of these microsatellites, combined with the distribution patterns

Table 5: The longest microsatellite motifs in the five edible fungal genomes.

	Longest microsatellites		
	Motif	Repeats	Size
V. volvacea	A/T	61	61
C. cinereus	AACCCT/ATTGGG	34	204
S. commune	AACCCT/ATTGGG	36	216
A. bisporus	AG/CT	34	68
P. ostreatus	AACCCT/ATTGGG	38	228

in both the coding and noncoding regions of the genome, may provide clues to the functionality of microsatellites in gene regulation.

At present, there is no standard cut-off limit for the minimum length of microsatellites [35]. Adopting slippage rate changes of around 10 bases for mono- and dinucleotide repetitions, universal thresholds of 8–10 bp and 7–10 bp were proposed for mononucleotide microsatellites in yeast [36] and eukaryotes [37], respectively. However, the threshold for human microsatellites was found to depend on their motif size (9 repeats for mononucleotide and 4 repeats for di-, tri-, and tetranucleotide microsatellites) [38]. Accordingly, and in order to compare our data with previous studies of other fungi, we identified mono-, di-, tri-, tetra-, penta- and hexanucleotide microsatellite motifs at minimum repeat numbers of 10, 6, 5, 5, 5, and 5, respectively. Of the 1346 microsatellites identified, 72.9% were embedded in noncoding DNA (corresponding to 51.23% of the genome assembly), and 61% were located in the intergenic regions (28.9% of the genome assembly). This distribution pattern, which is common in fungal genomes, has been attributed to negative selection against frame-shift mutations in coding regions [39] and possibly accounts for the small number of microsatellites in fungi. Another contributing factor is the relatively smaller amount of noncoding DNA in fungi, due to the high density of genes within the fungi genome, compared with higher eukaryotes. This distribution implies that microsatellites are generated in intergenic regions by duplication or transposition.

In addition to possible relevance as evolutionary neutral DNA markers [40], microsatellites have some functional significance, including effects on chromatin organization, regulation of gene activity, recombination, DNA replication, cell cycle, and mismatch repair systems [41, 42]. Firstly, we can estimate the function of a microsatellite by its position. Judging from previous experience, microsatellites located in CDSs can alter the function of the protein [20], those located in introns can affect gene transcription, those located in 5′UTRs can regulate gene expression, and those located in 3′UTRs may cause transcription slippage [43]. In *V. volvacea*, 55.6% of the total number of microsatellites were scattered in gene models and nearly half of these microsatellites were located in coding regions. Due to the high mutation rate of microsatellites, genes containing microsatellites in their coding regions would not be conserved. Close inspection of the great majority of microsatellites appearing in kinase-encoding genes was located in CDSs, suggesting these genes are capable of undergoing mutation. Microsatellites were also found in the introns, 5′UTRs, and 3′UTRs of CYPs, carbohydrate-degrading enzymes, dehydrogenases, and transport proteins. Earlier studies have shown that microsatellites located in promoter regions may affect gene activity [44]. In addition, some long microsatellites located in intergenic regions may have special functions. For instance, long microsatellites involved in sister chromatid cohesion, which indirectly assist kinetochore formation, are highly clustered in the centromere [45]. Excess numbers of microsatellite repeats may play important roles both in genomic stability and also in the evolution of additional genomic features. Consequently, the longest motifs (T)61, (CAC)14, and (GAG)13 in *V. volvacea* merit close attention with respect to possible functionality.

Comparative analyses of the microsatellite distribution in the genomes of *V. volvacea* and four edible fungi revealed that, in all cases, the majority of the microsatellites were mono-, di-, and trinucleotides, accounting for up to approximately 90% of all the microsatellites identified. However, their respective percentages varied in the different genomes. *V. volvacea* and *A. bisporus* showed a great affinity for mononucleotide repeats compared to the other three genomes in which trinucleotide microsatellites were the predominant type. Furthermore, microsatellites in the *S. commune*, *C. cinereus* and *P. ostreatus* genomes were comparatively longer. This was unexpected, especially in the case of *P. ostreatus* (the second smallest genome of the five), since an earlier study had concluded that microsatellites in larger genomes were longer compared with those in smaller genomes [27]. Hence, neither the abundance nor the length of the microsatellites in these fungi was correlated with genome size. In the case of microsatellite types, few trends were evident. A/T were the most frequent mononucleotide repeats in *A. bisporus*, *V. volvacea*, *C. cinereus*, and *P. ostreatus*, whereas C/G was most frequent in the *S. commune* genome. Among the dinucleotide motifs, AC/GT, AG/CT, and AT/AT were present at higher frequencies in *A. bisporus*, *V. volvacea*, *C. cinereus*, and *P. ostreatus*, whereas CG/CG was the most common in *S. commune*. The reason for the difference may be attributable in part to the higher GC content in the *S. commune* genome and to its distant phylogenetic position.

5. Conclusion

Comprehensive analysis of microsatellites in the *V. volvacea* and four other completely sequenced edible fungal genomes will provide better understanding of the nature of these important sequences. Such understanding of the characteristics of microsatellites in the genomes of *V. volvacea* and the other edible fungi will serve many useful purposes including the isolation and development of variable markers and will facilitate research on the role of microsatellites in genome organization.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This study was supported by the Shanghai Committee of Science and Technology, China (Grant no. 10dz2212400); Chinese National Science and Technology Support Program (Grant no. 2013BAD16B02); and Shanghai Key Scientific and Technological Project for Agriculture. The authors thank Dr. John Buswell for linguistic revision of the paper.

References

[1] D. Bao, M. Gong, H. Zheng et al., "Sequencing and comparative analysis of the straw mushroom (*Volvariella volvacea*) genome," *PLoS ONE*, vol. 8, no. 3, Article ID e58294, 2013.

[2] G. Thiribhuvanamala, S. Krishnamoorthy, K. Manoranjitham, V. Praksasm, and S. Krishnan, "Improved techniques to enhance the yield of paddy straw mushroom (*Volvariella volvacea*) for commercial cultivation," *African Journal of Biotechnology*, vol. 11, no. 64, pp. 12740–12748, 2012.

[3] Y. J. Cai, S. J. Chapman, J. A. Buswell, and S.-T. Chang, "Production and distribution of endoglucanase, cellobiohydrolase, and β- glucosidase components of the cellulolytic system of Volvariella volvacea, the edible straw mushroom," *Applied and Environmental Microbiology*, vol. 65, no. 2, pp. 553–559, 1999.

[4] S. T. Chang, "Mushroom research and development-equality and mutual benefit," *Mushroom Biology and Mushroom Products*, pp. 1–10, 1996.

[5] S. V. Kalava and S. G. Menon, "Protective efficacy of the extract of *volvariella volvacea* (bulliard ex fries) singer. against carbon tetrachloride induced hepatic injury," *International Journal of Pharmaceutical Sciences and Research*, vol. 3, no. 8, pp. 2849–2856, 2012.

[6] L. Ramkumar, T. Ramanathan, and J. Johnprabagaran, "Evaluation of nutrients, trace metals and antioxidant activity in *Volvariella volvacea* (Bull. Ex. Fr.) Sing," *Emirates Journal of Food and Agriculture*, vol. 24, no. 2, pp. 113–119, 2012.

[7] B. Z. Chen, F. Gui, B. G. Xie, F. Zou, Y. J. Jiang, and Y. J. Deng, "Sequence and comparative analysis of the MIP gene in Chinese straw mushroom, *Volvariella volvacea*," *Genome*, vol. 55, no. 9, pp. 667–672, 2012.

[8] D. Field and C. Wills, "Long, polymorphic microsatellites in simple organisms," *Proceedings of the Royal Society B*, vol. 263, no. 1367, pp. 209–215, 1996.

[9] G. Tóth, Z. Gáspári, and J. Jurka, "Microsatellites in different eukaryotic genomes: surveys and analysis," *Genome Research*, vol. 10, no. 7, pp. 967–981, 2000.

[10] S. Rufai, M. M. Hanafi, M. Y. Rafii, S. Ahmad, I. W. Arolu, and J. Ferdous, "Genetic dissection of new genotypes of drumstick tree (*Moringa oleifera Lam.*) using random amplified polymorphic DNA marker," *BioMed Research International*, vol. 2013, Article ID 604598, 6 pages, 2013.

[11] H. Ellegren, "Microsatellites: simple sequences with complex evolution," *Nature Reviews Genetics*, vol. 5, no. 6, pp. 435–445, 2004.

[12] E. Guichoux, L. Lagache, S. Wagner et al., "Current trends in microsatellite genotyping," *Molecular Ecology Resources*, vol. 11, no. 4, pp. 591–611, 2011.

[13] N. Mittal and A. K. Dubey, "Microsatellite markers—a new practice of DNA based markers in molecular genetics," *Pharmacognosy Reviews*, vol. 3, no. 6, pp. 235–246, 2009.

[14] C. Spampinato and D. Leonardi, "Molecular fingerprints to identify Candida species," *BioMed Research International*, vol. 2013, Article ID 923742, 10 pages, 2013.

[15] J. Labbé, C. Murat, E. Morin, F. Le Tacon, and F. Martin, "Survey and analysis of simple sequence repeats in the *Laccaria bicolor* genome, with development of microsatellite markers," *Current Genetics*, vol. 57, no. 2, pp. 75–88, 2011.

[16] M. J. Lawson and L. Zhang, "Distinct patterns of SSR distribution in the Arabidopsis thaliana and rice genomes," *Genome Biology*, vol. 7, no. 2, article R14, 2006.

[17] J. J. Rudd, J. Antoniw, R. Marshall, J. Motteram, B. Fraaije, and K. Hammond-Kosack, "Identification and characterisation of Mycosphaerella graminicola secreted or surface-associated proteins with variable intragenic coding repeats," *Fungal Genetics and Biology*, vol. 47, no. 1, pp. 19–32, 2010.

[18] C. Murat, C. Riccioni, B. Belfiori et al., "Distribution and localization of microsatellites in the Perigord black truffle genome and identification of new molecular markers," *Fungal Genetics and Biology*, vol. 48, no. 6, pp. 592–601, 2011.

[19] J. G. Gibbons and A. Rokas, "Comparative and functional characterization of intragenic tandem repeats in 10 Aspergillus genomes," *Molecular Biology and Evolution*, vol. 26, no. 3, pp. 591–602, 2009.

[20] K. J. Verstrepen, A. Jansen, F. Lewitter, and G. R. Fink, "Intragenic tandem repeats generate functional variability," *Nature Genetics*, vol. 37, no. 9, pp. 986–990, 2005.

[21] Y. C. Li, A. B. Korol, T. Fahima, and E. Nevo, "Microsatellites within genes: structure, function, and evolution," *Molecular Biology and Evolution*, vol. 21, no. 6, pp. 991–1007, 2004.

[22] S. Feng, H. Tong, Y. Chen et al., "Development of pineapple microsatellite markers and germplasm genetic diversity analysis," *BioMed Research International*, vol. 2013, Article ID 317912, 11 pages, 2013.

[23] C. Dutech, J. Enjalbert, E. Fournier et al., "Challenges of microsatellite isolation in fungi," *Fungal Genetics and Biology*, vol. 44, no. 10, pp. 933–949, 2007.

[24] J. Qian, H. Xu, J. Song, J. Xu, Y. Zhu, and S. Chen, "Genome-wide analysis of simple sequence repeats in the model medicinal mushroom Ganoderma lucidum," *Gene*, vol. 512, no. 2, pp. 331–336, 2013.

[25] J. Jurka and C. Pethiyagoda, "Simple repetitive DNA sequences from primates: compilation and analysis," *Journal of Molecular Evolution*, vol. 40, no. 2, pp. 120–126, 1995.

[26] C. Y. Li, L. Liu, J. Yang et al., "Genome-wide analysis of microsatellite sequence in seven filamentous fungi," *Interdisciplinary Sciences*, vol. 1, no. 2, pp. 141–150, 2009.

[27] H. Karaoglu, C. M. Y. Lee, and W. Meyer, "Survey of simple sequence repeats in completed fungal genomes," *Molecular Biology and Evolution*, vol. 22, no. 3, pp. 639–649, 2005.

[28] J. Ye, L. Fang, H. Zheng et al., "WEGO: a web tool for plotting GO annotations," *Nucleic Acids Research*, vol. 34, pp. W293–W297, 2006.

[29] P. Jarne and P. J. L. Lagoda, "Microsatellites, from molecules to populations and back," *Trends in Ecology and Evolution*, vol. 11, no. 10, pp. 424–429, 1996.

[30] G. Luikart, P. R. England, D. Tallmon, S. Jordan, and P. Taberlet, "The power and promise of population genomics: from genotyping to genome typing," *Nature Reviews Genetics*, vol. 4, no. 12, pp. 981–994, 2003.

[31] F. Martin, A. Kohler, C. Murat et al., "Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis," *Nature*, vol. 464, no. 7291, pp. 1033–1038, 2010.

[32] V. Albanèse, N. F. Biguet, H. Kiefer, E. Bayard, J. Mallet, and R. Meloni, "Quantitative effects on gene silencing by allelic variation at a tetranucleotide microsatellite," *Human Molecular Genetics*, vol. 10, no. 17, pp. 1785–1792, 2001.

[33] Y. D. Kelkar, N. Strubczewski, S. E. Hile, F. Chiaromonte, K. A. Eckert, and K. D. Makova, "What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats," *Genome Biology and Evolution*, vol. 2, no. 1, pp. 620–635, 2010.

[34] C. E. Pearson, K. N. Edamura, and J. D. Cleary, "Repeat instability: mechanisms of dynamic mutations," *Nature Reviews Genetics*, vol. 6, no. 10, pp. 729–742, 2005.

[35] E. Meglecz, G. Nève, E. Biffin, and M. G. Gardner, "Breakdown of phylogenetic signal: a survey of microsatellite densities in 454 shotgun sequences from 154 non model eukaryote species," *PLoS ONE*, vol. 7, no. 7, Article ID e40861, 2012.

[36] O. Rose and D. Falush, "A threshold size for microsatellite expansion," *Molecular Biology and Evolution*, vol. 15, no. 5, pp. 613–615, 1998.

[37] K. J. Dechering, K. Cuelenaere, R. N. H. Konings, and J. A. M. Leunissen, "Distinct frequency-distributions of homopolymeric DNA tracts in different genomes," *Nucleic Acids Research*, vol. 26, no. 17, pp. 4056–4062, 1998.

[38] Y. Lai and F. Sun, "The relationship between microsatellite slippage mutation rate and the number of repeat units," *Molecular Biology and Evolution*, vol. 20, no. 12, pp. 2123–2131, 2003.

[39] D. Metzgar, J. Bytof, and C. Wills, "Selection against frameshift mutations limits microsatellite expansion in coding DNA," *Genome Research*, vol. 10, no. 1, pp. 72–80, 2000.

[40] S. Temnykh, G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, and S. McCouch, "Computational and experimental analysis of microsatellites in rice (*Oryza sativa L.*): frequency, length variation, transposon associations, and genetic marker potential," *Genome Research*, vol. 11, no. 8, pp. 1441–1452, 2001.

[41] Y. C. Li, A. B. Korol, T. Fahima, A. Beiles, and E. Nevo, "Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review," *Molecular Ecology*, vol. 11, no. 12, pp. 2453–2465, 2002.

[42] S. Subramanian, R. K. Mishra, and L. Singh, "Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions," *Genome Biology*, vol. 4, no. 2, p. R13, 2003.

[43] D. E. Riley and J. N. Krieger, "UTR dinucleotide simple sequence repeat evolution exhibits recurring patterns including regulatory sequence motif replacements," *Gene*, vol. 429, no. 1-2, pp. 80–86, 2009.

[44] M. D. Vinces, M. Legendre, M. Caldara, M. Hagihara, and K. J. Verstrepen, "Unstable tandem repeats in promoters confer transcriptional evolvability," *Science*, vol. 324, no. 5931, pp. 1213–1216, 2009.

[45] T. D. Murphy and G. H. Karpen, "Localization of centromere function in a Drosophila minichromosome," *Cell*, vol. 82, no. 4, pp. 599–609, 1995.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 750961, 9 pages
http://dx.doi.org/10.1155/2014/750961

Research Article

De Novo Assembly and Characterization of *Sophora japonica* Transcriptome Using RNA-seq

Liucun Zhu,[1] Ying Zhang,[2] Wenna Guo,[1] Xin-Jian Xu,[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] *Department of Mathematics, 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 25 September 2013; Revised 22 November 2013; Accepted 25 November 2013; Published 2 January 2014

Academic Editor: Tao Huang

Sophora japonica Linn (Chinese Scholar Tree) is a shrub species belonging to the subfamily Faboideae of the pea family Fabaceae. In this study, RNA sequencing of *S. japonica* transcriptome was performed to produce large expression datasets for functional genomic analysis. Approximate 86.1 million high-quality clean reads were generated and assembled *de novo* into 143010 unique transcripts and 57614 unigenes. The average length of unigenes was 901 bps with an N50 of 545 bps. Four public databases, including the NCBI nonredundant protein (NR), Swiss-Prot, Kyoto Encyclopedia of Genes and Genomes (KEGG), and the Cluster of Orthologous Groups (COG), were used to annotate unigenes through NCBI BLAST procedure. A total of 27541 of 57614 unigenes (47.8%) were annotated for gene descriptions, conserved protein domains, or gene ontology. Moreover, an interaction network of unigenes in *S. japonica* was predicted based on known protein-protein interactions of putative orthologs of well-studied plant genomes. The transcriptome data of *S. japonica* reported here represents first genome-scale investigation of gene expressions in Faboideae plants. We expect that our study will provide a useful resource for further studies on gene expression, genomics, functional genomics, and protein-protein interaction in *S. japonica*.

1. Introduction

Sophora japonica Linn (Chinese Scholar Tree) is a shrub of the pea family Fabaceae. It grows into a lofty tree 10–20 m tall that produces a fine, dark brown timber. It is not only a kind of popular ornamental tree, but also a valuable nectar tree, offering delicious and healthy food. Moreover, dried flowers and buds of *Sophora japonica*, containing many kinds of components such as flavones, tetraglycosides, isoflavones, and isoflavone tetraglycosides [1], are used as useful herb to treat hemorrhoids and hematemesis in China, Japan, and Korea [2]. In spite of its medicinal and economic value, not much genomic or transcriptomic information is available for *S. japonica*. As of September 2013, only 74 nucleotide sequences and 35 proteins from *S. japonica* were available in GenBank. Hence, generation of genomic and transcriptome data is necessary to help further studies on *S. japonica*.

In the latest decade, the emergence of the next generation sequencing (NGS) technology offers a fast and effective way for generation of transcriptomic datasets in nonmodel species using various platforms such as Roche 454, Illumina HiSeq, and Applied Biosystems SOLiD [3–5]. Compared to the whole-genome sequencing, RNA-seq, which is considered as a cost-effective and ultra-high-throughput DNA sequencing technology, is a revolutionary advance in the functional genomic research [6]. In this approach, sequences of the expressed parts of the genome are produced [7] to identify genes [8] and explore the low abundance transcripts [9]. Due to the many advantages, RNA-seq is specifically attractive for nonmodel organisms without genomic sequences [10–13].

In this study, we used RNA-seq technology to investigate the transcriptome of S. *japonica* from three tissues. Using Illumina sequencing platform, a total of 86139654 reads

of S. *japonica* transcriptome were produced. Those were assembled into 57614 unigenes and annotated for functionality. Furthermore, the protein-protein interaction network of expressed genes in S. *japonica* was constructed. This is the first S. *japonica* and *Styphnolobium* genus transcriptome data generated by RNA-seq technology. The information provides a good resource for further gene expression, genomics, and functional studies in S. *japonica*.

2. Method

2.1. RNA Preparation and Sequencing. *S. japonica* (provided by the Yangzhou eight strange Memorial) was grown in an open-air place in Jiangsu Province, Eastern China. Total RNA was extracted using TRIzol method (Invitrogen) from three different tissues: tender shoots, young leaves, and flower buds. RNA was isolated from every tissue and mixed together in equal proportion for cDNA preparation.

The poly-A mRNA was isolated from the total RNA using poly-T oligo-attached magnetic beads (Illumina). After purification, fragmentation buffer (Ambion, Austin, TX) was added to digest the mRNA to produce small fragments. These small fragments were used as templates to synthesize the first-strand cDNA with superscript II (Invitrogen) and random hexamer primers. The synthesis of the second strand was performed in a solution containing the reaction buffer, dNTP, RNaseH, and DNA polymerase I using Truseq RNA sample preparation Kit. Next, these cDNA fragments were handled with end repair using T4 DNA polymerase, Klenow DNA polymerase, and T4 polynucleotide kinase (Invitrogen). Illumina's paired-end adapters were then ligated to the two ends of cDNA fragments. The adapter sequences were as follows: read1 adapter: AGATCGGAAGAGCACACGTC and read2 adapter: AGATCGGAAGAGCGTCGTGT. The products from this ligation reaction were electrophoresed on a 2% (w/v) agarose gel (certified low range ultragrade agarose from Bio-Rad) and purified according to appropriate size of DNA fragments suitable for Illumina sequencing. Then the sequencing library was constructed according to the protocol of the Paired-End Sample Preparation kit (Illumina). Sequencing was done with an Illumina Hiseq 2000. Raw read sequences are available in the Short Read Archive database from National Center for Biotechnology Information (NCBI) with the accession number SRR964825.

2.2. De Novo Assembly. After removal of adaptor sequences along with low quality reads and reads of larger than 5% unknown sequences, the resting were assembled into unitranscripts and unigenes by Trinity [14].

We used RSEM [15] to quantify expression levels of each unique transcript (see additional file 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/750961). Results were reported in units of TPM (transcripts per million). After counting fraction of each isoform, we used length × isoform percent as a standard to choose unigenes (see additional file 2).

2.3. Functional Annotation and Classification. All assembled unigenes, longer than 300 bps, were further analyzed to predict putative gene descriptions, conserved domains, gene ontology (GO) terms, and association with metabolic pathways. First of all, all the unigenes were searched in the protein databases including NCBI NR, Swiss-Prot, and clusters of orthologous groups (COG) [16] through BLASTALL procedure (ftp://ftp.ncbi.nih.gov/blast/executables/release/2.2.18/) with an E-value < $1.0E-6$. After obtaining the features of the best BLASTX hits from the alignments, putative gene names and "CDS" (coding DNA sequences) were determined. Subsequently, according to the NR annotation, we took advantage of Blast2GO [17] software to predict GO terms of molecular function, cellular component, and biological process. After obtaining GO annotation for all unigenes, GO functional classification of the unigenes performed using WEGO software [18] and exhibited the distribution of gene functions at the second level. Unigene sequences were also compared to the COG database to predict and classify possible gene functions based on orthologies. Association of unigenes with the KEGG pathways was determined using BLASTX against the Kyoto Encyclopedia of Genes and Genomes database [19]. The KEGG pathways annotation was performed in the KEGG Automatic Annotation Server (KAAS) (http:/www.genome.jp/tools/kaas/) [20].

To obtain the potential protein coding sequences from all unigenes, we first predicted all the open reading frames (ORFs). According to the BLASTP results against NR database, we chose the correct ORFs as potential protein coding sequences. And the longest ORFs from the unigenes without BLASTP results were considered as referential protein coding sequences (additional file 3).

2.4. Construction and Topological Analysis of Protein Interaction Network. The interaction network of unigenes in *S. japonica* was constructed in the form of nodes and edges where nodes represent genes and edges represent interactions between genes. First, we downloaded protein-protein interactions (PPI) and sequences of six species *Arabidopsis thaliana, Arabidopsis lyrata, Oryza sativa* subsp. *Japonica, Brachypodium distachyon, Populus trichocarpa*, and *Sorghum bicolor* from STRING database that is a precomputed database for the detection of protein-protein interactions [21]. Then, the protein sequences of genes from PPIs were searched against the unigenes datasets in our study to find homologies by TBLASTN (E-value < $1.0E-6$). The TBLASTN hits with identity >50% and covering query gene >80% were identified as the candidate interacting genes of the network. According to the known PPI network of the above six species, the interaction network of *S. japonica* was constructed using the homologous unigenes from the TBLASTN searches.

The topological features such as the degree distribution of nodes, degree correlation, clustering coefficient (C), and shortest path length (L) were determined for the resultant networks. To each node i of the network, we assigned a degree k_i, which is the number of its neighbors. We calculated the degree distribution of the giant component (i.e., the

Table 1: Summary of sequence assembly by trinity after Illumina sequencing.

	Number	Mean size (bp)	N50 size (bp)	Total nucleotides (bp)
Read	86139654	101	101	8700105054
Unique transcript	143010	1482	1155	211940997
Unigene	57614	901	545	51899592

probability $P(k)$ that a protein has k edges) [22] using the equation

$$P(k) = \frac{N(k)}{N}, \quad (1)$$

where N is the number of nodes and $N(k)$ is the number of nodes with degree k.

The degree correlation, which is characterized by analyzing the average degree of nearest neighbors $k_{nn,i}$ [23], is defined by

$$k_{nn,i} = \frac{1}{k_i} \sum_j a_{ij} k_j. \quad (2)$$

The clustering coefficient (C) was defined as the average probability with which two neighbors of a node were also neighbors to each other. For instance, if a node i had k_i links, and among its k_i nearest neighbors there were e_i links, then the clustering coefficient of i [23] was calculated using the equation

$$C_i = \frac{2e_i}{k_i(k_i - 1)}. \quad (3)$$

The shortest path length (L) between two nodes was defined as the minimum number of intermediate nodes that must be traversed to go from one node to another [23]. The average shortest path length was the shortest path length averaged over all the possible pairs of nodes in the network.

3. Result

3.1. De Novo Sequence Assembly of S. japonica Transcriptome. Total RNA from three different tissues (tender shoots, young leaves, and flower buds) was extracted and blended in equal proportions for Illumina sequencing. A total of 86.1 million high-quality clean reads with total of 8700105054 nucleotides (nt) sequences were produced with an average length of 101 bps for each short read (Table 1).

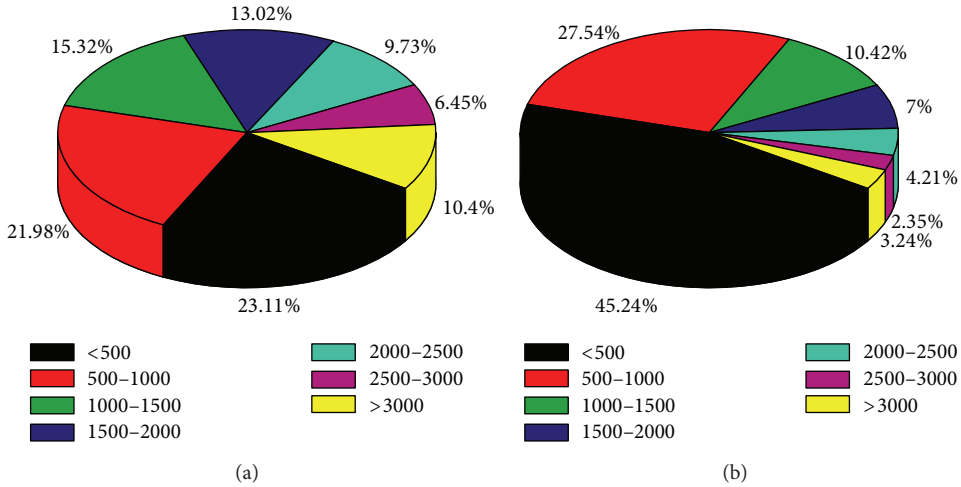
As a result of the absence of the genomic sequences of *S. japonica*, the transcripts were assembled *de novo* from all high-quality reads by Trinity [14]. A total of 143010 unique transcripts (UTs) were predicted from the clean sequence reads, with an average length of 1482 bps and an N50 of 1155 bps. The majority of UTs (33045) were between 100 and 500 bps, which accounted for 23.1% of total UTs shown in Figure 1(a). Then after removing redundancy, 57614 unigenes were generated with an average length of 901 bps. As shown in Figure 1(b), the length of the unigenes ranged from 300 bps to more than 3000 bps.

The quality score distribution across all bases and over all sequences was shown in additional files 4 and 5, revealing that most of the sequences have quality score larger than 30. To further evaluate the quality of the dataset, we compared the unigenes from *S. japonica* with other species using BLASTX (additional file 6). The result showed that more than half of unigenes that are having significant BLAST hits were mapped to soybean, which was consistent with our expectation.

3.2. Functional Annotation and Classification of S. japonica Transcriptome. In order to annotate the transcriptome of *S. japonica*, a total of 57614 unigenes were first examined against the NR database in NCBI using BLASTX with an E-value cut-off of $1e^{-6}$, which showed 27507 (47.7%) having significant BLAST hits (Table 2). The E-value distribution of significant hits revealed that 67.8% of matched sequences had strong homology (smaller than $1.0e-50$), while the other homologous sequences (32.2%) had E-values in the range of $1.0E-50$–$1.0E-6$ (Figure 2(a)). The distribution of sequence similarities represented that most of the BLASTX hits (95.3%) were in the range between 40% and 100%. Only 4.7% of hits had sequence similarity values less than 40% (Figure 2(b)).

The protein coding sequences of unigenes were also compared with the protein database at Swiss-Prot by BLASTX. A total of 20463 of 57614 unigenes (35.5%) showed hits at an E-value threshold of $\leq 1.0E-6$ (Table 2). More than half of the matched sequences (53.7%) had strong homologies with E-values of $\leq 1.0E-50$, and the remaining unigenes had E-values between $1.0E-50$ and $1.0E-6$ (Figure 2(c)). The distribution of sequence similarities against Swiss-Prot was different than that obtained against the NR database. While 75.0% of query sequences against Swiss-Prot had similarities between 40% and 100%, only 25.0% of sequences had strong homologies with <40% identity (Figure 2(d)). Thus by combining the results of sequence similarity searches from NR and Swiss-Prot database, we identified a final set of 27541 unigenes.

3.3. Gene Ontology (GO) Classification. GO terms were predicted for each assembled unigene to characterize functionality of gene products on the basis of their sequence similarities to known proteins in the Nr database. Of the 57614 unigenes of *S. japonica*, a total of 15063 unigenes were assigned to at least one of the three main GO categories: cellular component (11860, 20.6%), biological process (11643, 20.2%), and molecular function (11160, 19.4%). These GO terms were further subdivided into 51 sub-categories (Table 2, Figure 3, and additional file 7). Among these categories, the "cell," "cell part," "cellular process," "organelle," "metabolic process," "catalytic activity," and "binding" terms were found to have association with relatively more number of unigenes than other GO terms. The relative abundance of unigenes associated with cellular processes (9396) and metabolic processes (9010) in the biological processes category implied that the *S. japonica* tissues used in the study processed extensive metabolic activities.



(a) (b)

Figure 1: Overview of the *S. japonica* transcriptome assembly shown by pie graphs. The size distribution of the UTs (a) and unigenes (b) produced from *de novo* assembly of reads by trinity.

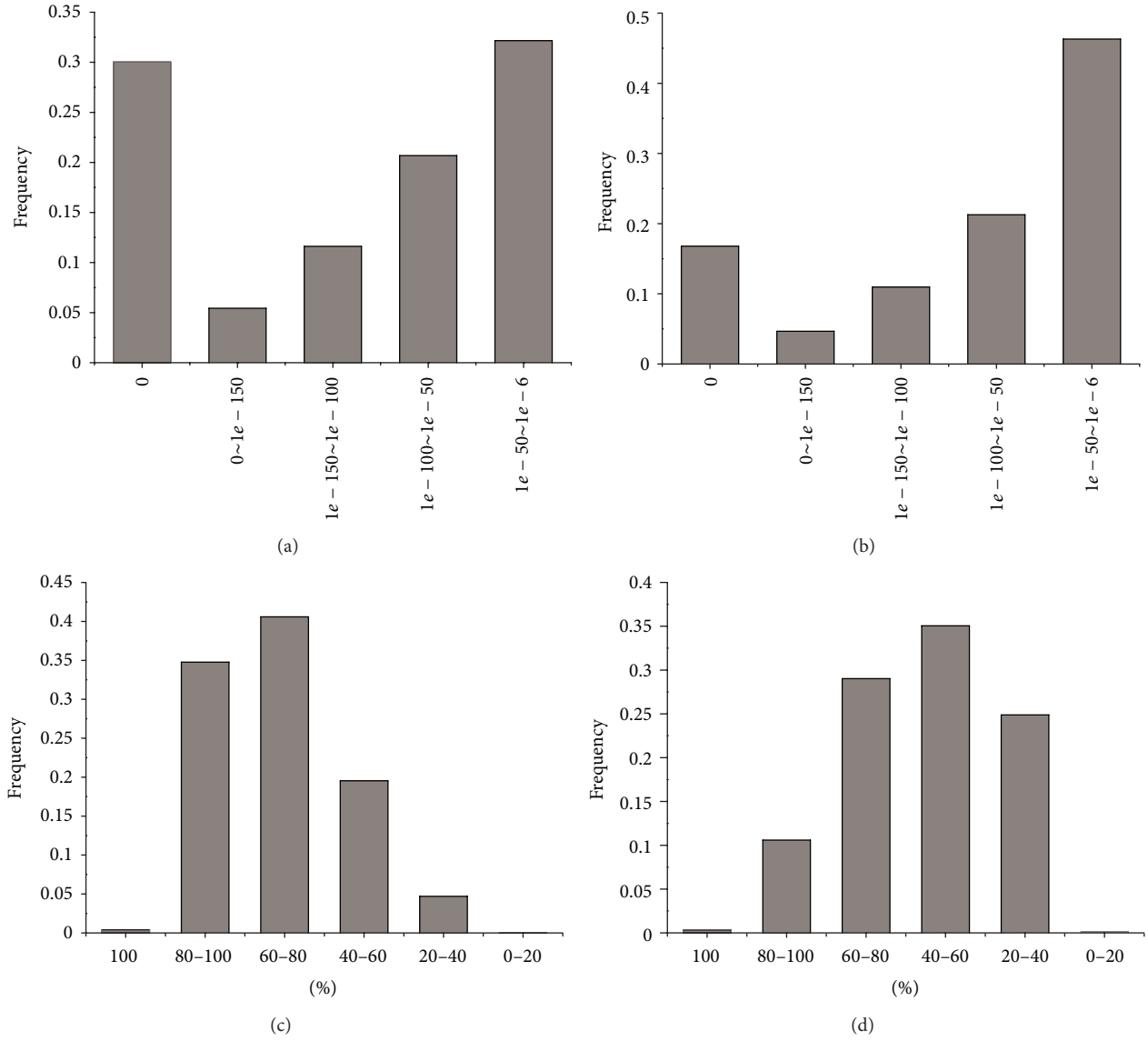


Figure 2: Unigene homology searches against NR and Swiss-prot databases. *E*-values proportional frequency distribution of BLAST hits against the NR database (a) and Swiss-prot database (c). Proportional frequency distribution of unigenes similarities against the NR database (b) and Swiss-Prot database (d) based on the best BLAST hits (*E*-value $\leq 1.0E - 5$).

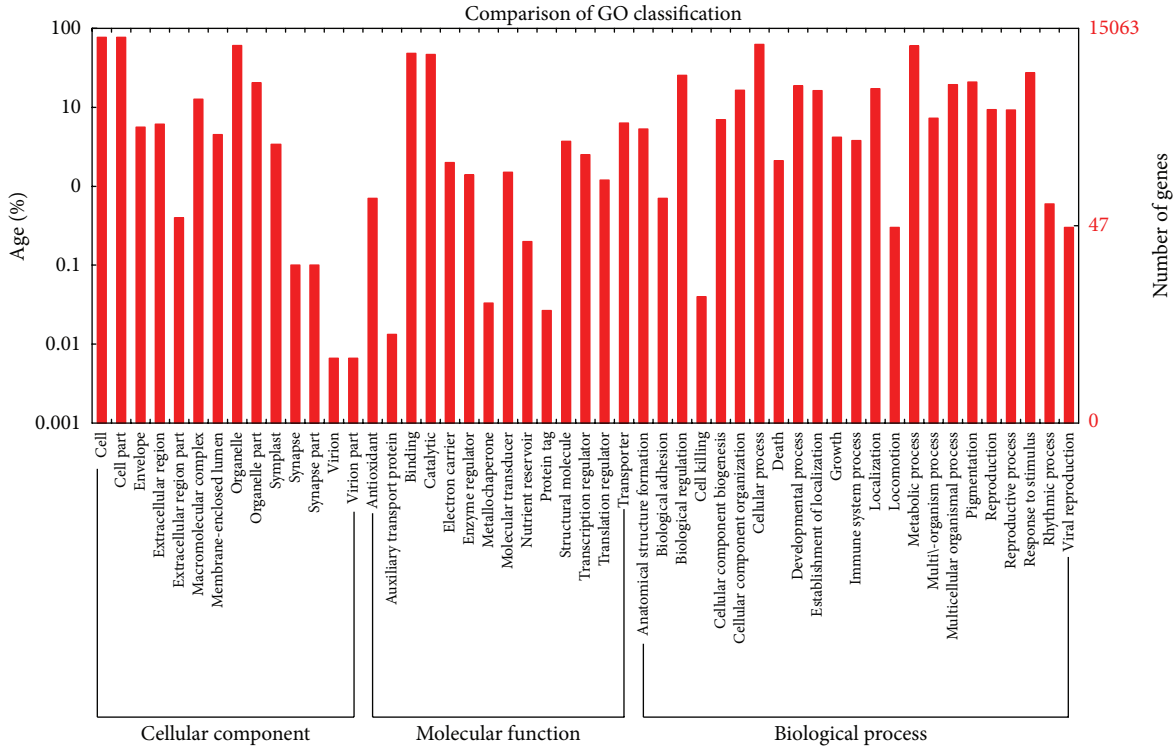


FIGURE 3: Gene ontology classification of the *S. japonica* transcriptome. Gene ontology (GO) terms associated with *S. japonica* unigenes based on significant hits against the NR database. They are summarized into three main GO categories (biological process, cellular component, and molecular function) and 51 subcategories.

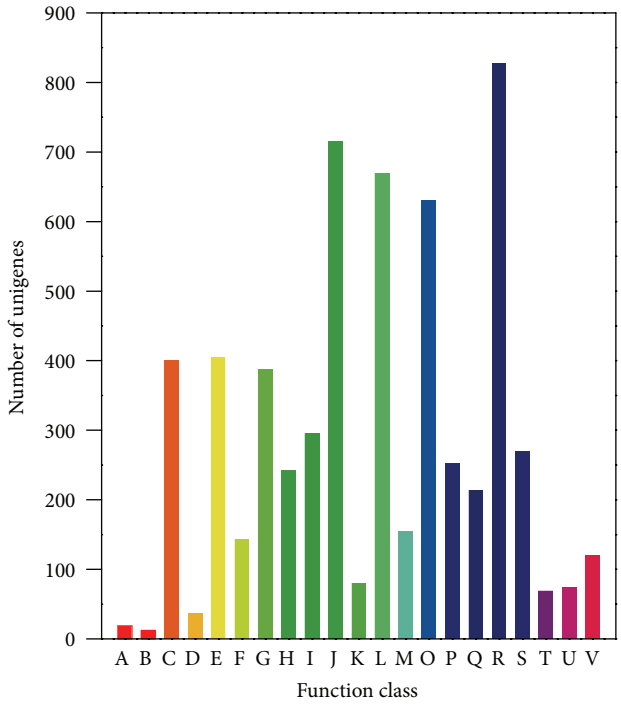
TABLE 2: Summary of annotation of *S. japonica* unigenes.

Category	Number	Percentage
Nr annotated unigenes	27507	47.74%
Swissprot annotated unigenes	20463	35.52%
GO classified unigenes	15063	26.14%
COG classified unigenes	5863	10.18%
KEGG classified unigenes	2869	4.98%

3.4. COG Classification. Cluster of Orthologous Groups (COG) database was used to classify the predicted proteins based on orthologous relationships of deduced amino sequences with 66 genomes, including bacteria, plants, and animals. Only individual proteins or groups of paralogs from at least three lineages involved in each COG were considered to be an ancient conserved domain. A total of 5863 *S. japonica* unigenes (10.2% of all unigenes) showed significant homology in the COG database. Since some of these unigenes were annotated with multiple COG functions, a total of 6012 functional annotations were predicted (E-value $\leq 1.0E-6$). Those were mapped to 21 COG clusters (Table 2, Figure 4, and additional file 8). The top five categories based on number of orthologies were (1) "general function prediction only" (13.8%); (2) "translation, ribosomal structure, and biogenesis" (11.9%); (3) "replication, recombination, and repair" (11.1%); (4) "posttranslational modification, protein turnover, and chaperones" (10.5%); and (5) "amino acid transport and metabolism" (6.7%). The two categories comprising "RNA processing and modification" and "chromatin structure and dynamics" consisted of 19 and 12 unigenes (0.5%), respectively, representing the two small COG classifications.

3.5. KEGG Pathway Mapping. To further predict the metabolic pathway in *S. japonica,* the assembled unigenes were annotated with corresponding enzyme commission (EC) numbers in the KAAS using *Arabidopsis thaliana* and *Oryza sativa* as references. A total of 2869 unigenes were mapped to 309 pathways corresponding to six KEGG modules: metabolism, genetic information processing, environmental information processing, cellular processes, and organismal systems and human diseases (additional file 9). Metabolic pathways had the largest number of unigenes (2155 members, 47.2%), followed by ribosome (158 members, 5.5%, ko03010), biosynthesis of amino acids (139 members, 4.8%, ko01230), carbon metabolism (130 members, 4.5%, ko01200), spliceosome (129 members, 4.5%, ko03040), protein processing in the endoplasmic reticulum (123 members, 4.3%, ko04141), plant hormone signal transduction (122 members, 4.3%, ko04075), purine metabolism (107 members, 3.7%, ko00230), and RNA transport (100 members, 3.5%, ko03013).

In conclusion, 27541 unigenes were annotated using NR, Swiss-Prot, COG, and KEGG databases. These unigenes had BLASTX scores with E-values $\leq 1.0E-6$. Among these, 1561 unigenes showed hits in all the four public databases (NR,



A: RNA processing and modification
B: chromatin structure and dynamics
C: energy production and conversion
D: cell cycle control, cell division, and chromosome partitioning
E: amino acid transport and metabolism
F: nucleotide transport and metabolism
G: carbohydrate transport and metabolism
H: coenzyme transport and metabolism
I : lipid transport and metabolism
J: translation, ribosomal structure, and biogenesis
M: cell wall/membrane/envelope biogenesis
O: posttranslational modification, protein turnover, and chaperones
P: inorganic ion transport and metabolism
Q: secondary metabolites biosynthesis, transport, and catabolism
K: transcription
L: replication, recombination, and repair
R: general function prediction only
S: function unknown
T: signal transduction mechanisms
U: intracellular trafficking, secretion, and vesicular transport
V: defense mechanisms

Figure 4: COG functional classification of the *S. japonica* transcriptome. Of 57614 unigenes in the NR database, 5863 unigenes show significant homologies to the COGs database (*E*-value ≤ 10−6) which were classified into 21 COG categories.

Swiss-Prot, COG, and KEGG) providing the best functional annotations of those unigenes (Table 2). These annotations provide a valuable resource to investigate further processes, structures, functions, and pathways of *S. japonica* in future studies.

3.6. Construction and Topological Analysis of Protein-Protein Interaction Network in S. japonica. The interaction network was constructed using the annotated unigenes of *S. japonica* in comparison with genes with at least one known protein-protein association and links of the six genomes in STRING database (additional file 10 and additional file 11). The network included one giant component and 88 small components. The giant component consisted of 1887 nodes connected via 7634 edges. Figure 5(a) showed the degree distribution $P(k) = 0.23\,\mathrm{K}^{-0.94}$ (the least square fit of associations) which implies the sale-free characteristics of the component. The degree-correlation of the giant component is shown in Figure 5(b). The decay behavior of *knn* with *k* suggests the disassortative mixing of nodes. We plotted the clustering coefficient of a node with *k* links in Figure 5(c) yet. The guideline is $C(k) \sim k^{-1}$, the scaling law which reflects the hierarchy of the giant component. Besides, we compared the clustering coefficient *C* and the shortest path length *L* of the giant component with *Erdős-Rényi*, *Watts-Strogatz,* and *Barabási-Albert* models of the same nodes *N* and links *E*. The data in Table 3 shows high clustering coefficient and small path length. Those suggest the small-world properties of the network.

Table 3: The average clustering coefficient (*C*) and shortest path length (*L*) of the giant component of the unigenes of *S. japonica* measured using Erdős-Rényi, Watts-Strogatz, and Barabási-Albert network models.

Item	*C*	*L*
Giant component	4.68*E* − 03	5.01
Erdős-Rényi	1.49*E* − 04	5.84
Watts-Strogatz	2.20*E* − 06	2.00
Barabási-Albert	4.34*E* − 04	3.35

4. Discussion

Sophora japonica Linn is an economically important species for several reasons. It is commonly used to afforest cites and highways for their adaptability to ecology and environment. It also provides useful products such as honey and lumber for human use [2]. Apart from such ecological and economical values of pagoda tree, it has a unique mythological importance to Chinese people. The pagoda tree mentioned in the famous Chinese idiom story "A Fond Dream of Nanke" is believed to still present in the yard of the Yangzhou eight strange Memorial at present, in Jiangsu Province. This story tells that more than one thousand years ago, a person named Nanke drunk and rested against the pagoda tree having a dream. In his dream, he became the prime minister of the kingdom of pagoda tree. After waking up, he found that the kingdom of pagoda tree was the nest of ants under the pagoda tree. Nowadays, people believe that the tree is over 2000 years old. However, very little research has been done with this important species to understand its genome. Recently, high-throughput RNA sequencing has offered a new avenue to generate abundant sequence information from any organism [24, 25]. The data obtained from RNA-seq projects are also helpful in inferring the basic biological, molecular, and cellular processes [19, 20]. Genomes of many plant species have been studied by *de novo* transcriptome analysis, such as willow [26], *Cocos nucifera* [27], tea plant

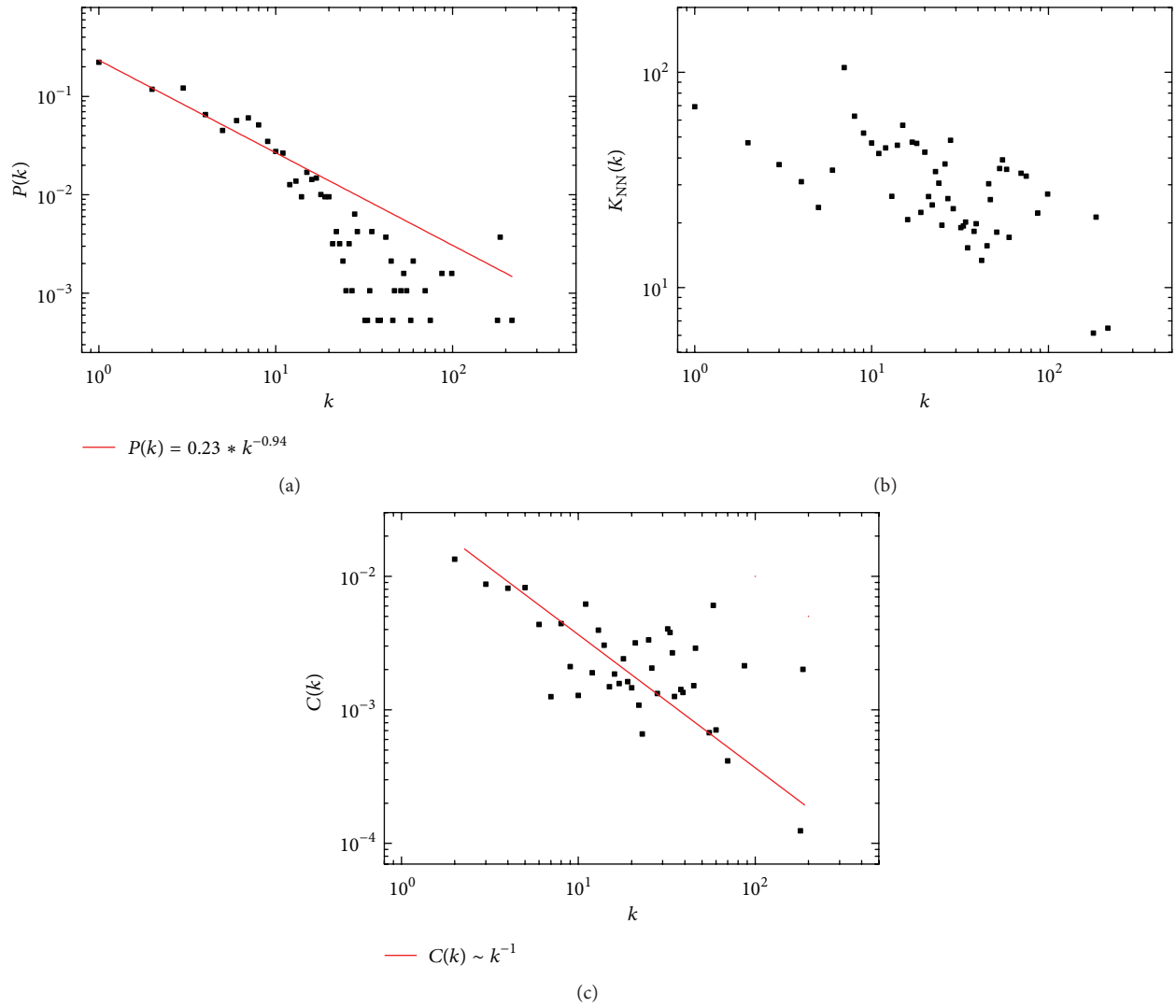


FIGURE 5: The topological analysis of the giant component of *S. japonica* protein interaction with 1887 nodes and 7634 edges. (a) Log-log plots of the node degree distribution with a power-law fit (red line). (b) Average nearest-neighbor degree k_{nn} as a function of the node degree k. (c) Log-log plots of the average clustering coefficient C as a function k with a guideline $C(k) \sim k^{-1}$ (red line).

[10], and pineapple [28]. In this study, we used Illumina RNA-seq technology to sequence the *Sophora japonica* plant transcriptome and predicted a large number of expressed genes in *S. japonica*. We obtained 8.7 G bps coverage with 86.1 million high-quality clean reads. Using *de novo* software Trinity, we generated 57614 unigenes. Our results revealed that 27541 unigenes (47.8% of all assembled unigenes) were functionally annotated and involved in different biological processes.

Using the sequences of the predicted unigenes, we constructed a protein-protein interaction network to understand gene interactions in *S. japonica*. We identified a giant and 88 small components of the network. The best fit of the degree distribution of the giant component demonstrated that it was a scale-free network in which a few proteins interacted with high connectivity [29]. Like other biological networks, the giant component displays disassortative mixing that ensures connection of high-degree nodes with low-degree nodes. It is likely that the disassortativity may reduce the proportion of the important edges among hubs and increase the stability of biological networks when compared to the assortative network.

In addition, the giant component of the network also exhibited small-world properties including the high clustering coefficient as well as the smaller and shortest path length (Table 3), suggesting that the neighbors of one node have close associations among each other in the network. The smaller shortest path length is an indicative of minimal distance between a node and its target to minimize energy involved in interactions between proteins. At the same time, the scaling law of the average clustering coefficient as a function of the degree (Figure 5(c)) indicates the hierarchical structure which reflects the evolutionary patterns associated with various organizational levels of the network.

The combination of many local variations, which affect the small but highly interacted nodes, slowly affects the properties of the larger but less interacted nodes [30]. Such a process during evolution ensures both stability and low energy consumption in an efficient protein-protein interaction.

5. Conclusion

In this study, we applied Illumina RNA sequencing and *de novo* assembly approach to study the *S. japonica* transcriptome for the first time. Totally, about 86.1 million reads assembled into 57614 unigenes were generated with an average length of 1321 bps. Among these unigenes, 27541 unigenes obtained annotation with gene descriptions from NR, Swiss-Prot, COG, and KEGG databases. This study demonstrated that the RNA-seq technology could be used as a rapid and efficient method for *de novo* transcriptome analysis of non-model plant organisms that provides a good resource of gene expression data for further analysis. A protein-protein interaction network of expressed genes was constructed in *S. japonica*. The topological analysis revealed that degree correlation of the giant component was disassortative and had small-world properties. This result implied that the protein-protein interaction network in *S. japonica* might have resulted from a long-term evolution to ensure both stability and low energy consumption protein-protein interactions.

Abbreviations

UT: Unique transcript
NR: NCBI non-redundant protein
COG: Cluster of Orthologous Groups
GO: Gene ontology
KEGG: Kyoto Encyclopedia of Genes and Genomes database
KAAS: KEGG Automatic Annotation Server
BLAST: Basic Local Alignment Search Tool
PPI: Protein-protein network.

Conflict of Interests

The authors declare that they have no conflict of interests.

Authors' Contribution

Liucun Zhu and Ying Zhang contributed equally to this work. Liucun Zhu conceived and designed the study, carried out data analysis, interpreted the entire results, and drafted the paper. Ying Zhang carried out data analysis and drafted the paper. Wenna Guo helped to draft the paper. Xin-Jian Xu carried out data analysis and helped to draft the paper. Qiang Wang participated in the design of the study and interpreted the results. All authors read and approved the final paper.

Acknowledgments

This work was supported by Grants from the Science and Technology Commission of Shanghai Municipality (12ZR1444200 and 13ZR1416800), National Natural Science Foundation of China (61103075), and Foundation for the Author of National Excellent Doctoral Dissertation of PR China (201134).

References

[1] J. M. Kim and H. S. Yun-Choi, "Anti-platelet effects of flavonoids and flavonoid-glycosides from *Sophora japonica*," *Archives of Pharmacal Research*, vol. 31, no. 7, pp. 886–890, 2008.

[2] H. Ishida, T. Umino, K. Tsuji, and T. Kosuge, "Studies on the antihemostatic substances in herbs classified as hemostatics in traditional Chinese medicine. I. On the antihemostatic principles in *Sophora japonica* L.," *Chemical and Pharmaceutical Bulletin*, vol. 37, no. 6, pp. 1616–1618, 1989.

[3] T. L. Parchman, K. S. Geist, J. A. Grahnen, C. W. Benkman, and C. A. Buerkle, "Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery," *BMC Genomics*, vol. 11, no. 1, article 180, 2010.

[4] M. Dassanayake, J. S. Haas, H. J. Bohnert, and J. M. Cheeseman, "Shedding light on an extremophile lifestyle through transcriptomics," *The New Phytologist*, vol. 183, no. 3, pp. 764–775, 2009.

[5] F. Alagna, N. D'Agostino, L. Torchia et al., "Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development," *BMC Genomics*, vol. 10, article 399, 2009.

[6] K. O. Mutz, A. Heilkenbrinker, M. Lonne, J. G. Walter, and F. Stahl, "Transcriptome analysis using next-generation sequencing," *Current Opinion in Biotechnology*, vol. 24, no. 1, pp. 22–30, 2013.

[7] T. T. Torres, M. Metta, B. Ottenwälder, and C. Schlötterer, "Gene expression profiling by massively parallel sequencing," *Genome Research*, vol. 18, no. 1, pp. 172–177, 2008.

[8] M. S. Clark, M. A. S. Thorne, F. A. Vieira, J. C. R. Cardoso, D. M. Power, and L. S. Peck, "Insights into shell deposition in the Antarctic bivalve Laternula elliptica: gene discovery in the mantle transcriptome using 454 pyrosequencing," *BMC Genomics*, vol. 11, no. 1, article 362, 2010.

[9] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[10] C. Shi, H. Yang, C. Wei et al., "Deep sequencing of the Camellia sinensis transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds," *BMC Genomics*, vol. 12, article 131, 2011.

[11] E. Novaes, D. R. Drost, W. G. Farmerie et al., "High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome," *BMC Genomics*, vol. 9, article 312, 2008.

[12] P. Gayral, J. Melo-Ferreira, S. Glemin et al., "Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap," *PLoS Genetics*, vol. 9, no. 4, Article ID e1003457, 2013.

[13] L. Wan, J. Han, M. Sang et al., "De novo transcriptomic analysis of an oleaginous microalga: pathway description and gene discovery for production of next-generation biofuels," *PLoS ONE*, vol. 7, no. 4, Article ID e35142, 2012.

[14] M. G. Grabherr, B. J. Haas, M. Yassour et al., "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nature Biotechnology*, vol. 29, no. 7, pp. 644–652, 2011.

[15] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, article 323, 2011.

[16] R. L. Tatusov, D. A. Natale, I. V. Garkavtsev et al., "The COG database: new developments in phylogenetic classification of proteins from complete genomes," *Nucleic Acids Research*, vol. 29, no. 1, pp. 22–28, 2001.

[17] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research," *Bioinformatics*, vol. 21, no. 18, pp. 3674–3676, 2005.

[18] J. Ye, L. Fang, H. Zheng et al., "WEGO: a web tool for plotting GO annotations," *Nucleic Acids Research*, vol. 34, pp. W293–W297, 2006.

[19] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The KEGG resource for deciphering the genome," *Nucleic Acids Research*, vol. 32, pp. D277–D280, 2004.

[20] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa, "KAAS: an automatic genome annotation and pathway reconstruction server," *Nucleic Acids Research*, vol. 35, pp. W182–W185, 2007.

[21] 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.

[22] R. M. Ferreira, J. L. Rybarczyk-Filho, R. J. Dalmolin et al., "Preferential duplication of intermodular hub genes: an evolutionary signature in eukaryotes genome networks," *PloS ONE*, vol. 8, no. 2, Article ID e56579, 2013.

[23] X. Xu and M. Zhou, "Rank-dependent deactivation in network evolution," *Physical Review E*, vol. 80, no. 6, Article ID 066105, 2009.

[24] H. Guo, J. A. Xian, A. L. Wang, C. X. Ye, and Y. T. Miao, "Trascriptome analysis of the Pacific white shrimp Litopenaeus vannamei exposed to nitrite by RNA-seq," *Fish and Shellfish Immunology*, vol. 35, no. 6, pp. 2008–2016, 2013.

[25] S. M. Gross, J. A. Martin, J. Simpson, M. J. Abraham-Juarez, Z. Wang, and A. Visel, "De novo transcriptome assembly of drought tolerant CAM plants, Agave deserti and Agave tequilana," *BMC Genomics*, vol. 14, article 563, 2013.

[26] J. Liu, T. Yin, N. Ye et al., "Transcriptome analysis of the differentially expressed genes in the male and female shrub willows (Salix suchowensis)," *PloS ONE*, vol. 8, no. 4, Article ID e60181, 2013.

[27] H. Fan, Y. Xiao, Y. Yang et al., "RNA-Seq analysis of Cocos nucifera: transcriptome sequencing and de novo assembly for subsequent functional genomics approaches," *PloS ONE*, vol. 8, no. 3, Article ID e59997, 2013.

[28] W. D. Ong, L. Y. Voo, and V. S. Kumar, "De novo assembly, characterization and functional annotation of pineapple fruit transcriptome through massively parallel sequencing," *PloS ONE*, vol. 7, no. 10, Article ID e46937, 2012.

[29] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[30] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 132724, 6 pages
http://dx.doi.org/10.1155/2013/132724

Research Article

Prediction of Drugs Target Groups Based on ChEBI Ontology

Yu-Fei Gao,[1] Lei Chen,[2] Guo-Hua Huang,[3] Tao Zhang,[3] Kai-Yan Feng,[4] Hai-Peng Li,[5] and Yang Jiang[1]

[1] *Department of Surgery, China-Japan Union Hospital of Jilin University, Changchun 130033, China*
[2] *College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*
[3] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[4] *Beijing Genomics Institute, Shenzhen Beishan Industrial Zone, Shenzhen 518083, China*
[5] *CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*

Correspondence should be addressed to Yang Jiang; jy7555@163.com

Received 15 September 2013; Accepted 28 October 2013

Academic Editor: Tao Huang

Most drugs have beneficial as well as adverse effects and exert their biological functions by adjusting and altering the functions of their target proteins. Thus, knowledge of drugs target proteins is essential for the improvement of therapeutic effects and mitigation of undesirable side effects. In the study, we proposed a novel prediction method based on drug/compound ontology information extracted from ChEBI to identify drugs target groups from which the kind of functions of a drug may be deduced. By collecting data in KEGG, a benchmark dataset consisting of 876 drugs, categorized into four target groups, was constructed. To evaluate the method more thoroughly, the benchmark dataset was divided into a training dataset and an independent test dataset. It is observed by jackknife test that the overall prediction accuracy on the training dataset was 83.12%, while it was 87.50% on the test dataset—the predictor exhibited an excellent generalization. The good performance of the method indicates that the ontology information of the drugs contains rich information about their target groups, and the study may become an inspiration to solve the problems of this sort and bridge the gap between ChEBI ontology and drugs target groups.

1. Introduction

Identification of target proteins of drugs is of importance in the drug discovery pipeline [1] because drugs exert their functions by hitting some proteins, that is, their target proteins, in human tissues. On the other hand, in addition to their therapeutic effects, most of the drugs have some undesirable side effects caused also by hitting some target proteins. If a drug with unclear undesirable side effects was brought into the market, it is a potential hazard to both pharmaceutical companies and their consumers. Thus, studying the target proteins of a drug is highly beneficial to the treatment of diseases and reduction of side effects. However, identification of drugs target proteins by experiments needs lots of time and money. It is necessary to establish effective computational methods to tackle this problem which can provide useful references.

Many efforts have been made to identify drugs target proteins in the past few years, such as docking simulations [2, 3], literature text mining [4], combination of chemical structure and protein structural information or functional information [5–8], side effect similarity [9], and so forth. In this paper, we attempted a novel method using the ontology information of compounds, which was similar to gene ontology of proteins, to identify drugs target proteins. With the discovery of novel candidate drugs, the quantity of all candidate pairs of drugs and target proteins is tremendously large, preventing researchers to carry out an exhaustive search of drugs target proteins. In view of this, a necessary step is to establish an effective method to reduce the candidate proteins for each query drug, that is, reducing the search space by deducing the kind of functions a drug may have. According to the data in KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/) [10],

Table 1: The number of drugs in each group.

Index	Target group	Number of drugs		
		Training dataset	Test dataset	Total
1	G Protein-coupled Receptors	272	35	307
2	Nuclear Receptors	82	13	95
3	Ion Channels	109	9	118
4	Enzymes	325	31	356
—	Total	788	88	876

the target proteins of drugs could be divided into the following five groups: (1) G Protein-coupled Receptors, (2) Cytokine Receptors, (3) Nuclear Receptors, (4) Ion Channels, and (5) Enzymes. If one can establish a method to correctly predict the target groups of a query drug, the possible target proteins would be limited only to the predicted group, facilitating further analyses.

In the past few years, many novel compounds have been discovered with the advance of combinatorial chemistry. To record these compounds, some online databases are established, such as KEGG [10], STITCH (Search Tool for Interactions of Chemicals) [11], and ChEBI (Chemical Entities of Biological Interest) [12], from which users can retrieve all sorts of information about the compounds, for example, their structures, activities, reactions, and so on. Furthermore, their information can also be used to infer the attributes of novel compounds [5, 7, 8, 13–15]. In the paper, we employed compound ontology information, named as ChEBI ontology, to infer the target group of a novel drug, that is, a predictor that was built to predict the target group of drugs based on ChEBI ontology. A benchmark dataset consisting of 876 drugs was established by collecting data in KEGG, from which a training dataset and a test dataset were obtained by splitting the data. Jackknife test demonstrates an overall prediction accuracy of 83.12% and independent test achieves a prediction accuracy of 87.50%, indicating that the predictor has excellent generalization. We hope that the predictor may facilitate the discovery of new therapeutic or undesirable effects of existing drugs.

2. Materials and Methods

2.1. Dataset. 2,795 drugs were retrieved from Chen and Zeng's study [8], which were downloaded from KEGG (http://www.genome.jp/kegg/) [10]. According to their target proteins, these drugs were classified into the following five groups: (1) G Protein-coupled Receptors, (2) Cytokine Receptors, (3) Nuclear Receptors, (4) Ion Channels, and (5) Enzymes. We then screened the data with the following rules: drugs without ChEBI ontology information were excluded, resulting in 895 drugs; drugs belonging to more than one group were excluded, resulting in 879 drugs; and because there were only 3 drugs in Cytokine Receptors—not enough to build an effective prediction model on the group, these drugs and the group were also excluded. Thus, we obtained a benchmark dataset **S** containing 876 drugs allocated into four groups. The distribution of these drugs is listed in column 5 of Table 1. The codes of the drugs in each group are available in Supplementary Material I available online at http://dx.doi.org/10.1155/2013/132724.

To evaluate the generalization of the predictor, the benchmark dataset **S** was divided into a training dataset $\mathbf{S}_{tr}$ and a test dataset $\mathbf{S}_{te}$, where $\mathbf{S}_{te}$ was constructed by randomly selecting 88 (10%) drugs in **S** and the rest in **S** comprised $\mathbf{S}_{tr}$. The number of drugs in each group in the training and test dataset was listed in columns 3 and 4 of Table 1, respectively.

2.2. Prediction Based on ChEBI Ontology. The term "ontology" derived from philosophy, meaning the theory or study of the basic characteristics of all reality. Since gene ontology, the established ontology information about proteins, is deemed as a very useful tool for investigating various attributes of proteins [16–21], similarly, the ontology information of compounds may also facilitate the study of various attributes of compounds.

ChEBI, a well-known compound database, contained some important ontology information about compounds named as ChEBI ontology [12]. It consists of four subontologies: (1) Molecular Structure, (2) Biological Role, (3) Application, and (4) Subatomic Particle, which may be suitable for the prediction of various attributes of compounds. The information of ChEBI ontology was retrieved from ftp://ftp.ebi.ac.uk/pub/databases/chebi/ontology/ ("chebi.obo", July 2012). Ontologies are controlled vocabularies which can be conceived as graph-theoretical structures consisting of "terms" forming the node set and "relations" of two terms forming the edge set [22]. Based on the "terms" and "relations" (including "is_a" and "relationship") in the obtained file, a graph with 31,813 nodes and 64,514 edges was established. As for the two terms, the smaller the distance is between them, the more intimate the "relations" are implicated between them. Thus, the distance of terms t and t', denoted by $d(t, t')$, would be used to measure the relationship of compounds.

For two compounds c_1 and c_2, $T(c_1)$ and $T(c_2)$ are an ontology term set of c_1 and c_2, respectively. The following formula was used to measure the functional relationship of c_1 and c_2:

$$Q(c_1, c_2) = \text{Mean}\{d(t_1, t_2) : t_1 \in T(c_1), t_2 \in T(c_2)\}. \quad (1)$$

The smaller the $Q(c_1, c_2)$ is, the stronger the functional relationship would be shared by c_1 and c_2.

For a query drug d_q, its target group was predicted according to the following steps.

(i) Find drugs in the training set $\mathbf{S}'$, say, $d_1, d_2, \ldots, d_s$, such that

$$Q\left(d_q, d_1\right) = Q\left(d_q, d_2\right) = \cdots = Q\left(d_q, d_s\right) = \min_{d \in \mathbf{S}'} Q\left(d_q, d\right). \quad (2)$$

(ii) The target groups of $d_1, d_2, \ldots, d_s$ were put into a voting system.

(iii) The target group with the most votes is deemed to be the predicted target group of d_q. Note that if more than one target group is receiving the most votes, randomly select one of them as the predicted result.

2.3. Prediction Based on Chemical Interaction. In recent years, the idea of "systems biology" is penetrating into the prediction of various attributes of proteins and compounds and is considered to be very useful [13, 14, 23–25]. The constructed methods were all based on the fact that interactive proteins and compounds often share common features. To define the interactive compounds, we downloaded the chemical interaction files from STITCH ((chemical_chemical.links.detailed.v3.1.tsv.gz) http://stitch.embl.de/download/chemical_chemical.links.detailed.v3.1.tsv.gz, http://stitch.embl.de/) [11], a well-known database including the interaction information of proteins and chemicals. In the obtained file, each interaction is composed of two chemicals and five kinds of scores. In detail, the first four kinds of scores are estimated according to the structures, activities, reactions, and cooccurrence in the literature of two chemicals [11], while the last kind of score is calculated by integrating the aforementioned four kinds of scores. It is reasonable to use the last kind of score to indicate the interactivity of two chemicals. Thus, it was adopted here to indicate the interactivity of two chemicals; that is, two chemicals are interactive chemicals if and only if the last kind of score of the interaction between them is greater than 0. For the later formulation, we denote the score of chemicals c_1 and c_2 by $I(c_1, c_2)$. In particular, if c_1 and c_2 are noninteractive chemicals, we set $I(c_1, c_2) = 0$.

As described above, the interactive compounds share common features with higher possibility than noninteractive ones. In view of this, the target group of a query drug d_q can be determined by its interactive compounds in the training set. The detailed procedure of the method is almost similar to that of the method in Section 2.2. Now, instead of (2), we used the following formula to select drugs in the first step

$$I\left(d_q, d_1\right) = I\left(d_q, d_2\right) = \cdots = I\left(d_q, d_s\right) = \max_{d \in \mathbf{S}'} I\left(d_q, d\right). \quad (3)$$

2.4. Jackknife Test. In statistical prediction, there are three cross-validation methods: independent dataset test, subsampling (or k-fold crossover) test, and jackknife test [13], which are often used to evaluate the performance of various classifiers. Among them, jackknife test is deemed the least arbitrary [13] because the test sample and training samples are always open. Furthermore, the classifier evaluated by jackknife test can always provide a unique result for a given dataset. Accordingly, it has been widely used to examine the performance of various classifiers in recent years [13, 26–36]. Here, we also adopted it to evaluate the current method.

Table 2: The correct prediction rates of the method based on ChEBI ontology.

Target group	Training dataset	Test dataset
G Protein-coupled Receptors	93.38%	100%
Nuclear Receptors	73.17%	69.23%
Ion Channels	60.55%	55.56%
Enzymes	84.62%	90.32%
Overall	83.12%	87.50%

3. Results and Discussions

As described in Section 2.1, the benchmark dataset $\mathbf{S}$ was divided into two datasets, $\mathbf{S}_{tr}$ and $\mathbf{S}_{te}$, consisting of 788 and 88 drugs, respectively. The method based on ChEBI ontology was applied to predict the target groups of drugs in these two datasets. The detailed results were given in the following sections.

3.1. Performance of the Predictor on the Training Dataset. As for the 788 drugs in the training dataset $\mathbf{S}_{tr}$, the predictor based on ChEBI ontology was evaluated by jackknife test. The prediction results were listed in column 2 of Table 2, from which we can see that the prediction accuracies for each target group were 93.38%, 73.17%, 60.55%, and 84.62%, respectively, while the overall prediction accuracy was 83.12%. Since there are four target groups investigated by the study, the average correct rate would be 25% if one identifies drugs target groups in $\mathbf{S}_{tr}$ by random guesses, which is much lower than the overall prediction accuracy obtained by our method. Compared to the results in Chen and Zeng's work [8], in which a similarity-based method was proposed to predict drugs target groups, our results are also very competitive because the prediction accuracies in their work were less than 80%. All of these suggest that the proposed predictor performs fairly well on the training dataset.

3.2. Performance of the Predictor on the Test Dataset. As for the 88 drugs in the test dataset $\mathbf{S}_{te}$, the predictor was modeled only based on the training dataset $\mathbf{S}_{tr}$ without involving $\mathbf{S}_{te}$. The prediction accuracies for each group and the overall accuracy were listed in column 3 of Table 2. It can be seen that the prediction accuracies for each group were 100%, 69.23%, 55.56%, and 90.32%, respectively, while the overall prediction accuracy was 87.50%, which is even better than that of the training dataset, indicating that the predictor has an excellent generalization.

Table 3: The prediction accuracies of the method based on ChEBI ontology and chemical interaction on the benchmark dataset evaluated by jackknife test.

Target group	Prediction based on ChEBI ontology	Prediction based on chemical interaction
G Protein-coupled Receptors	94.46%	83.71%
Nuclear Receptors	76.84%	82.11%
Ion Channels	61.02%	77.97%
Enzymes	86.24%	87.08%
Overall	84.70%	84.13%

Table 4: The 13 drugs which are closest to "D00146" in $\mathbf{S}_{tr}$.

Drug	Target group	Ontology term
D00089	G Protein-coupled Receptors	CHEBI:7872
D00101	G Protein-coupled Receptors	CHEBI:9937
D00176	G Protein-coupled Receptors	CHEBI:35940
D00284	G Protein-coupled Receptors	CHEBI:3901
D00291	G Protein-coupled Receptors	CHEBI:4450
D00410	Enzymes	CHEBI:44241
D00994	Enzymes	CHEBI:4759
D01002	G Protein-coupled Receptors	CHEBI:4025
D01163	G Protein-coupled Receptors	CHEBI:31554
D02783	G Protein-coupled Receptors	CHEBI:337298
D07431	G Protein-coupled Receptors	CHEBI:64628
D07759	G Protein-coupled Receptors	CHEBI:4024
D07905	G Protein-coupled Receptors	CHEBI:4822

3.3. Comparison of the Predictors Based on ChEBI Ontology and Chemical Interaction. The method based on chemical interaction described in Section 2.3 is popular for predicting various attributes of compounds [13, 14]. Thus, we compared the performances of these two methods in identifying drugs target groups as follows.

To compare the methods with the same datasets, all samples in the benchmark dataset **S** were used to make prediction; that is, two predictors were conducted to predict the target groups of samples in **S** evaluated by jackknife test. The prediction results obtained by these two methods were listed in Table 3. It is observed that the overall prediction accuracy for the predictor using ChEBI ontology was 84.70%, which is a little higher than that of the method using chemical interaction. In detail, the prediction accuracy for the target group "G Protein-coupled Receptors" obtained by the proposed method was much higher than the corresponding accuracy obtained by the method based on chemical interaction, the prediction accuracies for the target group "Enzymes" obtained by these two methods were almost the same, while the prediction accuracies for the rest two target groups obtained by the proposed method were lower than those obtained by the method based on chemical interaction. All of these indicate that the two predictors perform at the same level on the benchmark dataset **S**. Thus, it can be inferred that strong links may exist between ChEBI ontology and chemical interactions.

3.4. Analysis of the Relationship of Drugs Ontology Information and Their Target Group. From Sections 3.1–3.3, the ChEBI ontology information of compounds connects strongly with their targets' information. In this section, some examples are picked up to confirm this and to reinforce the understanding of using ChEBI to categorize drugs into their target groups.

The drug "D00146" is a sample in the training dataset $\mathbf{S}_{tr}$. Its target group is "G Protein-coupled Receptors" and it hits the ontology term "CHEBI:3892." According to the procedure of the method based on ChEBI ontology, 13 drugs in $\mathbf{S}_{tr}$ (listed in Table 4) were found, satisfying the function $Q(\bullet)$ to be minimum. It is observed that 11 out of 13 drugs are in the target group "G Protein-coupled Receptors" and the rest two drugs are in the target group "Enzymes." Thus, the target group "G Protein-coupled Receptors" got 11 votes, "Enzymes" got 2 votes, and the rest target groups did not get any votes. Accordingly, the target group of "D00146" is predicted to be "G Protein-coupled Receptors," which is indeed its true target group. Another example is the drug "D00387" in the test dataset $\mathbf{S}_{te}$, which is in the target group "Ion Channels." According to its ontology term "CHEBI:9674," we found 20 drugs in $\mathbf{S}_{tr}$, such that the function $Q(\bullet)$ achieved a minimum. These 20 drugs were listed in Table 5, from which we can see that 12 drugs are in target group "Ion Channels" and 8 drugs are in target group "G Protein-coupled Receptors." Thus, the result of "D00387" is predicted to be in the target group "Ion Channels." It is also predicted correctly.

Table 5: The 20 drugs which are closest to "D00387" in $\mathbf{S}_{tr}$.

Drug	Target group	Ontology term	Drug	Target group	Ontology term
D00225	Ion Channels	CHEBI:2611	D00293	Ion Channels	CHEBI:49575
D00300	G Protein-coupled Receptors	CHEBI:4636	D00311	Ion Channels	CHEBI:4858
D00430	Ion Channels	CHEBI:9073	D00494	G Protein-coupled Receptors	CHEBI:8461
D00506	Ion Channels	CHEBI:8069	D00549	Ion Channels	CHEBI:44915
D00669	G Protein-coupled Receptors	CHEBI:4637	D01177	G Protein-coupled Receptors	CHEBI:32091
D01205	G Protein-coupled Receptors	CHEBI:31472	D01310	Ion Channels	CHEBI:32124
D01372	Ion Channels	CHEBI:32315	D01485	G Protein-coupled Receptors	CHEBI:31981
D01657	Ion Channels	CHEBI:52993	D02419	G Protein-coupled Receptors	CHEBI:34720
D02624	Ion Channels	CHEBI:53760	D08283	Ion Channels	CHEBI:111762
D08473	G Protein-coupled Receptors	CHEBI:8802	D08690	Ion Channels	CHEBI:10125

The two examples in the above paragraph show that the target information of these drugs is indeed related to their ontology information. The good performance of the predictor demonstrated the validity of using ontology information to predict drugs target groups.

4. Conclusion

This study employed ChEBI ontology to categorize drugs based on their target proteins. The good performance of the method suggests that ontologies are good indicators of drugs target groups. However, only about 30% of the samples reported in KEGG were investigated in this study due to the lack of ontology information of most drugs. It is anticipated that the method would be more effective at the prediction with the development of ChEBI ontology and hopefully a multilabel classifier may be developed to allocate some drugs to more than one category in the near future.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Yu-Fei Gao and Lei Chen contributed equally to this work.

Acknowledgments

This work was supported by Grants from National Basic Research Program of China (2011CB510101, 2011CB510102), Innovation Program of Shanghai Municipal Education Commission (12YZ120, 12ZZ087), National Natural Science Foundation of China (31371335, 61202021, and 61373028), the Grant of "The First-Class Discipline of Universities in Shanghai", Natural Science Fund Projects of Jilin Province (201215059), Development of Science and Technology Plan Projects of Jilin Province (20100733, 201101074), SRF for ROCS, SEM (2009-36), Scientific Research Foundation (Jilin Department of Science and Technology, 200705314, 20090175, and 20100733), Scientific Research Foundation (Jilin Department of Health, 2010Z068), SRF for ROCS (Jilin Department of Human Resource and Social Security, 2012–2014), and Shanghai Educational Development Foundation (12CG55).

References

[1] J. Knowles and G. Gromo, "Target selection in drug discovery," *Nature Reviews Drug Discovery*, vol. 2, no. 1, pp. 63–69, 2003.

[2] A. C. Cheng, R. G. Coleman, K. T. Smyth et al., "Structure-based maximal affinity model predicts small-molecule druggability," *Nature Biotechnology*, vol. 25, no. 1, pp. 71–75, 2007.

[3] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, "A fast flexible docking method using an incremental construction algorithm," *Journal of Molecular Biology*, vol. 261, no. 3, pp. 470–489, 1996.

[4] S. Zhu, Y. Okuno, G. Tsujimoto, and H. Mamitsuka, "A probabilistic model for mining implicit "chemical compound-gene" relations from literature," *Bioinformatics*, vol. 21, no. 2, pp. ii245–ii251, 2005.

[5] 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.

[6] 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.

[7] L. Chen, Z.-S. He, T. Huang, and Y.-D. 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.

[8] L. Chen and W.-M. Zeng, "A two-step similarity-based method for prediction of drugs target group," *Protein and Peptide Letters*, vol. 20, pp. 364–370, 2013.

[9] 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.

[10] 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.

[11] 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.

[12] K. Degtyarenko, P. De matos, M. Ennis et al., "ChEBI: a database and ontology for chemical entities of biological interest," *Nucleic Acids Research*, vol. 36, no. 1, pp. D344–D350, 2008.

[13] L. Chen, W.-M. Zeng, Y.-D. Cai, K.-Y. 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.

[14] L.-L. Hu, C. Chen, T. Huang, Y.-D. 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.

[15] L. Chen, T. Huang, J. Zhang et al., "Predicting drugs side effects based on chemical-chemical interactions and protein-chemical interactions," *BioMed Research International*, vol. 2013, Article ID 485034, 8 pages, 2013.

[16] L. Chen, X. Shi, X. Kong, Z. Zeng, and Y.-D. Cai, "Identifying protein complexes using hybrid properties," *Journal of Proteome Research*, vol. 8, no. 11, pp. 5212–5218, 2009.

[17] 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.

[18] K.-C. Chou and Y.-D. Cai, "Prediction of protein subcellular locations by GO-FunD-PseAA predictor," *Biochemical and Biophysical Research Communications*, vol. 320, no. 4, pp. 1236–1239, 2004.

[19] C. E. Jones, U. Baumann, and A. L. Brown, "Automated methods of predicting the function of biological sequences using GO and BLAST," *BMC Bioinformatics*, vol. 6, article 272, 2005.

[20] M. A. Mahdavi and Y.-H. Lin, "False positive reduction in protein-protein interaction predictions using gene ontology annotations," *BMC Bioinformatics*, vol. 8, article 262, 2007.

[21] S. Carroll and V. Pavlovic, "Protein classification using probabilistic chain graphs and the Gene Ontology structure," *Bioinformatics*, vol. 22, no. 15, pp. 1871–1878, 2006.

[22] B. Smith, W. Ceusters, B. Klagges et al., "Relations in biomedical ontologies," *Genome biology*, vol. 6, no. 5, article R46, 2005.

[23] L. Hu, T. Huang, X. Shi, W.-C. Lu, Y.-D. 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.

[24] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular systems biology*, vol. 3, p. 88, 2007.

[25] 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.

[26] X. Shao, Y. Tian, L. Wu, Y. Wang, L. Jing, and N. Deng, "Predicting DNA- and RNA-binding proteins from sequences with kernel methods," *Journal of Theoretical Biology*, vol. 258, no. 2, pp. 289–293, 2009.

[27] 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.

[28] X. Xiao, J. Min, and P. Wang, "Predicting ion channel-drug interactions based on sequence-derived features and functional groups," *Journal of Bionanoscience*, vol. 7, pp. 49–54, 2013.

[29] R. G. Ramani and S. G. Jacob, "Prediction of P53 mutants (multiple sites) transcriptional activity based on structural (2D&3D) properties," *PLoS ONE*, vol. 8, Article ID e55401, 2013.

[30] G. S. Han, V. Anh, A. P. Krishnajith, and Y.-C. Tian, "An ensemble method for predicting subnuclear localizations from primary protein structures," *PLoS ONE*, vol. 8, Article ID e57225, 2013.

[31] Y. Matsuta, M. Ito, and Y. Tohsato, "ECOH: an enzyme commission number predictor using mutual information and a support vector machine," *Bioinformatics*, vol. 29, pp. 365–372, 2013.

[32] Z. Qiu, C. Qin, M. Jiu, and X. Wang, "A simple iterative method to optimize protein ligand-binding residue prediction," *Journal of Theoretical Biology*, vol. 317, pp. 219–223, 2012.

[33] Y.-N. Zhang, D.-J. Yu, S.-S. Li et al., "Predicting protein-ATP binding sites from primary sequence through fusing bi-profile sampling of multi-view features," *BMC Bioinformatics*, vol. 13, article 118, 2012.

[34] W. Chen and H. Lin, "Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine," *Computers in Biology and Medicine*, vol. 42, no. 4, pp. 504–507, 2012.

[35] L. Chen, W. Zeng -M, Y. Cai -D, 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.

[36] 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.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 703849, 15 pages
http://dx.doi.org/10.1155/2013/703849

Research Article

A Systems' Biology Approach to Study MicroRNA-Mediated Gene Regulatory Networks

Xin Lai,[1,2] Animesh Bhattacharya,[3] Ulf Schmitz,[1] Manfred Kunz,[3] Julio Vera,[2] and Olaf Wolkenhauer[1,4]

[1] *Department of Systems Biology and Bioinformatics, University of Rostock, 18051 Rostock, Germany*
[2] *Laboratory of Systems Tumor Immunology, Department of Dermatology, Faculty of Medicine, University of Erlangen-Nuremberg, Ulmenweg 18, 91054 Erlangen, Germany*
[3] *Department of Dermatology, Venereology and Allergology, University of Leipzig, 04155 Leipzig, Germany*
[4] *Institute for Advanced Study (STIAS), Wallenberg Research Centre at Stellenbosch University, Stellenbosch 7600, South Africa*

Correspondence should be addressed to Julio Vera; julio.vera-gonzalez@uk-erlangen.de and Olaf Wolkenhauer; olaf.wolkenhauer@uni-rostock.de

Received 31 July 2013; Revised 12 September 2013; Accepted 17 September 2013

Academic Editor: Tao Huang

MicroRNAs (miRNAs) are potent effectors in gene regulatory networks where aberrant miRNA expression can contribute to human diseases such as cancer. For a better understanding of the regulatory role of miRNAs in coordinating gene expression, we here present a systems biology approach combining data-driven modeling and model-driven experiments. Such an approach is characterized by an iterative process, including biological data acquisition and integration, network construction, mathematical modeling and experimental validation. To demonstrate the application of this approach, we adopt it to investigate mechanisms of collective repression on p21 by multiple miRNAs. We first construct a p21 regulatory network based on data from the literature and further expand it using algorithms that predict molecular interactions. Based on the network structure, a detailed mechanistic model is established and its parameter values are determined using data. Finally, the calibrated model is used to study the effect of different miRNA expression profiles and cooperative target regulation on p21 expression levels in different biological contexts.

1. Introduction

Although microRNAs (miRNAs) are physically small, they have been shown to play an important role in gene regulation [1]. Currently, an increasing number of studies are being carried out to deepen our understanding of miRNA regulatory mechanisms and functions. However, experimental approaches have limitations when dealing with complex biological systems composed of multiple layers of regulation such as the transcriptional and post-transcriptional regulation by transcription factors (TFs) and miRNAs [2]. Most experimental approaches focus on the identification of miRNA targets and the investigation of physiological consequences when perturbing miRNA expressions but are unsuited to provide a system-level interpretation for observed phenomena. Therefore, the introduction of a systematic approach, which can unravel the underlying mechanisms by which miRNAs exert their functions, becomes increasingly appealing.

The systems biology approach, combining data-driven modeling and model-driven experiments, provides a systematic and comprehensive perspective on the regulatory roles of miRNAs in gene regulatory networks [3–5]. To investigate a gene regulatory network, an iterative process of four steps is needed. (I) *Biological network construction*: a map is constructed that shows interactions among molecular entities (such as genes, proteins and RNAs), using information from literature and databases. (II) *Model construction*: depending on the biological problem investigated and experimental data available, the interaction map can be translated into a detailed mechanistic model that can simulate the temporal evolution of molecular entities. The values of parameters in this model can be determind from literature, databases or they are directly estimated from quantitative experimental data using

optimization methods. (III) *Computational experiments*: once a model is established, it can be simulated and/or analyzed for its general behavior. (IV) *Experimental validation*: model predictions together with biological explanations are integrated to guide the design of new experiments, which in turn validate or falsify the model. If model predictions are in agreement with the experiments, the model justifies the biological hypotheses behind it. In turn, these hypotheses, which provide reasonable explanations for the biological phenomenon, lead to an enhanced understanding of the gene regulatory network. Otherwise, the structure of the mathematical model is modified to generate new hypotheses and suggest new experiments.

The application of the systems biology approach to the analysis of a gene regulatory network is demonstrated with a case study of the regulation of p21 by multiple miRNAs [4]. The network combining putative targets of TF and miRNA regulation with experimentally proven molecular interactions was constructed and visualized. Next, the network was translated into a detailed mechanistic model, which was characterized and validated with experimental data. Finally, the integration of quantitative data and modeling helped us to generate and validate hypotheses about mechanisms of collective miRNA repression on p21.

2. Results

2.1. The Systems Biology Approach to Study miRNA-Mediated Gene Regulatory Networks

2.1.1. Mathematical Modeling. The aim of our analysis is to unravel the complex mechanisms by which gene regulatory networks involving miRNAs are regulated. We iteratively integrate data from literature, experiments and biological databases into a detailed mechanistic model of a gene regulatory network. The model is then used to formulate and test hypotheses about mechanisms of miRNA target regulation and cellular process-related variability. The methodology includes four steps which are briefly summarized in Table 1. In the coming sections we present these steps in detail.

(1) Data Retrieval. To construct a gene regulatory network composed by different levels of regulation, we collect information from different resources which are briefly described below, and more resources for data retrieval are introduced in Table 1.

(a) *Transcriptional level regulators*. Experimentally verified TFs for a gene can be extracted from literature or databases such as TRED, TRANSFAC, or HTRIdb [6–8]; the putative TFs, which are associated with conserved TF-binding sites residing in the promoter region of a gene, can for example, be extracted from the table of TFs with conserved binding sites in the UCSC genome browser or the TRANSFAC database [9]. miRGen 2.0 is a database that provides both predicted and experimentally verified information about miRNA regulation by TFs [10].

(b) *Post-transcriptional regulators*. Databases such as miRecords, Tarbase and miRTarBase are repositories of experimentally validated miRNA:gene interactions [11–13]. Predictions of miRNA:gene interactions are accumulated in databases like miRWalk [14].

(c) *Protein-protein interactions*. Both the Human Protein Reference Database (HPRD) [15] and the STRING database [16] document experimentally verified protein-protein interactions; besides, STRING also provides putative protein-protein interactions ranked by confidence scores. Further details about the exact mechanism of protein-protein interactions can be found in Reactome [17].

(2) Network Construction and Visualization. Based on the information collected, a gene regulatory network is constructed and visualized for providing an overview. For this purpose, we recommend CellDesigner which uses standardized symbols (Systems Biology Graphical Notation—SBGN) [18] for visualization and stores gene regulatory networks in the SBML format (Systems Biology Markup Language) [19]. CellDesigner also provides the possibility to simulate temporal dynamics of the gene regulatory network due to the integration of the SBML ODE (ordinary differential equation) solver. Besides, Cytoscape is another powerful tool for integration of biological networks and gene expression data [20].

For assessing the reliability of interactions considered in gene regulatory networks, confidence scores can be computed as being documented in our previous publication [4]. The factors that are used to determine the confidence score for molecular interactions can be: the number of publications reporting an interaction, experimental methods used to identify an interaction, interaction types and computational predictions. The computed confidence scores range from 0 to 1, where values towards 1 indicate higher confidence, whereas values towards 0 indicate lower confidence in a given interaction. For example, the confidence score for a miRNA:gene interaction can be calculated using the following equation:

$$S_{\text{miRNA:gene}} = \frac{w_p \cdot S_p + w_m \cdot S_m + w_{\text{bs}} \cdot S_{\text{bs}}}{w_p + w_m + w_{\text{bs}}}, \quad (1)$$

where $w_{\langle p,m,\text{bs}\rangle}$ are weights that are assigned to the scores which account for the number of publications (S_p), detection method (S_m) and the number of predicted binding sites (S_{bs}). The values of the weights can be assigned based on expert knowledge, and the higher the value of the weight is, the bigger impact it has on the confidence score for the interaction. The values of S_p and S_{bs} can be calculated using the logarithmic equation: $S_{\langle p,\text{bs}\rangle}(n) = \log_{m+1}(n)$, where n denotes the number of publications describing the miRNA:gene interaction or the number of binding sites that the miRNA has in the $3'$ UTR (untranslated region) of the gene. The value of m is a cut-off that represents the number of publications or binding sites required for $S_{\langle p,\text{bs}\rangle}$ to obtaining their maximum values. Various methods such as western blots, qRT-PCR and reporter assays can be applied to support the miRNA:gene interaction, but these methods

Table 1: Overview of the methodology. Key points in each step of the methodology and the main resources for constructing miRNA-mediated gene regulatory networks are given.

Step 1: data retrieval	
Regulation types	Resources
Transcriptional gene regulation	TRED (http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home): a database that provides an integrated repository for both cis- and transregulatory elements in mammals TRANSFAC (http://www.gene-regulation.com/pub/databases.html): a database that collects eukaryotic transcriptional regulation, comprising data on TFs, their target genes, and binding sites The UCSC table browser (http://genome.ucsc.edu/): a popular web-based tool for querying the UCSC Genome Browser annotation tables HTRIdb (http://www.lbbc.ibb.unesp.br/htri/): an open-access database for experimentally verified human transcriptional regulation interactions MIR@NT@N (http://maia.uni.lu/mironton.php/): an integrative resource based on a metaregulation network model including TFs, miRNAs, and genes PuTmiR (http://www.isical.ac.in/~bioinfo_miu/TF-miRNA.php): a database of predicted TFs for human miRNAs TransmiR (http://202.38.126.151/hmdd/mirna/tf/): a database of validated TF-miRNA interactions miRGen 2.0 (http://diana.cslab.ece.ntua.gr/mirgen/): a database of miRNA genomic information and regulation
Posttranscriptional gene regulation	miRecords (http://mirecords.biolead.org/): a resource for animal miRNA-target interactions Tarbase (http://www.microrna.gr/tarbase/): a database that stores detailed information for each miRNA-gene interaction, the experimental validation methodologies, and their outcomes miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/): a database that collects validated miRNA-target interactions by manually surveying the pertinent literature miRWalk (http://www.umm.uni-heidelberg.de/apps/zmf/mirwalk/): a comprehensive database that provides information on miRNAs from human, mouse, and rat, on their predicted as well as validated binding sites in target genes
Protein-protein interaction	HPRD (http://www.hprd.org/): a centralized platform to visually depict and integrate information pertaining to domain architecture, posttranslational modifications, interaction networks, and disease association for each protein in the human proteome STRING (http://string-db.org/): a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations MPPI (http://mips.helmholtz-muenchen.de/proj/ppi/): a collection of manually curated high-quality PPI data collected from the scientific literature by expert curators DIP (http://dip.doe-mbi.ucla.edu/dip/Main.cgi): a catalog of experimentally determined interactions between proteins IntAct (http://www.ebi.ac.uk/intact/main.xhtml): a platform that provides a database system and analysis tools for molecular interaction data Reactome (http://www.reactome.org/): an open-source, open access, manually curated, and peer-reviewed pathway database
GO annotation	Amigo GO (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi): the official GO browser and search engine miR2Disease (http://www.mir2disease.org/): a manually curated database that aims at providing a comprehensive resource of miRNA deregulation in various human diseases miRCancer (http://mircancer.ecu.edu): a miRNA-cancer association database constructed by text mining on the literature PhenomiR (http://mips.helmholtz-muenchen.de/phenomir/): a database that provides information about differentially expressed miRNAs in diseases and other biological processes miRGator (http://mirgator.kobic.re.kr/): a novel database and navigator tool for functional interpretation of miRNAs miRó (http://ferrolab.dmi.unict.it/miro): a web-based knowledge base that provides users with miRNA-phenotype associations in humans
Step 2: network construction and visualization	
(i) Visualize regulatory interactions in platforms such as CellDesigner and Cytoscape that support standardized data formats	
(ii) Calculate confidence scores for assessing reliability of interactions in gene regulatory networks	
Step 3: model construction and calibration	
(i) Formulate equations using rate equations	
(ii) Fix parameter values using available biological information	
(iii) Estimate the other unknown and immeasurable parameter values using optimization methods which can minimize the distance between model simulations and experimental data such as time course qRT-PCR and western blot data	
Step 4: model validation and analysis	
(i) Design new experiments and generate new data to verify the calibrated model	
(ii) Study complex properties and behavior of the system	

provide experimentalists with different levels of confidence, thus differing confidences can be reflected using different values for S_m based on the experience of experimentalists. Of note, although the confidence scores cannot be directly converted into a mathematical model, with the help of these scores we can discard non-reliable putative interactions to generate the ultimate version of a gene regulatory network. The final version of the network can be further analyzed to identify regulatory motifs like feedforward loops (FFLs), for example, with the help of the Cytoscape plugin NetDS [21]. Thereafter, the complete network or parts of it can be converted into a detailed mechanistic model which is described in detail in the following section.

(3) Model Construction and Calibration. After the construction, visualization and refinement of a gene regulatory network, it is converted into a detailed mechanistic model which enables the investigation of unanswered biological questions and validation of hypotheses. Ordinary differential equations (ODEs) describe how processes of synthesis, biochemical modification and/or degradation affect the temporal concentration profile of biochemical species like proteins, RNAs and metabolites. An ODE model can be constructed using appropriate kinetic laws such as the law of mass-action, which states that the rate of a chemical reaction is proportional to the probability that the reacting species collide. This collision probability is in turn proportional to the concentration of the reactants [22]. A general representation of ODE-based models using mass-action kinetics is given by the following equation

$$\frac{dx_i}{dt} = \sum_{\mu=1}^{m} c_{i\mu} \cdot k_\mu \cdot \prod_{j=1}^{n} x_j^{g_{\mu j}}, \quad i \in \{1, 2, \ldots, n\}, \tag{2}$$

where x_i represents state variables which denote the molar concentration of the ith biochemical specie. Every biochemical reaction μ is described as a product of a rate constant (k_μ) and biochemical species ($x_j, j \in \{1, 2, \ldots, n\}$) that are involved in this reaction. $c_{i\mu}$, the so-called stoichiometric coefficients, relate the number of reactant molecules consumed to the number of product molecules generated in the reaction μ. $g_{\mu j}$ denotes kinetic orders which are equal to the number of species of x_i involved in the biochemical reaction μ. The rate constants, kinetic orders and the initial conditions of state variables are defined as model parameters. Besides mass-action kinetics, other kinetic rate laws such as Michaelis-Menten kinetics, Hill equation and power-laws are also frequently used in ODE models.

After ODEs are formulated, the model requires to be calibrated, a process by which parameter values are adjusted in order to make model simulations match experimental observations as good as possible. To do so, there are two possible means: characterization of parameter values using available biological information or estimation of parameter values using optimization methods. Some parameter values can be directly measured or obtained from literature or databases. For example, the half-life ($t_{1/2}$) of some molecules (e.g., protein) can be measured *in vitro* via western blotting. This information can be used to characterize their degradation rate constants through the equation $k_{\text{deg}} = (\ln 2/t_{1/2})$. The database SABIO-RK provides a platform for modelers of biochemical networks to assemble information about reactions and kinetic constants [27]. However, for most model parameters, whose values cannot be measured in laboratories or be accessed from literature or databases, parameter estimation is a necessary process to characterize their values. Before running parameter estimation, initial parameter values and boundaries should be set within physically plausible ranges. To do so, the database BioNumbers provides modelers with key numbers in molecular and cell biology, ranging from cell sizes to metabolite concentrations, from reaction rates to generation times, from genome sizes to the number of mitochondria in a cell [28]. After parameter estimation, unknown parameter values are determined using optimization methods which can minimize a cost function that measures the goodness of fit of the model with respect to given quantitative experimental data sets. Parameter estimation using optimization algorithms is an open research field, in which several methods have been developed according to the nature and numerical properties of biological data analyzed. The discussion for choosing proper optimization methods for parameter estimation is beyond the scope of this paper, but the interested reader is referred to the paper published by Chou and Voit [29].

(4) Model Validation and Analysis. Usually, the model simulations are compared with the experimental data used for the parameter estimation, but a good agreement between both is not enough to guarantee the predictive ability of the model. Therefore, it is necessary to validate the model with data sets that are not used during the parameter estimation. This process is called model validation and can ensure more reliable and accurate model predictions. To do so, the data generated in new experiments or extracted from literature are compared with model simulations, which are obtained after configuring the model according to the new experimental settings. Once a model is validated, it can be used to perform predictive simulations, which are helpful to study the dynamic properties of biochemical systems, guide the design of new experiments in the laboratory and formulate additional hypotheses. In addition, tools such as sensitivity and bifurcation analysis can be used to study complex properties and behavior of the modeled system. Sensitivity analysis is used to evaluate the influence of model parameters (e.g., initial concentrations of the state variables and rate constants) on model outputs, such as the temporal behavior of network components [30]. Bifurcation analysis is employed to detect control parameters (also known as bifurcation parameters) whose variations are able to change drastically the dynamical properties of the biochemical system, as well as the stability of its fixed points [31]. The application of these tools to mathematical modeling is beyond the scope of this paper, but the interested reader is referred to the publication of Zhou et al. [32] and Marino et al. [33].

2.1.2. Experiment Methods. As mentioned in the previous sections, after a model is established, it can be calibrated

and validated using temporal experimental data. To do so, the data can either be derived from literature or generated to calibrate the model by own experiments. In case of ODE models, the most suitable data for model calibration is quantitative time-series data obtained from perturbation or quantitative dose-response experiments. The experiments, in which time-series data are measured for different regulators (such as miRNAs and TFs) of a gene regulatory network, can be obtained using the techniques described in the subsections below.

(1) qRT-PCR. Quantitative real time polymerase chain reaction (qRT-PCR) has been used to identify mRNAs regulated by overexpression or silencing of a specific miRNA [34, 35]. miRNAs typically exhibit their regulatory effects by associating with specific $3'$ UTR regions of the mRNAs called miRNA seed regions [34]. This association can lead either to a temporary inhibition of translation or complete degradation of the mRNA in which case qRT-PCR mediated detection is beneficial. For this, the cells are transfected with the miRNA and non-targeting control oligonucleotides at an appropriate concentration using either a lipid based transfection reagent or nucleofection. The transfected cells are then incubated for the necessary time periods (e.g., 24 h, 48 h, 72 h, etc.) after which lysates are prepared and total RNA is extracted. 1000–2000 ng of the RNA is then converted into cDNA using a reverse transcription kit. Taqman qRT-PCR is performed using 10–20 ng of cDNA and primers labeled with fluorescence probes to detect the transcriptional levels of the target mRNA. Housekeeping genes like GAPDH and HPRT are used as endogenous controls for data normalization. Relative expression at different time points is determined by comparing the Ct values of the miRNA transfected cells with Ct values of non-targeting control transfected cells and expressed as relative expression [36]. However, qRT-PCR in spite of being highly sensitive is applicable specifically under conditions of complete or partial degradation of the target mRNA. miRNA-mediated inhibition in translation can be better demonstrated by techniques like immunoblotting.

(2) Western Blot/Immunblotting. Western blotting or protein immunoblotting is a technique to detect the expression of a gene at protein level. This technique is particularly useful in determining the regulatory effects of a miRNA on expression of a target gene which is temporarily inhibited. For this the cells are transfected with a miRNA or an antagomiR as mentioned above and cell lysates are prepared at appropriate time points using protein lysis buffers (e.g., Radio-Immunoprecipitation Assay buffer or RIPA) containing lysis agents like Dithiothreitol (DTT) and protease inhibitors. The proteins from each sample are quantitated using Bradford or BCA reagents and compared with bovine serum albumin (BSA) standards for accurate protein estimation. 20–40 μg of protein is then loaded and resolved on a sodium dodecyl sulfate polyacrylamide gel (SDS-PAGE) along with a pre-stained protein marker. The protein bands are then transferred onto a nitrocellulose membrane followed by incubation with the appropriate primary and secondary antibodies linked to fluorescent dyes or horse radish peroxidase enzyme (HRP). The protein expression is then analyzed using either fluorescence or chemiluminiscence HRP substrates in gel documentation systems (e.g., LI-COR Odyssey). Housekeeping genes like β-actin or β-tubulin are used for protein normalization. The time point for maximum target gene suppression generally varies depending on the number of miRNA binding sites at the $3'$ UTR of the target gene and the extent of complementarity of the seed region [37]. Immunoblotting is a widely used technique to provide confirmatory evidence for the inhibitory effects of miRNA at the protein level, but it fails to explain the underlying interaction mechanisms.

(3) Reporter Gene Assay. As each miRNA can inhibit the expression of a large number of genes, regulation of a particular target gene may either be by direct interaction or be an indirect consequence of it. In direct regulation, a miRNA binds to the complementary sequences at the $3'$ UTR of a target gene and thereby suppresses its expression. As a consequence of this, the expression levels of a number of downstream genes (indirect targets) are also dysregulated making it crucial to differentiate between primary and secondary miRNA targets. To determine the interaction specificity, a reporter construct (luciferase) with intact or mutated $3'$ UTR of the target gene cloned at the $5'$ end is co-transfected into the cells along with the miRNA. The regulatory effect of the miRNA on the target gene expression is then measured using the expression of a reporter gene. In the absence of the appropriate binding sequences (mutated $3'$ UTR), the miRNA cannot suppress the reporter mRNA suggesting that the suppressive effect of the miRNA is mediated by a direct interaction. The reporter activity can be analyzed at different time points such as 24 hr, 48 hr and 72 hr to determine the time dependent suppression of a target gene expression by a miRNA.

2.2. Case Study: The Regulation of p21 by Multiple and Cooperative miRNAs

2.2.1. Construction and Visualization of p21 Regulatory Network. By using the approach described above, we investigated the regulation of p21 by its multiple targeting miRNAs. p21, also known as cyclin-dependent kinase inhibitor 1 (CDKN1A), is a transcriptional target of p53. It is required for proper cell cycle progression and plays a role in cell death, DNA repair, senescence and aging (reviewed in [38]). Interestingly, p21 was the first experimentally validated miRNA target hub, which is a gene that is simultaneously regulated by many miRNAs. This made it an ideal candidate for a case study of our approach [23]. To do so, we first constructed a p21 regulatory network using the following steps:

(a) We extracted miRNA-target interactions from the publication of Wu et al. [23] where a list of predicted p21-regulating miRNAs was subjected to experimental validation.

(b) Experimentally verified TFs of p21 were extracted from literature and putative TFs having conserved binding sites in the 5 kb upstream region of the p21 open reading frame were extracted from UCSC table

browser. A list of TFs controlling the expression of the miRNAs was constructed using information of experimentally proven TF-miRNA interactions extracted from TransmiR (release 1.0) [39]. In addition, we generated a list of putative TFs of miRNAs with binding sites in the 10 kb upstream region of the miRNA genes using information from the databases PuTmiR (release 1.0) [40] and MIR@NT@N (version 1.2.1) [41], and from the table of TFs with conserved binding sites in the UCSC genome browser (hg18) [9].

(c) Information about protein interactions was extracted from the Human Protein Reference Database (HPRD, release 9.0) [15] and the STRING database (release 9.0) [16]. Only the experimentally verified p21-protein interactions were used to construct the network.

(d) Additionally, we associated the TFs in the network to nine biological processes based on the Gene Ontology (GO) [42]. The corresponding GO terms were cell proliferation, cell apoptosis, immune response, inflammatory response, cell cycle, DNA damage, cell senescence, DNA repair and cell migration.

Next, we visualized the network in CellDesigner and computed a confidence score for each interaction in the network (Figure 1). The confidence scores provide us with the reliability of the interactions considered in p21 regulatory network. With the help of these scores, we discarded non-reliable interactions and constructed the mechanistic model accounting for p21 regulation by its targeting miRNAs. Besides, the interested experimentalists can further use this information to choose reliable interacting candidates of p21 for designing relevant experiments. The scores for each interaction of p21 regulatory network are shown in Table 2.

2.2.2. Mechanistic Modeling of p21 Regulation by Its Targeting miRNAs

(1) Model Construction. After constructing the regulatory network, a detailed mechanistic model of ODEs, which describes the biochemical reactions underlying the regulation of p21 was established. We chose the ODE modeling approach, as it is a simple formalism for describing temporal dynamics of biochemical systems and a wide range of tools are available to explore their properties. Precisely, the model considered the mRNA (mp21; (3)) and protein (p21; (6)) of the miRNA target hub p21, the p21-targeting miRNAs (miR_i; $i \in (1, \ldots, 15)$; (5)), and the complexes formed by p21 mRNA and miRNA, $[\mathrm{mp21} \mid \mathrm{miR}_i]$ (4). Altogether, the model is constituted by 32 state variables and 64 parameters:

$$\frac{d\mathrm{mp21}}{dt} = k_{\mathrm{syn}}^{\mathrm{mp21}} \cdot f_{\mathrm{act}}\left(\mathrm{TF}_{\mathrm{mp21}}\right) - \mathrm{mp21} \cdot \left(k_{\mathrm{deg}}^{\mathrm{mp21}} + \sum_i k_{\mathrm{ass}}^{\mathrm{complex}_i} \cdot \mathrm{miR}_i\right), \tag{3}$$

$$\frac{d\left[\mathrm{mp21} \mid \mathrm{miR}_i\right]}{dt} = k_{\mathrm{ass}}^{\mathrm{complex}_i} \cdot \mathrm{mp21} \cdot \mathrm{miR}_i - k_{\mathrm{deg}}^{\mathrm{complex}_i} \cdot \left[\mathrm{mp21} \mid \mathrm{miR}_i\right], \tag{4}$$

$$\frac{d\mathrm{miR}_i}{dt} = k_{\mathrm{syn}}^{\mathrm{miR}_i} \cdot f_{\mathrm{act}}\left(\mathrm{TF}_{\mathrm{miR}_i}\right) - \mathrm{miR}_i \cdot \left(k_{\mathrm{deg}}^{\mathrm{miR}_i} + k_{\mathrm{ass}}^{\mathrm{complex}_i} \cdot \mathrm{mp21}\right), \tag{5}$$

$$\frac{\mathrm{dp21}}{dt} = k_{\mathrm{syn}}^{\mathrm{p21}} \cdot \mathrm{mp21} + k_{\mathrm{deg}}^{\mathrm{p21}} \cdot \mathrm{p21}, \tag{6}$$

$$\mathrm{mp21}_{\mathrm{Total}} = \mathrm{mp21} + \sum_i \left[\mathrm{mp21} \mid \mathrm{miR}_i\right]. \tag{7}$$

For mp21, processes considered in the model were: (i) its synthesis ($k_{\mathrm{syn}}^{\mathrm{mp21}}$) mediated by TFs ($f_{\mathrm{act}}(\mathrm{TF}_{\mathrm{mp21}})$), (ii) its degradation ($k_{\mathrm{deg}}^{\mathrm{mp21}}$), and (iii) its association with a miRNA ($k_{\mathrm{ass}}^{\mathrm{complex}_i}$). For each miR_i, processes considered were: (i) the synthesis ($k_{\mathrm{syn}}^{\mathrm{miR}_i}$) mediated by TFs ($f_{\mathrm{act}}(\mathrm{TF}_{\mathrm{miR}_i})$), (ii) the degradation ($k_{\mathrm{deg}}^{\mathrm{miR}_i}$), and (iii) the association with the p21 mRNA target ($k_{\mathrm{ass}}^{\mathrm{complex}_i}$). For each $[\mathrm{mp21} \mid \mathrm{miR}_i]$ complex, processes considered were: (i) the formation of the complex by a miR_i and the p21 mRNA ($k_{\mathrm{ass}}^{\mathrm{complex}_i}$), and (ii) the complex degradation ($k_{\mathrm{deg}}^{\mathrm{complex}_i}$). For p21, processes considered were: (i) its synthesis ($k_{\mathrm{syn}}^{\mathrm{p21}}$), and (ii) its degradation ($k_{\mathrm{deg}}^{\mathrm{p21}}$). An additional algebraic equation accounting for the total measurable amount of p21 mRNA ($\mathrm{mp21}_{\mathrm{Total}}$) was also included. The SBML file of the model is available for download at http://www.sbi.uni-rostock.de/uploads/tx_templavoila/p21_TargetHub_03092013.xml.

(2) Model Calibration and Validation. For model calibration, we fixed some parameter values using published data and estimated the other unknown parameters using the time-series data published by Wu et al. [23], in which the p21 mRNA (northern plot) and protein levels (western plot) were measured 48 hr after transfection of individual p21-targeting miRNAs into human embryonic kidney 293 cells. The unknown parameter values were estimated using an iterative method combining global (particle swarm pattern search) [43] and local (downhill simplex method in multi-dimensions) [44] optimization algorithms. For each miR_i considered in the model, the method minimizes the distance between model simulations and experimental data using the following cost function

$$F_{\mathrm{cost}}^{\mathrm{miR}_i} = \frac{\left[\mathrm{mp21}_{\mathrm{sim}}^{\mathrm{miR}_i}(t) - \mathrm{mp21}_{\mathrm{exp}}^{\mathrm{miR}_i}(t)\right]^2}{\left(\delta_i^{\mathrm{mp21}}\right)^2} + \frac{\left[\mathrm{p21}_{\mathrm{sim}}^{\mathrm{miR}_i}(t) - \mathrm{p21}_{\mathrm{exp}}^{\mathrm{miR}_i}(t)\right]^2}{\left(\delta_i^{p21}\right)^2}, \quad i \in [1, \ldots, 15], \tag{8}$$

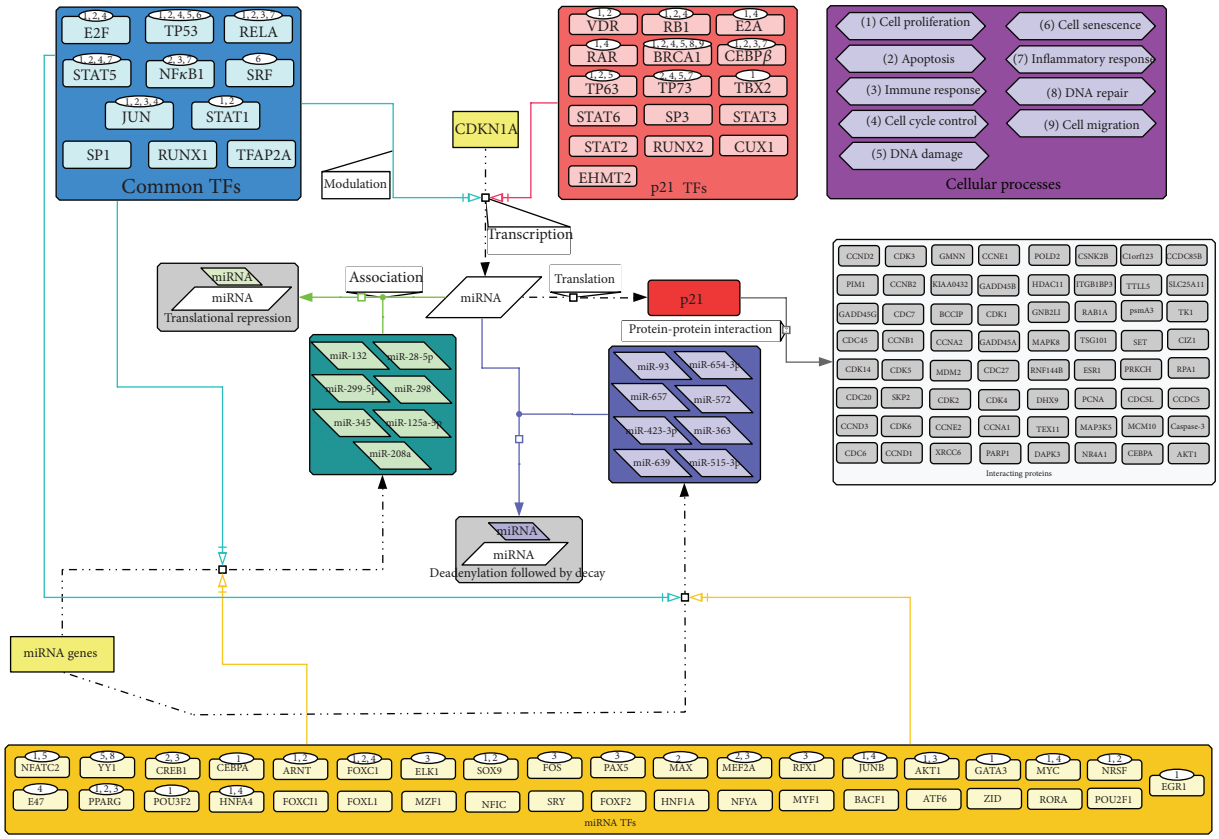


FIGURE 1: p21 regulatory network. The network contains several layers of regulators of p21: TFs (light blue and red boxes), miRNAs (dark blue and green boxes), and proteins (grey boxes). In each big box, there are small boxes which represent individual components of this layer of regulation. miRNAs are classified into two groups according to the mechanisms by which the expression of p21 is repressed. One group causes p21 translation repression (green box). These miRNAs bind to p21 mRNA resulting in the repressed translation of p21 but unchanged mRNA expression level. The other group of miRNAs (dark blue box) decreases the stability of p21 mRNA by modifying its structure, leading to mRNA decay and finally the downregulation of p21. TFs are classified into three groups: p21 TFs (red), miRNA TFs (yellow) and their common TFs (light blue). The p21 interacting-proteins are framed in the grey boxes. The purple boxes represent nine processes, and the TFs associated with these processes are indicated in the ellipses above them using corresponding figures. This data is adapted from our previous publication [4].

where $\mathrm{mp21}_{\mathrm{sim}}^{\mathrm{miR}_i}(t)$ and $\mathrm{mp21}_{\mathrm{exp}}^{\mathrm{miR}_i}(t)$ represent the simulated p21 mRNA and protein expression levels for each miR_i at time point t. $\mathrm{p21}_{\mathrm{sim}}^{\mathrm{miR}_i}(t)$ and $\mathrm{p21}_{\mathrm{exp}}^{\mathrm{miR}_i}(t)$ represent the measured value for each miR_i at time point t, and their standard deviations are δ_i^{mp21} and δ_i^{p21}. Here, t is the time point (48 hr) after overexpression of the individual miRNAs in embryonic kidney 293 cells at which the expression levels of the p21 and its mRNA were measured [23]. The model calibration results are shown in Figure 2(a) and the obtained parameter values are listed in Table 3.

Experimental results showed that a stronger repression of the target gene can occur when two miRNA binding sites on the target mRNA are in close proximity [24, 45]. To test the consequences of this hypothesis, we predicted cooperative miRNA pairs for p21, with seed site distances between 13–35 nt in the p21 3′ UTR. To substantiate the cooperative effect associated with pairs of miRNAs, we introduced a group of new state variables ([mp21 | miR_i | miR_j]) into the original model. These state variables account for the ternary complexes composed of p21 mRNA and two putatively cooperating miRNAs (miR_i and miR_j). For these new variables, processes considered are: (i) the association of p21 mRNA with miR_i and miR_j into a complex ($k_{\mathrm{ass}}^{\mathrm{co\text{-}complex}_{i,j}}$), and (ii) the degradation of the complex ($k_{\mathrm{deg}}^{\mathrm{co\text{-}complex}_{i,j}}$). After expansion, the corresponding modified and new ODEs are listed below:

$$\begin{aligned}\frac{d\mathrm{mp21}}{dt} &= k_{\mathrm{syn}}^{\mathrm{mp21}} \cdot f_{\mathrm{act}}\left(\mathrm{TF}_{\mathrm{mp21}}\right) \\ &\quad - \mathrm{mp21} \cdot \left(k_{\mathrm{deg}}^{\mathrm{mp21}} + \sum_i k_{\mathrm{ass}}^{\mathrm{complex}_i} \cdot \mathrm{miR}_i \right. \\ &\quad \left. + \sum_{i,j} k_{\mathrm{ass}}^{\mathrm{co\text{-}complex}_{i,j}} \cdot \mathrm{miR}_i \cdot \mathrm{miR}_j \right),\end{aligned} \tag{9}$$

Table 2: The confidence scores for the interactions in p21 regulatory network. The network consists of four types of interactions: miRNA-p21, TF-miRNA, TF-p21, and p21-protein.

(a) miRNA-p21 interaction scores

miRNA	miR-125a-5p	miR-132	miR-208a	miR-28-5p	miR-298	miR-299-5p	miR-345	miR-363
Score	0.73	0.73	0.73	0.80	0.73	0.80	0.73	0.73
miRNA	miR-423-3p	miR-515-3p	miR-572	miR-639	miR-654-3p	miR-657	miR-93	
Score	0.85	0.73	0.73	0.73	0.73	0.80	0.95	

(b) TF-miRNA interaction scores

	Score		Score		Score		Score		Score		Score
miR-208		miR-132		miR-657		miR-125a-5p		miR-28-5p		miR-93	
MAX	0.27	EGR1	0.93	TFAP2A	0.34	EGR1	0.86	MZF1	0.28	TFAP2A	0.34
EGR1	0.29	CREB1	0.86	SP1	0.34	PAX5	0.24	E47	0.27	SOX9	0.24
ARNT	0.27	AREB6	0.27	MZF1	0.34	NRSF	0.24	FOXC1	0.31	GATA3	0.27
RFX1	0.24	E47	0.24	MAX	0.29	ZID	0.24	FOXI1	0.24	MZF1	0.31
ATF6	0.24	ELK1	0.27	USF1	0.27	PPARG	0.27	SRY	0.27	SP1	0.34
YY1	0.27	EGR2	0.29	EGR1	0.34	GATA3	0.34	JUN	0.29	POU2F1	0.24
JUNB	0.24	FOS	0.24	RELA	0.29	MZF1	0.34	POU2F1	0.24	EGR1	0.29
POU2F1	0.24	NFIC	0.24	miR-654-3p		TP53	0.34	SRF	0.27	CEBPA	0.24
NFIC	0.27	RELA	0.27	YY1	0.27	NFIC	0.34	CEBPA	0.29	NFYA	0.24
miR-345		miR-423-3p		BACH1	0.24	miR-299-5p		RORA	0.24	RUNX1	0.24
SP1	0.34	BACH1	0.27	FOXL1	0.29	CREB1	0.24	miR-363		STAT1	0.24
RELA	0.27	STAT1	0.27	SRY	0.24	SRY	0.27	E2F1	0.62	E2F1	0.85
NFATC2	0.24	POU3F2	0.24	NFATC2	0.24	SRF	0.29	FOXC1	0.33	MYC	0.84
HNF4A	0.31	STAT5A	0.27	FOS	0.24	RUNX1	0.27	MZF1	0.34	miR-298	
NFYA	0.24	SRF	0.27	POU2F1	0.27	miR-572		miR-639		JUNB	0.31
POU2F1	0.24	MEF2A	0.27	NFκB1	0.24	FOXF2	0.24	NFYA	0.24		
NFIC	0.31	POU2F1	0.24	HNF1A	0.24						

(c) TF-p21 interaction scores

TF (verified)	SP1	SP3	E2F1	Runx1	Runx2	STAT1	E2A	STAT5	TP53	TP63	TP73
Score	1.00	1.00	0.77	0.77	0.77	0.72	0.72	0.67	0.67	0.67	0.67
TF (verified)	STAT3	CUX1	TFAP2A	BRCA1	VDR	RARA	C/EBPα	C/EBPβ	RB1	Tbx2	EHMT2
Score	0.66	0.66	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60
TF (putative)	NFκB1	RELA	STAT2	STAT6	SRF						
Score	0.44	0.44	0.44	0.44	0.43						

(d) p21-protein interaction scores

Protein	TP53	PCNA	CASP3	CCNA1	CCND1	SKP2	BCCIP	CCNA2	CCND2	CCNE2	AKT1	C1orf123	CCDC85B	CCNB1	CCNB2	CCND3
Score	1.00	0.92	0.87	0.87	0.87	0.87	0.81	0.81	0.81	0.81	0.72	0.72	0.72	0.72	0.72	0.72
Protein	CCNE1	CDC45	CDC5L	CDC6	CDC7	CDK1	CDK14	CDK2	CDK3	CDK4	CDK6	CEBPA	CIZ1	CSNK2A1	CSNK2B	DAPK3
Score	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
Protein	ESR1	GADD45A	GADD45B	GADD45G	GMNN	GNB2L1	HDAC11	ITGB1BP3	MAP3K5	MAPK8	MCM10	PARP1	PIM1	POLD2	PSMA3	RAB1A
Score	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
Protein	SET	SLC25A11	STAT3	TEX11	TK1	TSG101	TTLL5	XRCC6	CDC20	CDC27	CDK5	DHX9	MDM2	NR4A1	PRKCH	RPA1
Score	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56

Verified: experimentally verified interaction; putative: predicted interaction.

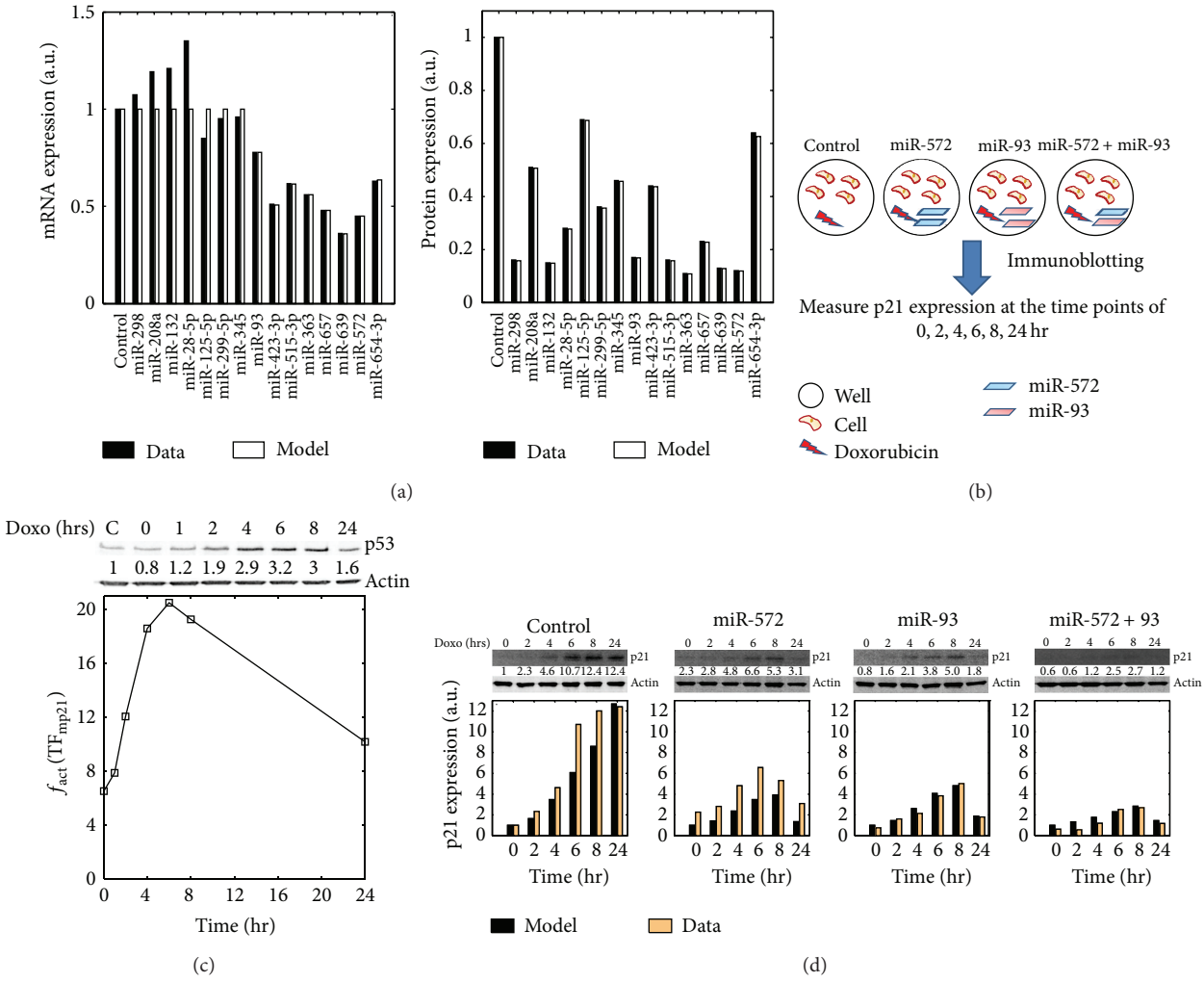


Figure 2: Model calibration and validation. (a) Model calibration. The figures show the relative change of the p21 mRNA and protein expression levels after overexpression of the indicated miRNAs (Model: model simulation; Data: experimental data). These data were normalized to the control group in which the p21 mRNA and protein expression levels were measured when the miRNAs were normally expressed (a.u.: arbitrary unit). (b) Experimental workflow. In the experiments, Sk-Mel-147 cells were seeded in six well plates. Then, mature miRNA mimics were transfected individually at a concentration of 100 nM (miR-572 and miR-93) or in combination at 50 nM each (miR-572 + miR-93). After 48 hr transfection with miRNA mimics, the cells were pulse treated with 250 nM doxorubicin for 1 hour after which normal growth medium was replenished. The immunblotting were performed to measure p21 expression at 0, 2, 4, 6, 8 and 24 hr post-doxorubicin treatment. (c) Temporal dynamics of p21 transcriptional function. Afterdoxorubicin treatment, the expression of p53, a TF of p21, was measured using immunblotting and these data were used to characterize the transcriptional function of p21 using MATLAB linear interpolation function. (d) Model validation. We measured the expression of p21 protein in response to genotoxic stress in the four scenarios as described in the main text. The measured data (Data) were compared with the model simulations (Model). The figures (a), (c) and (d) are adapted from our previous publication [4].

$$\frac{d\left[\text{mp21} \mid \text{miR}_i \mid \text{miR}_j\right]}{dt} = k_{\text{ass}}^{\text{co-complex}_{i,j}} \cdot \text{mp21} \cdot \text{miR}_i \cdot \text{miR}_j - k_{\text{deg}}^{\text{co-complex}_{i,j}} \cdot \left[\text{mp21} \mid \text{miR}_i \mid \text{miR}_j\right]. \quad (10)$$

To model stronger repression of the target gene by cooperating miRNAs, we assumed a stronger association rate constant for the complex [mp21 | miR$_i$ | miR$_j$] which is equal to the sum of their individual association rate constants ($k_{\text{ass}}^{\text{co-complex}_{i,j}} = k_{\text{ass}}^{\text{complex}_i} + k_{\text{ass}}^{\text{complex}_j}$). Similarly, the degradation rates of the complexes [mp21 | miR$_i$ | miR$_j$] were assumed to be equal to the sum of degradation rate constants of single miRNA binding complexes ($k_{\text{deg}}^{\text{co-complex}_{i,j}} = k_{\text{deg}}^{\text{complex}_i} + k_{\text{deg}}^{\text{complex}_j}$). However, it has to be noted that these added equations are an abstract description of miRNA cooperativity, because the details of this mechanism are not yet known.

To experimentally validate the capability of our model to predict the relative p21 concentrations regulated by cooperative miRNAs, we selected miR-572 and miR-93 as a case study.

Table 3: Initial concentrations of model variables and model parameter values. Based on the experimental data, the p21-targeting miRNAs verified by Wu et al. [23] were divided into two groups: the translation repression group (marked with asterisk) and the mRNA deadenylation group. A miRNA was classified into the mRNA deadenylation group if its overexpression can result in 20% or more downregulation of the p21 mRNA level (i.e., p21 mRNA level ≤ 0.8; the basal level is 1); otherwise, it was classified into the translation repression group. For the translation repression group, only $k_{ass}^{complex_i}$ was estimated and $k_{deg}^{complex_i}$ was fixed. For the other group, both $k_{deg}^{complex_i}$ and $k_{ass}^{complex_i}$ were estimated. The initial concentrations of p21 and mp21 were set to 1, and this value was used as their basal expression levels. During the parameter estimation, the initial concentrations of p21-targeting miRNAs were set to 100, because in the publication [23] the expression levels of p21 and mp21 were measured after the individual introduction of the p21-targeting miRNAs with amount of 100 nM. Due to the lack of biological information to characterize the transcriptional activation function (f_{act}) of p21 and its targeting miRNAs, the corresponding functions were assumed to be 1 for simplicity. The data is adapted from our previous publication [4].

Initial concentration of variables and TF functions

Variable	Description	Initial concentration (a.u.)
p21	p21 protein	1
mp21	p21 mRNA	1
$miR_{i=1,\ldots,15}$	p21-targeting miRNAs	100
$[mp21 \mid miR_{i=1,\ldots,15}]$	Complexes formed by miR_i and mp21	0
$f_{act}(TF_{mp21})$	p21's transcriptional activation function	1
$f_{act}\left(TF_{miR_i\ (i=1,\ldots,15)}\right)$	The transcriptional activation function of miR_i	1

Fixed parameter values

Parameter	Description	Value (hr^{-1})	Reference
k_{syn}^{mp21}	Synthesis rate constant of mp21	0.1155	fixed
k_{deg}^{mp21}	Degradation rate constant of mp21	0.1155	[24]
$k_{syn}^{miR_i}$ $(i = 1, \ldots, 15)$	Synthesis rate constant of miR_i	0.0289	fixed
$k_{deg}^{miR_i}$ $(i = 1, \ldots, 15)$	Degradation rate constant of miR_i	0.0289	[25]
k_{syn}^{p21}	Synthesis rate constant of p21	1.3863	fixed
k_{deg}^{p21}	Degradation rate constant of p21	1.3863	[26]

Estimated parameter values

miRNA (state variable)	$k_{deg}^{complex_i}$ $(i = 1, \ldots, 15)$ (hr^{-1})	$k_{ass}^{complex_i}$ $(i = 1, \ldots, 15)$ $(a.u.^{-1} \cdot hr^{-1})$	$F_{cost}^{miR_i}$ $(i = 1, \ldots, 15)$	Experimental data of p21 (protein, mRNA ± SD)
miR-298 (miR_1)*	0.1155	0.0254	$3.4e-004$	(0.16, 1.074 ± 0.025)
miR-208a (miR_2)*	0.1155	0.0041	$2.0e-003$	(0.51, 1.192 ± 0.022)
miR-132 (miR_3)*	0.1155	0.0275	$2.4e-003$	(0.15, 1.21 ± 0.147)
miR-28-5p (miR_4)*	0.1155	0.0119	$5.9e-003$	(0.28, 1.35 ± 0.06)
miR-125-5p (miR_5)*	0.1155	0.0018	$1.8e-003$	(0.69, 0.85 ± 0.051)
miR-299-5p (miR_6)*	0.1155	0.0080	$1.8e-004$	(0.36, 0.95 ± 0.038)
miR-345 (miR_7)*	0.1155	0.0051	$1.1e-004$	(0.46, 0.96 ± 0.039)
miR-93 (miR_8)	0.1564	0.0235	$4.1e-014$	(0.17, 0.7776 ± 0.03)
miR-423-3p (miR_9)	0.9118	0.0055	$2.8e-009$	(0.44, 0.5102 ± 0.11)
miR-515-3p (miR_{10})	0.2098	0.0253	$1.2e-013$	(0.16, 0.616 ± 0.037)
miR-363 (miR_{11})	0.2261	0.0399	$2.2e-014$	(0.11, 0.56 ± 0.15)
mR-657 (miR_{12})	0.3465	0.0158	$2.1e-014$	(0.23, 0.48 ± 0.12)
miR-639 (miR_{13})	0.4305	0.0327	$1.8e-017$	(0.13, 0.36 ± 0.084)
miR-572 (miR_{14})	0.3039	0.0360	$9.4e-023$	(0.12, 0.45 ± 0.044)
miR-654-3p (miR_{15})	9.7485	0.0024	$3.0e-014$	(0.64, 0.63 ± 0.053)

These two miRNAs were chosen, because their predicted target sites in the p21 3′ UTR are in close proximity to each other and thereby, they can induce cooperative repression on p21 as suggested in [24]. The experiments were performed as follows:

(i) Melanoma cells (Sk-Mel-147) were transfected with the mature miRNA mimics of the two miRNAs either individually (100 nM) or in combination (50 nM each), whereas untreated cells were used as control (Figure 2(b)).

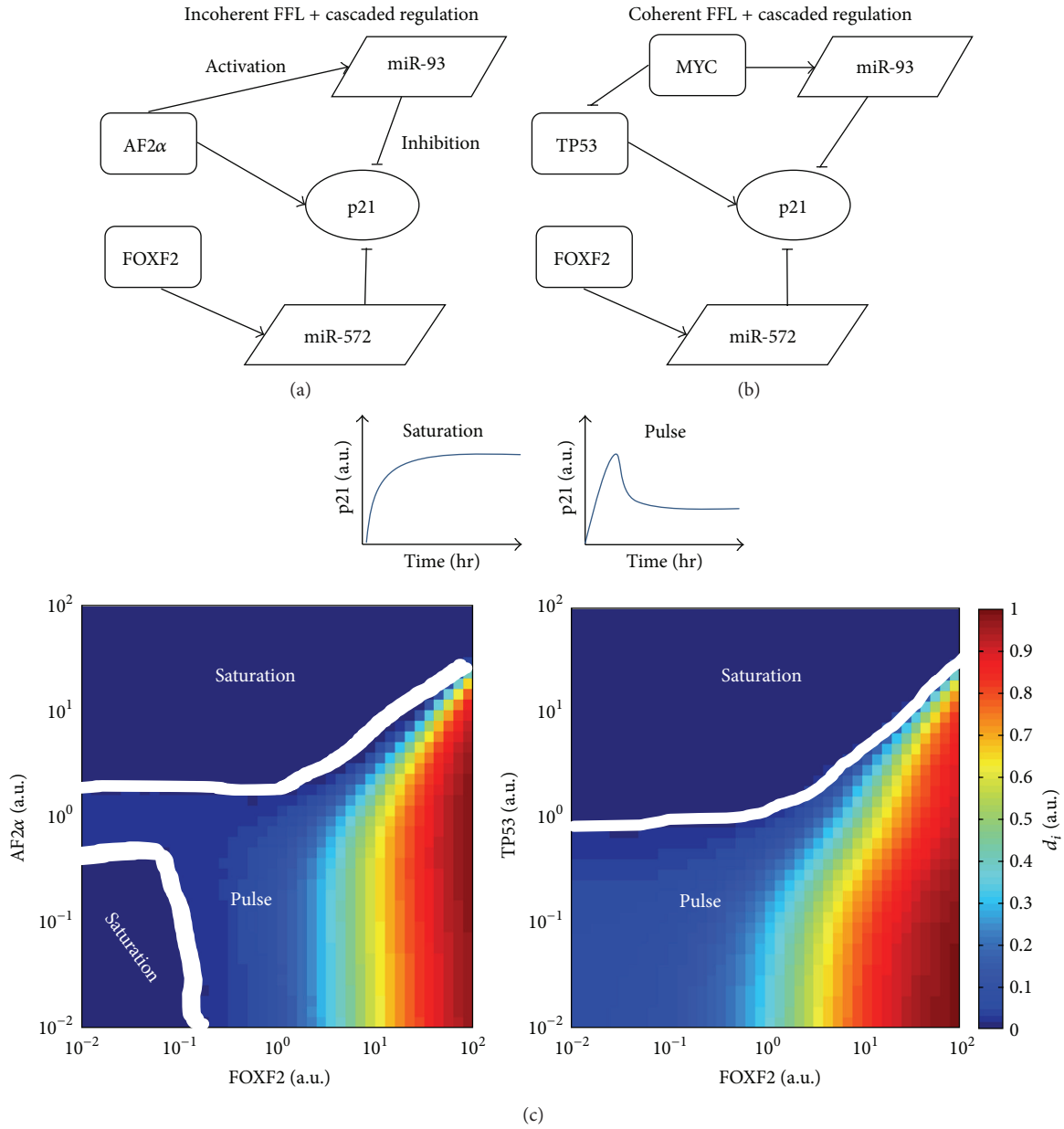


Figure 3: Different p21 dynamics in different network motifs. We ran simulations to show the different dynamics of p21 for two different network motifs (a) and (b). Through simulations, two dynamical patterns of p21 were identified: saturation and pulse ((c), top). For each network motif, the corresponding distributions of the two dynamical patterns were plotted ((c), bottom). For different combinations of the transcriptional strengths, the normalized distance (d_i) between peaks (p_i) and steady states (ss_i) of p21 is determined by the equation $d_i = (p_i - ss_i)/p_{max}$, $p_{max} = \max(p_1, \ldots, p_n)$. If $d_i = 0$, for the corresponding combination of transcriptional strengths the p21 dynamics is saturation, otherwise it is pulse. The regions showing different dynamical patterns of p21 are separated using the white lines.

(ii) Next, the cells were treated with doxorubicin, a genotoxic-stress inducing agent. The agent can upregulate the expression of p53, which is a known TF of p21, and therefore it can result in the upregulation of p21 (Figure 2(c)).

(iii) After doxorubicin treatment, the expression levels of p21 were measured by immunoblotting at different time points (0, 2, 4, 6, 8, 24 hr). The p21 expression values were normalized based on the p21 expression level in the control group measured at time point 0 hr (Figure 2(d)).

By doing so, we obtained the p21 response after genotoxic stress in four scenarios: (1) endogenous miRNA expression (Control); (2) overexpression of miR-572; (3) overexpression of miR-93; and (4) both miRNAs moderately overexpressed. Thereafter, we derived a model of seven ODEs based on the original equations which was configured according to the designed experiments, making the simulation results

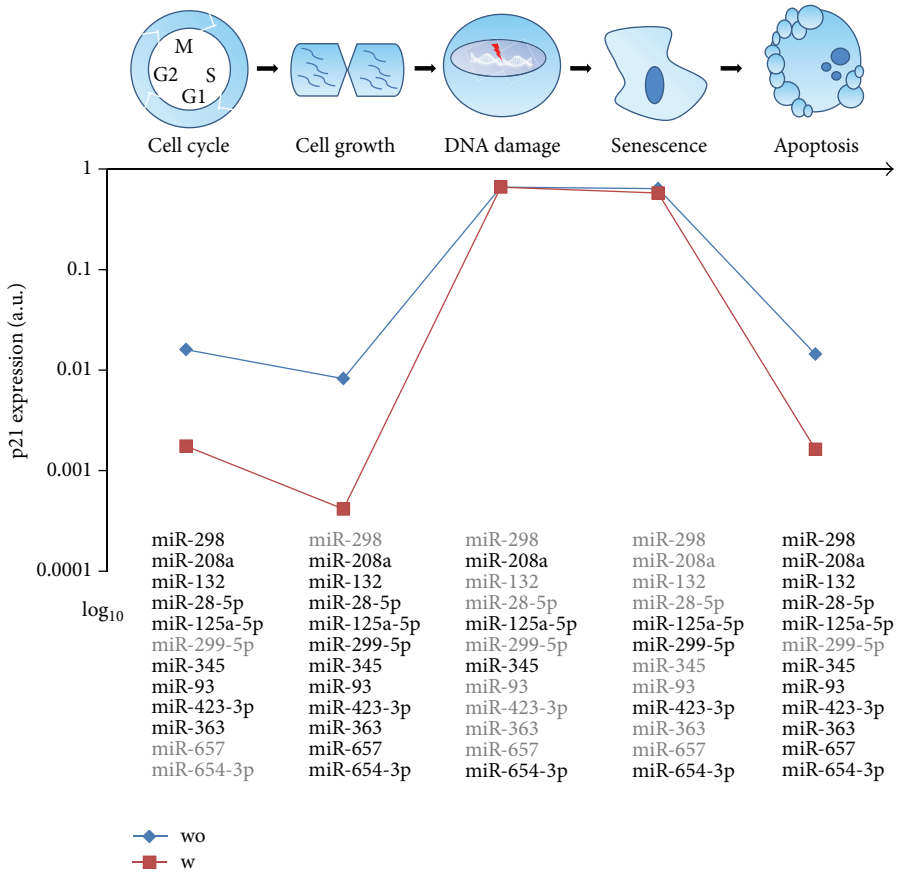


FIGURE 4: p21 expression regulated by cooperative miRNAs for different cellular processes. The associations of the miRNAs with these cellular processes were derived from GO terms of their TFs. A miRNA was supposed to be expressed (in bold black font) in a cellular process only if its TF is related to the corresponding GO term of this process. The p21 expression levels are computed for each process with (w)/without (wo) considering the cooperative effect among the p21 targeting miRNAs.

comparable with the experimental data. As shown in the Figure 2(c), the simulations are in good agreement with the experimental observations. The individual overexpression of miR-572 or miR-93 led to the reduction of the upregulation of p21 response after genotoxic stress induction. The two miRNAs cause different degrees of repression due to their different repression efficiencies on p21. Interestingly, the combined overexpression of both miRNAs induced the strongest downregulation of p21, and therefore verifying the hypothesis of their cooperative regulation of p21. Above all, the results not only validated the model but also demonstrated the ability of our method to identify cooperative miRNA pairs for p21.

(3) Predictive Simulations. As there are abundant of network motifs such as FFLs in p21 regulatory network and these network motifs are important for determining p21 dynamics, it is interesting to investigate the dynamics of p21 in network modules where FFLs are involved. To do so, two network modules including both miR-93 and miR-572, and their TFs were exemplified. In Figure 3(a), the network module contains an incoherent FFL composed by AF2α, miR-93 and p21, and a cascaded regulation in which p21 is repressed by FOXF2 via miR-572; in Figure 3(b), the same cascaded regulation together with a coherent FFL composed of TP53, MYC, miR-93 and p21 forms another regulatory module of p21. By modulating the transcriptional strengths of the two miRNAs by their TFs, two types of p21 dynamics were identified: saturation and pulse (Figure 3(c), top). In the former, the p21 expression increases and reaches its steady state at the highest level; in contrast, in the latter the p21 expression increases to a peak and thereafter drops to a steady state at lower level. For the two different network modules, various combinations of transcriptional strengths of the two miRNAs lead to different distributions of the two p21 dynamical patterns. For the network module of an incoherent FFL plus the cascaded regulation (Figure 3(c), bottom left), the saturation pattern appears in two distinct regions: one with weakened transcriptional strength of miR-93 and the other with enhanced transcriptional strength of miR-93; for the other module (Figure 3(c), bottom right), the saturation pattern only appears in the region, in which the transcriptional strength of miR-93 is enhanced. Taken together, the results showed that for the two different network motifs the dynamical pattern of p21 is changing according to

different combinations of its upstream regulators, suggesting the adaptation of p21 dynamics for different biological contexts.

Furthermore, we performed a number of simulations to show the influence of miRNA regulation on p21 expression levels in different cellular processes arranged in a consecutive manner. In this procedure, a cell is first in the process of cell cycle followed by proliferation (cell growth), then the cell responses to DNA damage and enters into the process of senescence, and finally apoptosis is initiated. As shown in Figure 4, during the cell cycle process, the p21 expression is low due to the activation of most of its targeting miRNAs; when the cell starts proliferating, the p21 expression declines to an even lower level because of more activated miRNAs in this process; after responding to DNA damage, the p21 expression soars to a high level, which is caused by the activation of its TFs like p53 and fewer expressed miRNAs.; although the p21 expression keeps at a similar level while the cell is undergoing senescence, the expressed miRNAs are different from the previous process; finally, the p21 expression decreases again to a low level due to the reemergence of most of its targeting miRNAs and the cell enters apoptosis. Interestingly, the model simulations are also consistent with experimental observations: under non-stressed condition the low expression level of p21 is needed for cell proliferation; the upregulation of p21 happens after response to DNA damage via p53 and the increased p21 expression further results in cell cycle arrest leading to senescence and apoptosis [25]. Besides, when considering the effect of cooperating miRNAs, the p21 expression levels were indistinguishable from the previous simulation of DNA damage response. However, for the other processes p21 expression levels were significantly lower compared to the simulations without considering the cooperative effect of the miRNAs. Above all, these results indicated that selective expression of cooperative miRNAs could be adopted by cells to ensure diverse expression levels of p21 to meet the requirements of different cellular processes.

3. Conclusions and Discussion

In this paper, we presented a systems biology approach, combining data-driven modeling and model-driven experiments, to investigate the role of miRNA-mediated repression in gene regulatory networks. This approach provides a systematic way to gain a deeper understanding of the regulation of target genes by mutiple and cooperative miRNAs. Using the regulation of p21 by multiple miRNAs as a case study, we showed how the ODE-based model, which is calibrated and validated by means of experimental data, is suitable for predicting the temporal dynamics of molecular concentrations involved in biochemical systems.

Provided there are sufficiently rich quantitative data sets avaliable to characterize the model, the use of the methodology here shown can be extended to more complex regulatory networks, involving multiple targets, cooperating TFs and miRNAs and signaling pathways displaying cross-talk via post-tranlational modifications. In this case, the critical element is the quality and quantitiy of the available data. Insufficiency and low quality of experimental data can cause errors in the process of model construction and overfitting in parameter estimation can lead to uncertainties in the model predictions. We believe that the quick development of quantitative high throughput techniques such as transcriptomics, proteomics and miRNomics will facilitate the construction and characterization of larger miRNA-mediated regulatory networks.

Other modeling frameworks than ODE-based models can be used to describe biological systems, such as probabilistic (e.g., Bayesian) or logical (e.g., Boolean) models. Importantly, different modeling frameworks have different properties and perform well regarding different perspectives and levels of mechanistic details of biochemical systems [46]. For example, Bayesian models are helpful in the construction of connections in signaling networks and can reveal the most likely underlying structure of the network in a probabilistic manner. Boolean models use binary values (0 and 1) and logical gates (AND, NOT, and OR) to describe activities of network components and the information flow among them. We believe that in the coming future, hybrid models, which consist of modeling framework and experimental technique specific sub-modules, will provide the necessary compromise between quantitative/qualitative accuracy and scalability for the investigation of large biochemical networks [47].

Disclosures

The authors declare that they have no competing financial interests.

Authors' Contribution

Julio Vera and Olaf Wolkenhauer are equal contributers.

Acknowledgments

The authors would like to acknowledge the funding from the German Federal Ministry of Education and Research (BMBF): eBio-miRSys (0316175A to Xin Lai and Julio Vera), eBio-SysMet (0316171 to Olaf Wolkenhauer) and Gerontosys-ROSAge (0315892A to Olaf Wolkenhauer).

References

[1] A. E. Pasquinelli, "MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship," *Nature Reviews Genetics*, vol. 13, no. 4, pp. 271–282, 2012.

[2] J. Vera, X. Lai, U. Schmitz, and O. Wolkenhauer, "MicroRNA-regulated networks: the perfect storm for classical molecular biology, the ideal scenario for systems biology," *Advances in Experimental Medicine and Biology*, vol. 774, pp. 55–76, 2013.

[3] S. Nikolov, J. Vera, U. Schmitz, and O. Wolkenhauer, "A model-based strategy to investigate the role of microRNA regulation in cancer signalling networks," *Theory in Biosciences*, vol. 130, no. 1, pp. 55–69, 2011.

[4] X. Lai, U. Schmitz, S. K. Gupta et al., "Computational analysis of target hub gene repression regulated by multiple and cooperative miRNAs," *Nucleic Acids Research*, vol. 40, no. 18, pp. 8818–8834, 2012.

[5] X. Lai, O. Wolkenhauer, and J. Vera, "Modeling miRNA regulation in cancer signaling systems: miR-34a regulation of the p53/Sirt1 signaling module," *Methods in Molecular Biology*, vol. 880, pp. 87–108, 2012.

[6] C. Jiang, Z. Xuan, F. Zhao, and M. Q. Zhang, "TRED: a transcriptional regulatory element database, new entries and other development," *Nucleic Acids Research*, vol. 35, supplement 1, pp. D137–D140, 2007.

[7] L. A. Bovolenta, M. L. Acencio, and N. Lemke, "HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions," *BMC Genomics*, vol. 13, article 405, 2012.

[8] V. Matys, O. V. Kel-Margoulis, E. Fricke et al., "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes," *Nucleic Acids Research*, vol. 34, database issue, pp. D108–D110, 2006.

[9] D. Karolchik, R. Baertsch, M. Diekhans et al., "The UCSC genome browser database," *Nucleic Acids Research*, vol. 31, no. 1, pp. 51–54, 2003.

[10] P. Alexiou, T. Vergoulis, M. Gleditzsch et al., "miRGen 2.0: a database of microRNA genomic information and regulation," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D137–D141, 2010.

[11] 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, supplement 1, pp. D105–D110, 2009.

[12] S.-D. Hsu, F.-M. Lin, W.-Y. Wu et al., "miRTarBase: a database curates experimentally validated microRNA-target interactions," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D163–D169, 2011.

[13] P. Sethupathy, B. Corda, and A. G. Hatzigeorgiou, "TarBase: a comprehensive database of experimentally supported animal microRNA targets," *RNA*, vol. 12, no. 2, pp. 192–197, 2006.

[14] H. Dweep, C. Sticht, P. Pandey, and N. Gretz, "miRWalk—database: prediction of possible miRNA binding sites by "walking" the genes of three genomes," *Journal of Biomedical Informatics*, vol. 44, no. 5, pp. 839–847, 2011.

[15] T. S. Keshava Prasad, R. Goel, K. Kandasamy et al., "Human protein reference database—2009 update," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D767–D772, 2009.

[16] 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, supplement 1, pp. D561–D568, 2011.

[17] D. Croft, G. O'Kelly, G. Wu et al., "Reactome: a database of reactions, pathways and biological processes," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D691–D697, 2011.

[18] N. le Novère, M. Hucka, H. Mi et al., "The systems biology graphical notation," *Nature Biotechnology*, vol. 27, no. 8, pp. 735–741, 2009.

[19] A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi, and H. Kitano, "CellDesigner 3.5: a versatile modeling tool for biochemical networks," *Proceedings of the IEEE*, vol. 96, no. 8, pp. 1254–1265, 2008.

[20] M. S. Cline, M. Smoot, E. Cerami et al., "Integration of biological networks and gene expression data using Cytoscape," *Nature Protocols*, vol. 2, no. 10, pp. 2366–2382, 2007.

[21] D.-H. Le and Y.-K. Kwon, "NetDS: a Cytoscape plugin to analyze the robustness of dynamics and feedforward/feedback loop structures of biological networks," *Bioinformatics*, vol. 27, no. 19, pp. 2767–2768, 2011.

[22] E. O. Voit, *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*, Cambridge University Press, Cambridge, UK, 2000.

[23] S. Wu, S. Huang, J. Ding et al., "Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 3′ untranslated region," *Oncogene*, vol. 29, no. 15, pp. 2302–2308, 2010.

[24] P. Sætrom, B. S. E. Heale, O. Snøve Jr., L. Aagaard, J. Alluin, and J. J. Rossi, "Distance constraints between microRNA target sites dictate efficacy and cooperativity," *Nucleic Acids Research*, vol. 35, no. 7, pp. 2333–2342, 2007.

[25] Y.-S. Jung, Y. Qian, and X. Chen, "Examination of the expanding pathways for the regulation of p21 expression and activity," *Cellular Signalling*, vol. 22, no. 7, pp. 1003–1012, 2010.

[26] W. Wang, H. Furneaux, H. Cheng et al., "HuR regulates p21 mRNA stabilization by UV light," *Molecular and Cellular Biology*, vol. 20, no. 3, pp. 760–769, 2000.

[27] U. Wittig, R. Kania, M. Golebiewski et al., "SABIO-RK—database for biochemical reaction kinetics," *Nucleic Acids Research*, vol. 40, database issue, pp. D790–D796, 2012.

[28] R. Milo, P. Jorgensen, U. Moran, G. Weber, and M. Springer, "BioNumbers—the database of key numbers in molecular and cell biology," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D750–D753, 2010.

[29] I.-C. Chou and E. O. Voit, "Recent developments in parameter estimation and structure identification of biochemical and genomic systems," *Mathematical Biosciences*, vol. 219, no. 2, pp. 57–83, 2009.

[30] A. Saltelli, K. Chan, and E. M. Scott, *Sensitivity Analysis*, John Wiley & Sons, New York, NY, USA, 1st edition, 2000.

[31] S. Strogatz, *Nonlinear Dynamics and Chaos: Applications to Physics, Biology, Chemistry, and Engineering: With Applications to Physics, Biology, Chemistry and Engineering*, Westview Press, Boulder, Colo, USA, 2000.

[32] P. Zhou, S. Cai, Z. Liu, and R. Wang, "Mechanisms generating bistability and oscillations in microRNA-mediated motifs," *Physical Review E*, vol. 85, no. 4, apart 1, Article ID 041916, 9 pages, 2012.

[33] S. Marino, I. B. Hogue, C. J. Ray, and D. E. Kirschner, "A methodology for performing global uncertainty and sensitivity analysis in systems biology," *Journal of Theoretical Biology*, vol. 254, no. 1, pp. 178–196, 2008.

[34] 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.

[35] J. Krützfeldt, N. Rajewsky, R. Braich et al., "Silencing of microRNAs in vivo with 'antagomirs'," *Nature*, vol. 438, no. 7068, pp. 685–689, 2005.

[36] M. W. Pfaffl, G. W. Horgan, and L. Dempfle, "Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR," *Nucleic Acids Research*, vol. 30, no. 9, p. e36, 2002.

[37] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, 2009.

[38] T. Abbas and A. Dutta, "p21 in cancer: intricate networks and multiple activities," *Nature Reviews Cancer*, vol. 9, no. 6, pp. 400–414, 2009.

[39] J. Wang, M. Lu, C. Qiu, and Q. Cui, "TransmiR: a transcription factor microRNA regulation database," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D119–D122, 2010.

[40] S. Bandyopadhyay and M. Bhattacharyya, "PuTmiR: a database for extracting neighboring transcription factors of human microRNAs," *BMC Bioinformatics*, vol. 11, article 190, 2010.

[41] A. Le Béchec, E. Portales-Casamar, G. Vetter et al., "MIR@NT@N: a framework integrating transcription factors, microRNAs and their targets to identify sub-network motifs in a meta-regulation network model," *BMC Bioinformatics*, vol. 12, article 67, 2011.

[42] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[43] A. I. F. Vaz and L. N. Vicente, "A particle swarm pattern search method for bound constrained global optimization," *Journal of Global Optimization*, vol. 39, no. 2, pp. 197–219, 2007.

[44] W. H. Press, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge, UK, 2007.

[45] J. G. Doench and P. A. Sharp, "Specificity of microRNA target selection in translational repression," *Genes and Development*, vol. 18, no. 5, pp. 504–511, 2004.

[46] B. Kholodenko, M. B. Yaffe, and W. Kolch, "Computational approaches for analyzing information flow in biological networks," *Science Signaling*, vol. 5, no. 220, 2012.

[47] F. M. Khan, U. Schmitz, S. Nikolov et al., "Hybrid modeling of the crosstalk between signaling and transcriptional networks using ordinary differential equations and multi-valued logic," *Biochimica et Biophysica Acta*, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 870795, 7 pages
http://dx.doi.org/10.1155/2013/870795

Research Article

Prediction of Gene Phenotypes Based on GO and KEGG Pathway Enrichment Scores

Tao Zhang,[1] Min Jiang,[2] Lei Chen,[3] Bing Niu,[4] and Yudong Cai[1]

[1] *Institute of Systems Biology, Shanghai University, 99 ShangDa Road, Shanghai 200444, China*
[2] *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*
[3] *College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*
[4] *College of Life Science, Shanghai University, 99 ShangDa Road, Shanghai 200444, China*

Correspondence should be addressed to Bing Niu; bingniu@shu.edu.cn and Yudong Cai; cai_yud@126.com

Received 19 August 2013; Accepted 23 September 2013

Academic Editor: Tao Huang

Observing what phenotype the overexpression or knockdown of gene can cause is the basic method of investigating gene functions. Many advanced biotechnologies, such as RNAi, were developed to study the gene phenotype. But there are still many limitations. Besides the time and cost, the knockdown of some gene may be lethal which makes the observation of other phenotypes impossible. Due to ethical and technological reasons, the knockdown of genes in complex species, such as mammal, is extremely difficult. Thus, we proposed a new sequence-based computational method called *k*NNA-based method for gene phenotypes prediction. Different to the traditional sequence-based computational method, our method regards the multiphenotype as a whole network which can rank the possible phenotypes associated with the query protein and shows a more comprehensive view of the protein's biological effects. According to the prediction result of yeast, we also find some more related features, including GO and KEGG information, which are making more contributions in identifying protein phenotypes. This method can be applied in gene phenotype prediction in other species.

1. Introduction

Recognition of gene phenotypes of proteins is a central challenge of the modern genetics to modulate protein functions and biological processes, and many well-known diseases, such as HIV [1–4], cancers [5–8], chronic liver diseases [9], and Gaucher disease [10], are all closed to protein phenotypes. Hence, determination of protein's phenotypes is quite fundamental and essential in systems biology and proteomics. Except for phenotypes attributes, there are also many other multilabel attributes of proteins, such as subcellular locations [11–13] and multiple functional types of antimicrobial peptides. Multilabel molecule biosystems are very common.

During the past decades, numerous efforts have been made in the prediction of gene phenotype of yeast protein based on the following approaches: experimental methods and computational methods. As for experimental approaches, the high-throughput phenotype assays [14, 15] combining with gene perturbation technology [16, 17] provide fast identification for active gene in a response [18]. For example, using yeast mutant strain collections identifies the phenotypes [19]. However, due to the high complexity of phenotypes, it is both costly and time-consuming to determine protein phenotypes by experiments. Sometimes, the results derived from experiment are even of high false rates [20]. Computational methods provide important complementary tools for this problem. Many studies based on sequence-based methods and network-based methods have been made in protein's gene phenotypes identification [21–23]. In this research, we presented a new sequence-based method called *k*NNA-based method to predict gene phenotypes.

2. Materals and Methods

2.1. Benchmark Dataset. In this study, 6,732 proteins of yeast were taken from CYGD (the MIPS Comprehensive Yeast

Table 1: Breakdown of 1462 budding yeast proteins according to their 11 phenotypes.

Tag	Phenotype category	Number of proteins
T_1	Conditional phenotypes	536
T_2	Cell cycle defects	272
T_3	Mating and sporulation defects	198
T_4	Auxotrophies, carbon, and nitrogen utilization defects	266
T_5	Cell morphology and organelle mutants	535
T_6	Stress response defects	147
T_7	Carbohydrate and lipid biosynthesis	46
T_8	Nucleic acid metabolism defects	219
T_9	Sensitivity to amino acid analogs and other drugs	124
T_{10}	Sensitivity to antibiotics	43
T_{11}	Sensitivity to immunosuppressants	14
Total	—	2,400

Genome Database [24], which collects information on the molecular structure and functional network of the budding yeast. After removing those without sequences, information, or phenotype annotations, the remaining 1,462 composed the benchmark dataset S. According to their phenotypes, these proteins were classified into the following 11 categories: (I) conditional phenotypes, (II) cell cycle defects, (III) mating and sporulation defects, (IV) auxotrophies, carbon and nitrogen utilization defects, (V) cell morphology and organelle mutants, (VI) stress response defects, (VII) carbohydrate and lipid biosynthesis, (VIII) nucleic acid metabolism defects, (IX) sensitivity to amino acid anaglogs and other drugs, (X) sensitivity to antibiotics. (XI) sensitivity to immunosuppressants. Let us use $T_1, T_2, \ldots, T_{11}$ to represent the tags of the 11 phenotypic categories, where T_1 denotes "conditional phenotypes," T_2 denotes "cell cycle defects," and so forth (see column 1 and 2 of Table 1 for the correspondence of tags and phenotypic categories). Thus, the benchmark dataset S can be formulated as

$$S = S_1 \cup S_2 \cup \cdots \cup S_{11}, \tag{1}$$

where S_i represents the set of proteins with tag T_i. The IDs of proteins in each S_i are available online in Supplementary Material at http://dx.doi.org/10.1155/2013/870795. From Table 1, we can see that the total number of proteins in each category is much larger than the total number of proteins investigated in this study, this means that some proteins are associated with multiple phenotypes. Like the cases in dealing with the proteins or compounds with multiple attributes [25–29], the proposed method could predict multiclassification phenotypes.

2.2. Feature Construction. The first important step to build an efficient prediction model is to encode each sample by numeric vector. Here, to catch the information of protein phenotype, Gene Ontology (GO) and KEGG enrichment scores were employed to represent the protein, which have been used in some biological problems [30, 31]. Their detailed definition can be found at [30, 31].

2.3. Protein Representation and Feature Reduction. Each protein was represented with 4682 features which include 4583 GO enrichment scores and 99 KEGG enrichment scores. However, among the 4,682 features, some features were with little relationship to the target, which may bring noises to the prediction model. Therefore, these features should be removed. Before removing the irrelevant features, the following formula was used to adjust all features to a standard scale:

$$U_{ij} = \frac{\left(u_{ij} - u_j\right)}{T_j}, \tag{2}$$

where T_j and u_j are the standard deviation and mean value of the jth feature, while u_{ij} and U_{ij} are the original value and standardized value of the ith sample on the jth feature.

After the transformation, the correlation coefficient between each feature with the target vector was computed and those with correlation coefficient less than 0.1 were discarded. Finally, 989 features remained. Within these 989 features, there were 947 Gene Ontology (GO) enrichment scores and 42 KEGG enrichment scores. Thus, each protein P_z was finally represented by a 989-D vector.

2.4. mRMR Method. Minimum Redundancy Maximum Relevance (mRMR), first proposed by Peng et al. [32], is an effective algorithm to identify discriminative features. The detailed algorithm of mRMR can be found at [32] and its program can be downloaded from http://penglab.janelia.org/proj/mRMR/.

mRMR has been widely used in the areas of bioinformatics [25, 33–36].

2.5. Prediction Model

2.5.1. kNNA-Based Method. Nearest neighbor algorithm is effective in solving classification and optimization problems in the field of bioinformatics due to its simplicity. It is adopted here to construct the multilabel prediction classifier.

Within k-NNA method, we used the cosine of the angle between two vectors to measure the similarity between them as follows:

$$\mathrm{Cos}\left\langle p_x, p_y \right\rangle = \frac{\vec{p}_x \cdot \vec{p}_y}{\left\|\vec{p}_x\right\| \cdot \left\|\vec{p}_y\right\|}, \tag{3}$$

where $\vec{p}_x \cdot \vec{p}_y$ represents the inner product between the n-dimensional vector of protein p_x and p_y and $\|p\|$ is the modulus of the vector.

For a query protein, k proteins in the training set which are closest to the query protein are first identified and are denoted by $p_1, p_2, \ldots, p_k$. Then, the categories of the query protein can be inferred from the categories of the k nearest proteins identified. The procedure of the methodology is described in detail as follows.

(a) Identifying the k nearest neighbors of the query protein, denoted by $p_1, p_2, \ldots, p_k$, with the k cosines of angle values as $w_1, w_2, \ldots, w_k$.

(b) Then, the following formula:

$$S\left(P \Longrightarrow j\right) = \sum_{i=1}^{k} w_i \cdot t_{p_i,j} \quad \left(j = 1, 2, \ldots, 11\right) \tag{4}$$

is used to calculate the probability that the query protein P belongs to the jth category, where $t_{p_i,j}$ is the item in t_{p_i} of protein p_i.

The probabilities (the scores of the 11 categories) calculated above are sorted in descending order for each query protein as

$$D^{\downarrow}\left\{S\left(P_z \Longrightarrow j\right) \mid j = 1, 2, \ldots, 11\right\} = V = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_j \\ \vdots \\ \mu_{10} \\ \mu_{11} \end{bmatrix}. \tag{5}$$

(c) The corresponding category labels of the category scores are denoted as

$$P^{D^{\downarrow}} = \left[P^{\mu_1}, P^{\mu_2}, \ldots, P^{\mu_i}, \ldots, P^{\mu_{11}}\right] \quad (i = 1, 2, \ldots, 11), \tag{6}$$

where P^{μ_i} is the class that scores ith in $D^{\downarrow}$.

2.5.2. Comparison with RPC-Based Method. In the ranking by pairwise comparison (RPC) method, for each pair of labels, a data is allocated to the pair of labels if the data belong to one and only one of the two labels (not both). Given q category labels, because there are $C_q^2 = q \cdot (q-1)/2$ possible pairwise combinations of the labels, data subsets, each for corresponding pairwise labels discrimination, are generated.

Given a new instance, all pairwise classifiers are trained to predict its label, and the ranking of the labels is obtained by counting the votes of each label, where if the instance is classified into a label, the label receives one vote.

Each dataset contains those examples of D that are annotated by at least one of the two corresponding labels, but not both. A binary classifier that learns to discriminate between the two labels is trained from each of these data sets. Given a new instance, all binary classifiers are invoked, and ranking is obtained by counting the votes received by each label.

2.6. Evaluation

(a) Jackknife Testing. Three methods are often used to evaluate a prediction model, including (1) independent test dataset, (2) subsampling (K-fold) test, and (3) jackknife Test. The first method uses unseen data for testing, which needs a large quantity of data. The second method partitions the training set into k portions, then taking each portion of the data as the test data and the others $(k-1)$ as the training data. The third one, also named as leave-one-out method, leaves each sample out in turn as the test data and others as the training data. To maximize the quantity of the training data, jackknife test is used to test the predictor developed in the paper; that is, each protein is in turn knocked out as the query protein, and the remaining ones as the training data of the kNNA-based method.

(b) Metric. Let us define $t_{z,P_z^{\mu_i}} = 1$ as protein P_z being correctly predicted to class $p_z^{\mu_1}$; otherwise, $t_{z,p_z^{\mu_i}} = 0$.

The ith prediction accuracy A^i is calculated as follows (the ith order predictions in $P^{D^{\downarrow}}$):

$$A^i = \frac{\sum_{j=1}^{m} t_{j,p_j^{\mu_i}}}{m}, \tag{7}$$

where m is the number of the training data.

2.7. Incremental Feature Selection. Incremental feature selection (IFS) is often used to search out an optimal feature subset that performs best. Specifically, features in the ranked feature set are added one by one from higher to lower rank and the first n features that perform best are regarded as the optimal features. When one feature is added, a new feature subset is constructed. Thus, given N features, N feature subsets will be constructed, where the ith -order feature subset is

$$S_i = \left\{f_1, f_2, \ldots, f_i\right\} \quad (1 \le i \le 989), \tag{8}$$

in which f_i represents the ith feature taken from the mRMR ranking.

Each feature subset is used to make prediction and the feature subset (first n features) that performs best is deemed as the optimal feature subset.

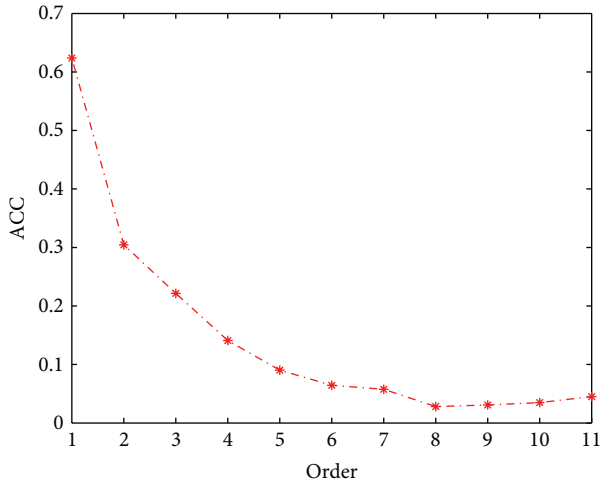


Figure 1: The curve showing the trend of the 11 order prediction accuracies.

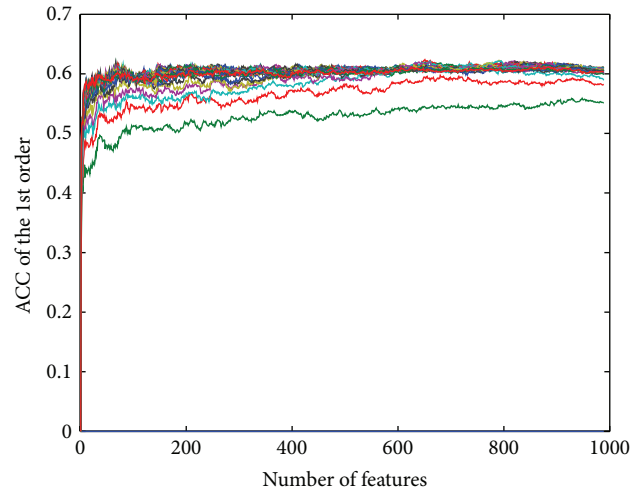


Figure 2: 30 IFS curves of kNNA-based method corresponding to different values of k.

3. Results and Discussion

3.1. Results

3.1.1. mRMR Results. We apply mRMR method to the dataset, and obtain two tables for the features (see Supplementary Material). One is called MaxRel feature table that ranks the features based on their relevance to the class of samples and the other is called mRMR feature table that lists the ranked features by the maximum relevance and minimum redundancy to the class of samples. Such list of ranked features was to be used in the following IFS procedure for the optimal features set selection.

3.1.2. Performance of kNNA-Based Method. The first-order prediction accuracy of Jackknife test is 62.38%, while $k = 17$ (k-NN) and $n = 651$ (number of optimal features). More details of the 11 order prediction accuracies by using kNNA-based method are listed in Table 2 and Figure 1. IFS curve of kNNA-based method can be seen in Figure 2, which contains 30 curves corresponding to different values of k, and their detailed computing results of accuracy (ACC) can be seen at Supplementary Material. We highlighted the peak area of these curves to find optimal k in Figure 3.

3.1.3. Performance of RPC-Based Method. Firstly, we classify the total labels into $55(C_{11}^5)$ sublabels. Select the sample which meets the demands that one sample belongs to one and only one of the two labels (not both). Then, 55 binary subsets were constructed. Three well-known binary classification algorithms including RandomForest, SMO, and Dagging were applied to build the prediction model. The prediction results are summarized in Table 3.

3.1.4. Comparison with RPC-Based Method. We compared the first-order prediction accuracy of our method with the first-order prediction accuracy of RPC-based method. It can be found that the first-order prediction accuracies of RPC-based method using Dagging, RandomForest, and SMO are all lower than our kNNA-based method.

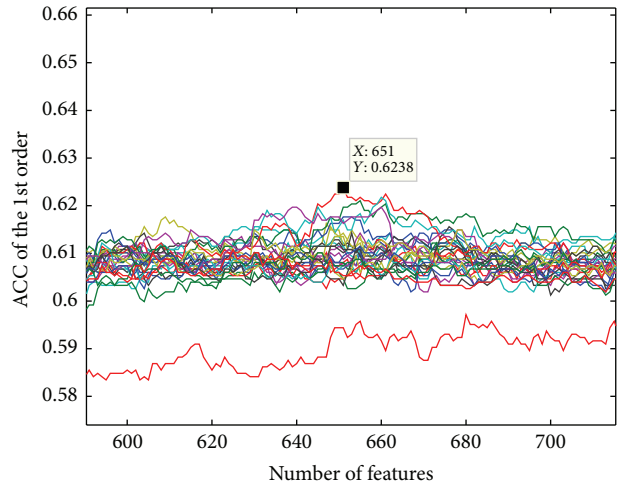


Figure 3: The peak and its coordinate of these IFS curves.

3.2. Discussion.

To illustrate the biological meanings of the selected optimal feature subset, we firstly classified GO terms into three kinds: the biological process, cellular component, and molecular function GO terms. The 622 GO terms in the mRMR feature list were mapped to the Gene Ontology (GO) terms, the children of the three root GO terms. The figures show the frequency of each GO term in the feature subset, and display the ratio of the number of each GO term to the scale of the number of its children terms.

3.2.1. Biological Process GO Terms. In BP frequency, the top five GO biological process terms are GO:0009987: cellular process (399), GO:0008152: metabolic process (316),

Table 2: The 11 order prediction accuracies by *k*NNA-based method.

	Method order										
	1	2	3	4	5	6	7	8	9	10	11
*k*NN-based method (ACC)	62.38	30.44	22.16	14.09	9.03	6.43	5.75	2.8	3.08	3.49	4.51

Table 3: The 11 order prediction accuracies by RPC-based methods (Dagging, RandomForest, SMO).

	Methods order										
	1	2	3	4	5	6	7	8	9	10	11
Dagging	60.05	33.58	21.96	13.75	10.53	8.28	6.57	3.56	2.6	1.85	1.44
RandomForest	58.62	34.2	22.3	14.7	9.92	7.66	5.95	5.2	3.28	1.5	0.82
SMO	56.16	34.68	21.55	14.84	10.88	7.8	6.36	4.65	3.21	2.26	1.78

GO:0019740: nitrogen utilization (216), GO:0065007: biological regulation (136), and GO:0050789: regulation of biological process (131). In BP percentage, the top five GO biological processes are GO:0019740: nitrogen utilization (4.20%), GO:0071840: cellular component organization or biogene (3.57%), GO:0000003: reproduction (2.94%), GO:0022414: reproductive process (2.88%), and GO:0009987: cellular process (2.04%). For both GO biological process term number and percentage distribution analysis, the GO terms corresponding to the nitrogen utilization (GO:0019740) and cellular process (GO:0009987) were highlighted within the top five GO terms. This indicates that proteins assigned with these two GO terms may affect protein phenotype determination greatly. This conclusion is consistent with the common knowledge that specific cellular biological activities of the proteins confer with special phenotypes. It was also reported by Granek and Magwene that two key signaling networks: the filamentous growth MAP kinase cascade and the Ras-cAMP-PKA pathway, can regulate the yeast colony morphology response [37]. Additionally, the yeast cell wall integrity pathway was involved in resistance of the yeast *Saccharomyces cerevisiae* to the biocide polyhexamethylene biguanide [38].

The highlight of nitrogen utilization (GO:0019740) suggests that the nitrogen utilization, which is essential for life survival and development, may have more definite affection on protein phenotype. Nutrient stresses trigger a variety of developmental switches in the budding yeast *Saccharomyces cerevisiae*. It was demonstrated that low levels of carbon combined with abundant nitrogen trigger complex colony formation in yeast [37].

3.2.2. Cellular Component GO Terms. In CC frequency, the top six GO cellular component terms are GO:0005623: cell (171), GO:0044464: cell part (169), GO:0043226: organelle (135), GO:0044422: organelle part (103), GO:0032991: macromolecular complex (84), and GO:0031974: membrane-enclosed lumen (39). In CC percentage, the top six GO cellular component terms are GO:0031974: membrane-enclosed lumen (12.4%), GO:0044422: organelle part (8.42%), GO:0043226: organelle (8.4%), GO:0032991: macromolecular complex (5.20%), GO:0044464: cell part (4.77%), and GO:0005623: cell (4.20%). For both GO cellular component term number and percentage distribution analysis, the GO terms corresponding to the organelle (GO:0043226) and organelle part (GO:0044422) were highlighted within the top six GO terms. It may be concluded that proteins located in all cellular organelles should be guaranteed. It suggests that organelles, which have specific structural and functional attributes, may possess more definite protein phenotype to carry out their specific functions. This also implicated that proteins assigned to these GO terms could contribute relatively more to the overall protein phenotype determination. For example, the communication between mitochondrial and nuclear loci (i.e., *COX1-MSY1* and *Q0182-RSM7*) showed significant reductions in the absence of mitochondrial encoded reverse transcriptase machinery [39]. The inclusion of macromolecular complex (GO:0032991) suggests that proteins expressing some phenotype need to interact with each other to function together and that macromolecular complex should certainly determine the phenotype of proteins. The inclusion of membrane-enclosed lumen (GO:0031974) also suggests that proteins assigned to this cellular component could greatly contribute to protein phenotype, because most of the cellular organelles are enclosed by membrane, such as mitochondrial and nucleus.

3.2.3. Molecular Function GO Terms. In MF frequency, the top six GO molecular function terms are GO:0003824: catalytic activity (79), GO:0005488: binding (69), GO : 0001071: nucleic acid binding transcription factor activity (40), GO:0000988: protein binding transcription factor activity (14). GO:0065009: regulation of molecular function (8), and GO:0005215: transporter activity (7). Proteins assigned to these three GO terms required binding or interaction to carry out their structural or functional activities. This suggests that proteins assigned to these six GO terms contributed profoundly to the protein phenotype. In MF percentage, the top six GO molecular function terms are GO:0009055: electron carrier activity (25%), GO:0016530: metallochaperone activity (25%), GO:0045182: translation regulator activity (14.3%), GO:0005198: structural molecule activity (11.8%), GO:0001071: nucleic acid binding transcription factor activity

(9.0%), GO:0005488: binding (3.99%), and GO:0016209: antioxidant activity (3.85%). The relatively small base number made protein GO terms influencing protein phenotype relatively more enriched in the top six molecular function GO terms, especially in electron carrier activity (GO:0009055) and metallochaperone activity (GO:0016530). The highlight of electron carrier activity (GO:0009055) may be attributed to the relatively limited and definite function of these proteins. It was reported that some ontology drug can interact with the electron transport chain (ETC) to generate high levels of ROS within the organelle and consequently cell leads to death [40]. The highlight of metallochaperone activity (GO:0016530) may be ascribed to that metalloprotein used to express specific function with metallochaperone and metallic ion. In all bacteria, a panel of metalloregulatory proteins controls the expression of genes encoding membrane transporters and metal trafficking proteins [41]. Because of the large base number of the top six GO terms in MF frequency, they have relatively lower enrichment within the top eight GO terms in MF percentage.

Authors' Contribution

Tao Zhang and Min Jiang contributed equally to this research.

Acknowledgments

This work was supported by grants from the National Basic Research Program of China (2011CB510101, 2011CB510102), the National Natural Science Foundation of China (31371335), the Innovation Program of Shanghai Municipal Education Commission (12ZZ087), the Leading Academic Discipline Project of Shanghai Municipal Education Commission "Molecular Physiology," the grant of "The First-class Discipline of Universities in Shanghai," and the Foundation for The Excellent Youth (SHU10022).

References

[1] M. van Houtte, G. Picchio, K. van der Borght, T. Pattery, P. Lecocq, and L. T. Bacheler, "A comparison of HIV-1 drug susceptibility as provided by conventional phenotyping and by a phenotype prediction tool based on viral genotype," *Journal of Medical Virology*, vol. 81, no. 10, pp. 1702–1709, 2009.

[2] A. V. Vasil'ev, E. V. Kazennova, and M. R. Bobkova, "Prediction of phenotype R5/X4 of HIV-1 variants circulating in Russia, by using computer methods," *Voprosy Virusologii*, vol. 54, no. 3, pp. 17–21, 2009.

[3] S. Xu, X. Huang, H. Xu, and C. Zhang, "Improved prediction of coreceptor usage and phenotype of HIV-1 based on combined features of V3 loop sequence using random forest," *Journal of Microbiology*, vol. 45, no. 5, pp. 441–446, 2007.

[4] H. Vermeiren, E. Van Craenenbroeck, P. Alen, L. Bacheler, G. Picchio, and P. Lecocq, "Prediction of HIV-1 drug susceptibility phenotype from the viral genotype using linear regression modeling," *Journal of Virological Methods*, vol. 145, no. 1, pp. 47–55, 2007.

[5] T.-Y. Lin, J. T.-C. Chang, H.-M. Wang et al., "Proteomics of the radioresistant phenotype in head-and-neck cancer: GP96 as a novel prediction marker and sensitizing target for radiotherapy," *International Journal of Radiation Oncology Biology Physics*, vol. 78, no. 1, pp. 246–256, 2010.

[6] T. F. Bathen, L. R. Jensen, B. Sitter et al., "MR-determined metabolic phenotype of breast cancer in prediction of lymphatic spread, grade, and hormone status," *Breast Cancer Research and Treatment*, vol. 104, no. 2, pp. 181–189, 2007.

[7] S. R. Lakhani, J. S. Reis-Filho, L. Fulford et al., "Prediction of BRCA1 status in patients with breast cancer using estrogen receptor and basal phenotype," *Clinical Cancer Research*, vol. 11, no. 14, pp. 5175–5180, 2005.

[8] T. Dwyer, J. M. Stankovich, L. Blizzard et al., "Does the addition of information on genotype improve prediction of the risk of melanoma and nonmelanoma skin cancer beyond that obtained from skin phenotype?" *American Journal of Epidemiology*, vol. 159, no. 9, pp. 826–833, 2004.

[9] L. A. Piruzyan, I. B. Korshunov, N. V. Morozova, N. E. Pyn'ko, and L. A. Radkevich, "Prediction of chronic liver diseases on the basis of the N-acetyltransferase 2 phenotype," *Doklady Biochemistry and Biophysics*, vol. 395, no. 1–6, pp. 84–87, 2004.

[10] P. D. Whitfield, P. Nelson, P. C. Sharp et al., "Correlation among genotype, phenotype, and biochemical markers in Gaucher disease: implications for the prediction of disease severity," *Molecular Genetics and Metabolism*, vol. 75, no. 1, pp. 46–55, 2002.

[11] G.-Z. Li, X. Wang, X. Hu, J.-M. Liu, and R.-W. Zhao, "Multilabel learning for protein subcellular location prediction," *IEEE Transactions on NanoBioscience*, vol. 11, no. 3, pp. 237–243, 2012.

[12] 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.

[13] 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 and Peptide Letters*, vol. 20, no. 3, pp. 309–317, 2013.

[14] B. L. Drees, V. Thorsson, G. W. Carter et al., "Derivation of genetic interaction networks from quantitative phenotype data," *Genome Biology*, vol. 6, no. 4, p. R38, 2005.

[15] A. M. Dudley, D. M. Janse, A. Tanay, R. Shamir, and G. M. Church, "A global view of pleiotropy and phenotypically derived gene function in yeast," *Molecular Systems Biology*, vol. 1, Article ID 2005.0001, 11 pages, 2005.

[16] 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.

[17] E. A. Winzeler, H. Liang, D. D. Shoemaker, and R. W. Davis, "Functional analysis of the yeast genome by precise deletion and parallel phenotypic characterization," *Novartis Foundation Symposium*, vol. 229, pp. 105–109, 2000, discussion 109–111.

[18] G. W. Carter, S. Prinz, C. Neou et al., "Prediction of phenotype and gene expression for combinations of mutations," *Molecular Systems Biology*, vol. 3, p. 96, 2007.

[19] B. Scherens and A. Goffeau, "The uses of genome-wide yeast mutant collections," *Genome Biology*, vol. 5, no. 7, article 229, 2004.

[20] K. L. McGary, I. Lee, and E. M. Marcotte, "Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes," *Genome Biology*, vol. 8, no. 12, article R258, 2007.

[21] J. Cedano, P. Aloy, J. A. Perez-Pons, and E. Querol, "Relation between amino acid composition and cellular location of proteins," *Journal of Molecular Biology*, vol. 266, no. 3, pp. 594–600, 1997.

[22] W. Resch, N. Hoffman, and R. Swanstrom, "Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks," *Virology*, vol. 288, no. 1, pp. 51–62, 2001.

[23] S. Pillai, B. Good, D. Richman, and J. Corbeil, "A new perspective on V3 phenotype prediction," *AIDS Research and Human Retroviruses*, vol. 19, no. 2, pp. 145–149, 2003.

[24] U. Güldener, M. Münsterkötter, G. Kastenmüller et al., "CYGD: the comprehensive yeast genome database," *Nucleic Acids Research*, vol. 33, pp. D364–D368, 2005.

[25] L. Chen, W.-M. Zeng, Y.-D. Cai, K.-Y. 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.

[26] P. Gao, Q. P. Wang, L. Chen, and T. Huang, "Prediction of human genes' regulatory functions based on proteinprotein interaction network," *Protein and Peptide Letters*, vol. 19, no. 9, pp. 910–916, 2012.

[27] L. Hu, T. Huang, X. Shi, W.-C. Lu, Y.-D. 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.

[28] L.-L. Hu, T. Huang, Y.-D. Cai, and K.-C. Chou, "Prediction of body fluids where proteins are secreted into based on protein interaction network," *PLoS ONE*, vol. 6, no. 7, Article ID e22989, 2011.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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*, pp. 1–9, 2011.

[34] L. Hu, W. Cui, Z. He et al., "Cooperativity among short amyloid stretches in long amyloidogenic sequences," *PLoS ONE*, vol. 7, no. 6, Article ID e39369, 2012.

[35] 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.

[36] 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, no. 4, Article ID e33393, 2012.

[37] J. A. Granek and P. M. Magwene, "Environmental and genetic determinants of colony morphology in yeast," *PLoS Genetics*, vol. 6, no. 1, Article ID e1000823, 2010.

[38] C. Elsztein, R. M. de Lucena, and M. A. de Morais Jr., "The resistance of the yeast Saccharomyces cerevisiae to the biocide polyhexamethylene biguanide: involvement of cell wall integrity pathway and emerging role for YAP1," *BMC Molecular Biology*, vol. 12, article 38, 2011.

[39] C. D. M. Rodley, R. S. Grand, L. R. Gehlen, G. Greyling, M. B. Jones, and J. M. O'Sullivan, "Mitochondrial-nuclear DNA interactions contribute to the regulation of nuclear transcript levels as part of the inter-organelle communication system," *PLoS ONE*, vol. 7, no. 1, Article ID e30943, 2012.

[40] R. K. Blackman, K. Cheung-Ong, M. Gebbia et al., "Mitochondrial electron transport is the cellular target of the oncology drug Elesclomol," *PLoS ONE*, vol. 7, no. 1, Article ID e29798, 2012.

[41] H. Reyes-Caballero, G. C. Campanello, and D. P. Giedroc, "Metalloregulatory proteins: metal selectivity and allosteric switching," *Biophysical Chemistry*, vol. 156, no. 2-3, pp. 103–114, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 914971, 7 pages
http://dx.doi.org/10.1155/2013/914971

Research Article

A Quantitative Analysis of the Impact on Chromatin Accessibility by Histone Modifications and Binding of Transcription Factors in DNase I Hypersensitive Sites

Peng Cui,[1,2] Jing Li,[1,2] Bo Sun,[1,2] Menghuan Zhang,[1] Baofeng Lian,[1,2] Yixue Li,[1,2] and Lu Xie[2]

[1] *School of Life Science and Biotechnology, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China*
[2] *Shanghai Center for Bioinformation Technology, 1278 Ke Yuan Road, Shanghai 201203, China*

Correspondence should be addressed to Yixue Li; yxli@sibs.ac.cn and Lu Xie; xielu@scbit.org

Received 31 July 2013; Accepted 3 September 2013

Academic Editor: Tao Huang

It is known that chromatin features such as histone modifications and the binding of transcription factors exert a significant impact on the "openness" of chromatin. In this study, we present a quantitative analysis of the genome-wide relationship between chromatin features and chromatin accessibility in DNase I hypersensitive sites. We found that these features show distinct preference to localize in open chromatin. In order to elucidate the exact impact, we derived quantitative models to directly predict the "openness" of chromatin using histone modification features and transcription factor binding features, respectively. We show that these two types of features are highly predictive for chromatin accessibility in a statistical viewpoint. Moreover, our results indicate that these features are highly redundant and only a small number of features are needed to achieve a very high predictive power. Our study provides new insights into the true biological phenomena and the combinatorial effects of chromatin features to differential DNase I hypersensitivity.

1. Introduction

In eukaryotes, DNA is organized into chains of nucleosomes, each of which consists of about 146 bp of DNA wrapped around an octamer of four types of histones [1]. The packaging of chromatin into nucleosomes provides a repressive environment for many DNA-binding proteins and plays an important role in the regulation of transcription [2]. However, some domains in chromatin are depleted of nucleosomes and exhibit highly accessible structure. These nucleosome-free regions are supersensitive to the cleavage of DNase I [3] and are known as DNase I hypersensitive sites (DHSs). They are predominantly found in many active genes and cis-regulatory elements [4]. The dynamic alterations of "openness" in chromatin play important roles in many biological processes, including transcription [5], replication [2], and differentiation [6].

Traditionally, the experimental technique of choice to discover the DNase I hypersensitive sites is Southern blotting [7]. However, this low-throughput method is not able to study large chromosomal regions at a time and cannot represent the "openness" of chromatin in a quantitative manner. The significance of differential accessibility in DNase I hypersensitive sites is unknown, but it may reflect some important biological phenomena like histone modifications and protein occupation [8]. Even until now genome-wide quantitative analyses of the relationship between chromatin accessibility and chromatin features in DNase I hypersensitive sites are rare. By taking advantage of the abundant datasets of the ENCODE project [9], we analyzed genome-wide localization data of DNase I hypersensitive sites and 33 chromatin features in human embryonic stem cell (H1hesc) cell line. All datasets were generated by recently developed genome-wide high throughput experimental techniques, such as Chip-seq [10, 11] and DNase-seq [12].

It is generally accepted that histone modifications and the binding of transcription factors are two main effectors for the "openness" of chromatin. Previous studies have shown

that histone modifications and transcription factors tend to occur near or just in the DNase I hypersensitive sites [8, 13]. Recently, two studies, one in K562 cell line and the other in Drosophila embryonic cells, have demonstrated that transcription factor binding sites and the chromatin accessibility are highly correlated with each other [6, 13]. Although these studies provided important information, so far, quantitative analysis of the combinatorial effects of different chromatin features and the biological significance of differential hypersensitivity is still unclear. In this work, we built support vector regression (SVR) models to directly predict the "openness" of chromatin in DNase I hypersensitive sites using combined chromatin features. Our results indicate that both histone modification features and transcription factor binding features are predictive for chromatin accessibility with high accuracy and these chromatin features are highly redundant.

2. Materials and Methods

2.1. Datasets. All datasets are from ENCODE project, which aims to build a comprehensive list of functional elements in the human genome [9]. The 10 histone modifications (HMs) and binding sites of 23 transcription factors (TFs) were quantified using Chip-seq and downloaded from the tracks of UCSC Genome Browser at ENCODE/Broad Institute and ENCODE/Stanford/Yale/USC/Harvard. The chromatin accessibility dataset was measured using DNase-seq and downloaded from ENCODE/OpenChrom (Duke/UNC/UTA). Each dataset includes the genome-wide sequencing signals and regions of statistically enriched signal (peaks). Peaks can be viewed as locations of chromatin features and DNase I hypersensitive sites, respectively, and the values of DNase-seq signals represent chromatin accessibility. We must note that DNase I hypersensitivity could not be simply viewed as binary property (peaks versus nonpeaks) but rather continuous values (sequencing signals) representing differential chromatin accessibility. These datasets come from the common H1hesc cell line.

2.2. Mapping HM and TF Binding Peaks to the DNase I Hypersensitive Sites. We obtained genomic locations of 33 chromatin feature profiles, all together including 582489 histone modification peaks (10 HMs) and 443217 transcription factor binding peaks (23 TFs). For each profile, we mapped the peaks of feature onto the genome and examined whether it localized in open chromatin or not. The presence or absence of chromatin feature within accessible chromatin was decided by overlap or nonoverlap with DNase-seq peaks. If there was any amount of overlap within accessible chromatin (DNase-seq peaks), we counted as a presence [13]. Then, we calculated the percentage of the peaks occurring in the DNase I hypersensitive sites for each feature.

2.3. Supervised Learning Methods for Chromatin Accessibility Prediction. To investigate the quantitative relationship between chromatin accessibility and these chromatin features in DNase I hypersensitive sites, we constructed support vector regression (SVR) models for HM and TF binding features, respectively. Concretely, in every DNase I hypersensitive site, we calculated the maximum signal of DNase-Seq and the corresponding maximum signal of Chip-Seq for each chromatin feature. For the sake of figuring out whether the maximum signal exhibits largest prediction power or not, as a comparison, we also calculated the average signal of Chip-seq and DNase-seq for each hypersensitive region. Then, SVR model was built to predict the chromatin accessibility using signals of these chromatin features. SVR is a machine learning algorithm based on statistical theory for regression problems [14, 15]. We implemented this algorithm using the "e1071" *R* package [16].

In order to reduce the computation cost, we randomly selected 5000 DHSs for our samples. The sample size analysis indicated that the prediction power increased only moderately after the size reached 2000. So, the sample size of 5000 is big enough to represent the entire dataset (Supporting Information S1 which is available online at http://dx.doi.org/10.1155/2013/914971). We used the 10-fold cross-validation method to evaluate the prediction power. Specifically, we randomly split our sample dataset into 10 equal size subsets. Among them, 9 subsets were used as training data and the remaining subset was treated as the validation data for testing the model. This process was repeated 10 times and each subset could only be used once as the validation data. After that, we combined the results and plotted the regression relationship between predicted signals and the actual DNase-seq signals. Then, the coefficient of determination (R^2) [17] was computed indicating how well these data points fit the line. R^2 is also a frequently used measure of the proportion of total variation of outcomes explained by the model. We chose the square root of the coefficient of determination (R) as our prediction power.

2.4. Analysis of the Importance for Each Chromatin Feature and the Combinatorial Effects of Different Features. To estimate which feature exhibits the maximal prediction power, we predicted the chromatin accessibility using only one feature. And to investigate whether HM features and TF binding features are redundant, we next predicted the "openness" of chromatin using all features. We also explored the combinatorial effects of these features. All possible one-feature (C_{33}^1), two-feature (C_{33}^2), and three-feature (C_{33}^3) models were evaluated by their performance.

2.5. Model Comparison Analysis. Instead of SVR algorithm, we also explored the quantitative relationship between chromatin features and chromatin accessibility with linear regression model. Similarly, HM features, TF binding features, and HM+TF feature combinations were applied to linear regression model, respectively. The coefficient of determination of the predicted signals and the actual DNase-seq signals were calculated and compared with the SVR models. In order to identify whether the maximum signals or the average signals exhibit largest prediction power, we also applied these models with the average signals of chromatin features to predict the average signals of DNase-seq.

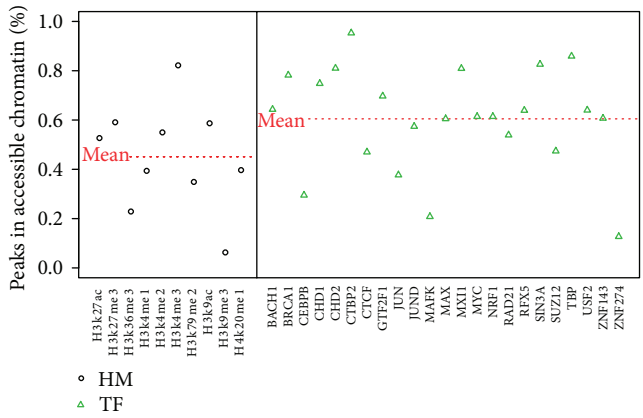


Figure 1: The percentages of histone modification (HM) features and transcription factor (TF) binding features within accessible chromatin regions. The black circle and green triangle represent HM features and TF binding features, respectively. The two red lines represent the mean percentages for HMs and TFs, respectively.

3. Results

3.1. The Localization Preference of Chromatin Features. We analyzed genome-wide localizations of 33 Chip-seq profiles in the human embryonic stem cell line (H1hesc) from ENCODE project [9], including 10 histone modifications, and the binding sites of 23 transcription factors. For each profile, we mapped the peaks of Chip-seq dataset to the DNase I hypersensitive sites (see Section 2). Figure 1 shows the percentage of the peaks within the accessible chromatin for each feature. We observed that different chromatin feature exhibits different preference to chromatin accessibility. For histone modifications, H3k4me3 exerts the largest preference of accessible chromatin. 82.2% H3k4me3 peaks located in DHS. On the contrary, most H3k9me3 occurred out of DHS (93.7%), which indicated that H3k9me3 was associated with heterochromatin [18]. Compared to histone modifications, a majority of transcription factors tend to bind onto accessible chromatin, which suggests that the process of transcription requires an open chromatin structure [19]. The mean percentage of transcription factors locating in DHS is 60.5%, higher than that of histone modifications (45.1%).

3.2. Predicting Chromatin Accessibility Using Histone Modification Features. In order to examine the quantitative relationship between chromatin accessibility and HM features in a combinatorial manner, we constructed SVR model to predict the "openness" of chromatin in DNase I hypersensitive sites using all histone modification features. We can see from Figure 2(a) that there is a linear relationship between predicted signals and the actual DNase-Seq signals. The coefficient of determination (R^2) is 0.58 indicating that histone modification features explain about 58% variance of chromatin accessibility.

We next examined the prediction power for every histone modification feature. Figure 2(b) shows that H3k4me2, H3k4me3, and H3k9ac exhibit the most important effects to chromatin accessibility ($R = 0.67, 0.66, 0.63$, resp.). These histone modifications are generally enriched in the promoters of expressed genes [20] and the open chromatin structure plays an important role in regulating the complex transcription process. On the other hand, H3k9me3 and H3k36me3 exhibit the least prediction powers ($R = 0.30$, 0.23, resp.), which suggests that these modifications are associated with heterochromatin [21, 22]. Interestingly, H3k27ac and H4k20me1, which are the most predictive histone modifications for gene expression levels [23], are not the most important features associated with chromatin accessibility.

3.3. Predicting Chromatin Accessibility Using Transcription Factor Binding Features. Previous studies have shown that transcription factors tend to bind onto open chromatin and they are highly correlated with each other [6, 13]. To investigate the quantitative relationship of the binding of transcription factors and the chromatin accessibility in a combinatorial manner, we next applied our SVR model to all TF binding features. As shown in Figure 3(a), the TF model achieves a coefficient of determination (R^2) of 0.58 which is equal to that achieved by HM model. These TF binding features can also explain about 58% variance of chromatin accessibility.

For the prediction power of particular TF binding feature, there is a difference with that of histone modifications; that is, most transcription factors exhibit important effects to chromatin accessibility (Figure 3(b)). This is consistent with their functions because transcription factors directly control the complex transcription process [24] which requires an open chromatin environment. However, a small group of features exhibit lower prediction powers, such as SUZ12, CTCF, and ZNF274 ($R = 0.37, 0.36, 0.33$, resp.). ZNF274 and SUZ12 are known to be transcriptional repressors [25, 26]. CTCF has many roles, such as transcriptional repression, insulator function, and imprinting genetic information [27]. These factors are not so important to contribute to the "openness" of chromatin.

3.4. Chromatin Features Are Highly Redundant to Chromatin Accessibility. The previous analyses suggest that both histone modification features and transcription factor binding features are predictive for chromatin accessibility with high accuracy in DNase I hypersensitive sites. So, there is a question that whether the prediction power will increase if we use all these features. To address this question, we directly predicted the "openness" of chromatin using all features. As shown in Figure 4(a), the coefficient of determination ($R^2 = 0.66$) is a little higher (8%) than using only HM or TF binding features, which indicates that these two types of features are highly redundant. To check the importance of different features and their combinatorial effects, we tried to build models with all possible combinations of one to three features (Figure 4(b)). Focusing on the three-feature combinations (5456 models), we found that the least prediction power combinations (H3k36me3, H3k9me3, and ZNF274, $R = 0.45$) could achieve about 56% prediction power of the full model ($R = 0.81$). And there are 110 combinations achieving more

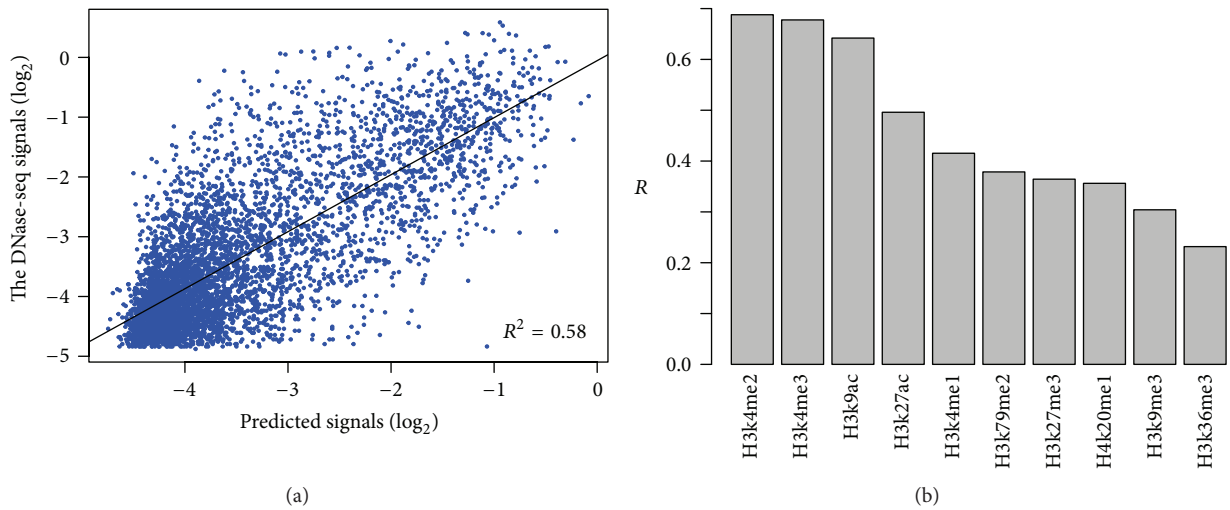


Figure 2: Prediction accuracy of chromatin accessibility using HM features. (a) Scatter plot of predicted versus experimentally measured DNase-seq signals using all HM features. The black line represents the linear fit between predicted and measured signals (R^2, coefficient of determination). (b) Prediction powers (R, the square root of coefficient of determination) of the SVR models using only one particular HM feature.

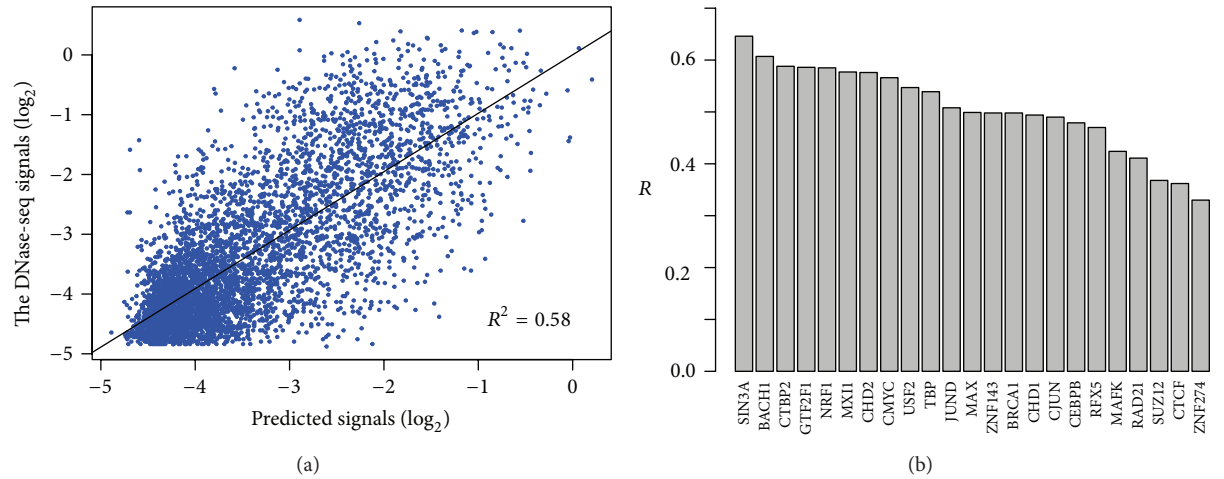


Figure 3: Prediction accuracy of chromatin accessibility using TF binding features. (a) Scatter plot of predicted versus experimentally measured DNase-seq signals using all TF binding features. The black line represents the linear fit between predicted and measured signals (R^2, coefficient of determination). (b) Prediction powers (R, the square root of coefficient of determination) of the SVR models using only one particular TF binding feature.

than 90% prediction power of the full one. These analyses indicate that most of these features are highly redundant for chromatin accessibility.

By examining the 110 high prediction power combinations, we found that seven chromatin features, H3k4me2, H3k4me3, H3k9ac, H4k20me1, SIN3A, ZNF143, SUZ12, were significantly enriched ($P < 0.01$, hypergeometric test) in the set of 110 models. Interestingly, all these features showed high prediction powers in the one-feature models except H4k20me1 and SUZ12. H4k20me1 is a particular one, which has been reported for the most predictive histone modification for gene expression [23]. SUZ12 is a part of Polycomb Repressive Complex 2 (PRC2) and may be involved in chromatin silencing with noncoding RNA [25]. The mechanisms of how SUZ12 influences chromatin structure are unknown; however, it may exert distinct impact on chromatin accessibility compared with other features.

3.5. Comparison with Other Models. In this study, we chose the SVR algorithm and the maximum signal in every hypersensitive region to model the relationship between chromatin features and chromatin accessibility. Generally, the SVR algorithm is a nonlinear regression method. We also have explored modeling using linear regression model and the average signal in every region. As shown in Table 1, prediction powers of models using average signal are significantly lower than the corresponding maximum signal models. And in either situation, the SVR models exhibit higher prediction power than linear models. Our results indicate that the "openness" of chromatin is determined by the maximum

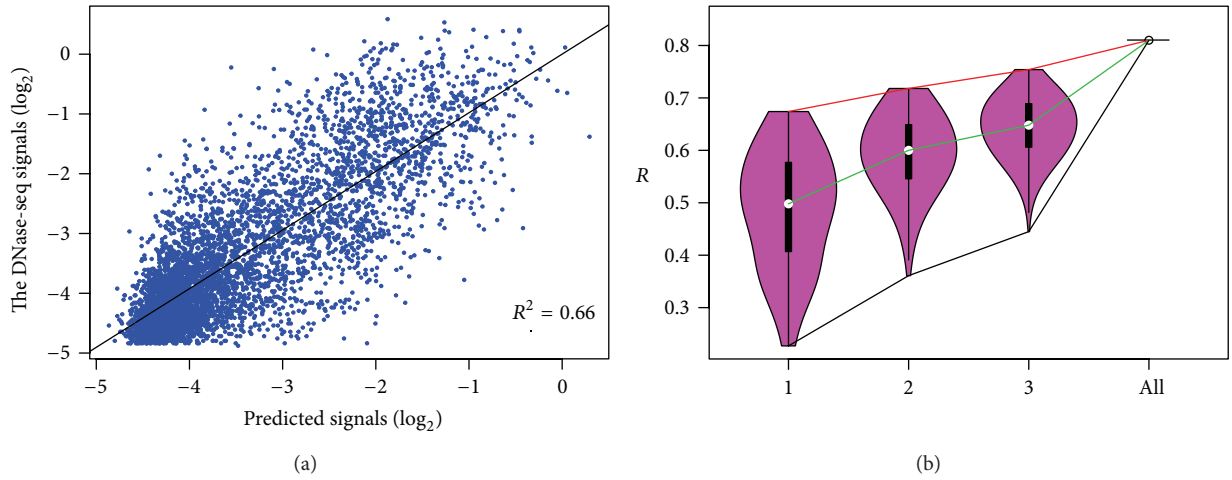


Figure 4: Redundancy of HM features and TF binding features. (a) Scatter plot of predicted versus experimentally measured DNase-seq signals using all HM and TF binding features. The black line represents the linear fit between predicted and measured signals (R^2, coefficient of determination). (b) Comparison of prediction powers (R, the square root of coefficient of determination) between all possible one-feature, two-feature, three-feature models, and the full model in H1hesc.

Table 1: Comparison of prediction powers with different models. The prediction power is represented as the square root of the coefficient of determination (R) for predicted and the actual DNase-seq signals. LM: linear regression model.

Model	SVR (max_signal)	LM (max_signal)	SVR (avg_signal)	LM (avg_signal)
HM	0.76	0.69	0.70	0.56
TF	0.76	0.63	0.69	0.53
HM + TF	0.81	0.73	0.75	0.61

signal of features and their relationships are assumed as a nonlinear relevance.

4. Discussion

In this work, we presented quantitative analyses of the relationship of histone modifications and the binding of transcription factors to chromatin accessibility separately and combinedly in DNase I hypersensitive sites. We first examined the percentage of feature peaks within DNase hypersensitive sites (DHSs) in human embryonic stem cell (H1hesc) line. We found that different chromatin features showed different location preference in DHS. This may be due to the particular function of different chromatin features. Thurman et al. have done similar analysis in K562 cell line [13] for TF binding features. In our analysis, we find that the percentage of transcription factors within DHS is significantly lower in the H1hesc cell line than that in K562 cell line. The reason may be as follows: in order to maintain the "stemness" state, most genes are repressed in the stem cell compared to the cancer cell line K562. This phenomenon means that the degree to which chromatin features occur in accessible chromatin may differ according to different cellular circumstances.

Our results demonstrate that both histone modification (HM) features and transcription factor (TF) binding features account for nearly 58% variance of chromatin accessibility in H1hesc cell line. For histone hallmarks, many activators of gene expression exhibited important impact on the "openness" of chromatin, such as H3k4me [21] and histone acetylations [28]. The hallmarks of repressors for gene expression such as H3k9me3 [21] show lower prediction powers. Unexpectedly, the transcription elongation hallmark H3k36me3 [29] shows the least prediction power. This is consistent with the viewpoint of a recently published paper. Chantalat et al. [22] argued that H3k36me3 is associated with constitutive and facultative heterochromatin. For TF binding features, the majority of TFs showed an important impact on chromatin accessibility except some transcriptional repressors, such as ZNF274 and SUZ12. This may indicate that the complex transcription process requires open chromatin environment [19].

It is generally accepted that cellular factors regulate the complex dynamic change of chromatin structure in a collective manner. We have shown that these features is highly redundant to predict chromatin accessibility and a small subgroup of features are able to achieve a very high prediction power. However, the mechanism of how these features cooperatively impact the openness of chromatin is still unclear, and we must note that our analysis could not reveal the "cause" or "consequence" relationship of HM and TF binding features to chromatin accessibility. Histone modifications play an important role in creating and maintaining the accessible chromatin environment [30] and may act as docking sites for transcription factors [31]. Some pioneer TFs tend to bind onto the genome and create an accessible site, such as FoxA1 [32] which is the best known pioneer transcription factor. Then, more transcription factors tend to bind onto the opening site and the DNase I hypersensitive site is created. As an extension, future work could explore the mechanisms of how these features cooperatively regulate open chromatin structure and their causal relationships, based on increased datasets.

5. Conclusion

We present genome-wide quantitative analysis of the impact of chromatin features to chromatin accessibility in DNase I hypersensitive sites. Our findings indicate that both histone modifications and the binding of transcription factors could explain nearly 58% variation of the "openness" of chromatin structure. The combinatorial effect analyses reveal that these chromatin features are highly redundant for prediction and H3k4me2, H3k4me3, H3k9ac, SIN3A, and ZNF143 show closest association with chromatin accessibility. Our results provide insights into the systematic effects of chromatin features to differential chromatin accessibility.

Abbreviations

DHS: DNase I hypersensitive site
HM: Histone modification
SVR: Support vector regression.

Acknowledgments

This work was funded by National Key Basic Research Program (2010CB912702, 2011CB910204), National Hi-Tech Program (2012AA020201), Key Infectious Disease Project (2012ZX10002012-014), and National Natural Science Foundation of China (31070752, 31000582).

References

[1] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8 Å resolution," *Nature*, vol. 389, no. 6648, pp. 251–260, 1997.

[2] J. D. Anderson and J. Widom, "Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites," *Journal of Molecular Biology*, vol. 296, no. 4, pp. 979–987, 2000.

[3] C. Dingwall, G. P. Lomonossoff, and R. A. Laskey, "High sequence specificity of micrococcal nuclease," *Nucleic Acids Research*, vol. 9, no. 12, pp. 2659–2674, 1981.

[4] D. S. Gross and W. T. Garrard, "Nuclease hypersensitive sites in chromatin," *Annual Review of Biochemistry*, vol. 57, pp. 159–197, 1988.

[5] P. N. Cockerill, "Structure and function of active chromatin and DNase I hypersensitive sites," *FEBS Journal*, vol. 278, no. 13, pp. 2182–2210, 2011.

[6] X.-Y. Li, S. Thomas, P. J. Sabo, M. B. Eisen, J. A. Stamatoyannopoulos, and M. D. Biggin, "The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding," *Genome Biology*, vol. 12, article R34, 2011.

[7] Q. Lu and R. Bruce, "DNaseI hypersensitivity analysis of chromatin structure," in *Epigenetics Protocols*, pp. 77–86, Humana Press, 2004.

[8] A. P. Boyle, S. Davis, H. P. Shulha et al., "High-resolution mapping and characterization of open chromatin across the genome," *Cell*, vol. 132, no. 2, pp. 311–322, 2008.

[9] I. Dunham, A. Kundaje, S. F. Aldred et al., "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, pp. 57–74, 2012.

[10] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology," *Nature Reviews Genetics*, vol. 10, no. 10, pp. 669–680, 2009.

[11] E. R. Mardis, "ChIP-seq: welcome to the new frontier," *Nature Methods*, vol. 4, no. 8, pp. 613–614, 2007.

[12] L. Song and G. E. Crawford, "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells," *Cold Spring Harbor Protocols*, 2010.

[13] R. E. Thurman, E. Rynes, R. Humbert et al., "The accessible chromatin landscape of the human genome," *Nature*, vol. 489, pp. 75–82, 2012.

[14] C. Nello and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.

[15] C. Cheng, K.-K. Yan, K. Y. Yip et al., "A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets," *Genome Biology*, vol. 12, no. 2, article R15, 2011.

[16] D. Evgenia, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel, "Misc functions of the Department of Statistics (e1071), TU Wien," R package pp. 1–5, 2008.

[17] B. S. Everitt, *Cambridge Dictionary of Statistics*, Cambridge University Press, New York, NY, USA, 2nd edition, 2002.

[18] J. Bartkova, P. Moudry, Z. Hodny, J. Lukas, R. D. Meyts, and J. Bartek, "Heterochromatin marks HP1γ, HP1α and H3K9me3, and DNA damage response activation in human testis development and germ cell tumours," *International Journal of Andrology*, vol. 34, no. 4, pp. e103–e113, 2011.

[19] D. Sproul, N. Gilbert, and W. A. Bickmore, "The role of chromatin structure in regulating the expression of clustered genes," *Nature Reviews Genetics*, vol. 6, no. 10, pp. 775–781, 2005.

[20] K. Regha, M. A. Sloane, R. Huang et al., "Active and repressive chromatin are interspersed without spreading in an imprinted gene cluster in the mammalian genome," *Molecular Cell*, vol. 27, no. 3, pp. 353–366, 2007.

[21] R. Margueron and D. Reinberg, "Chromatin structure and the inheritance of epigenetic information," *Nature Reviews Genetics*, vol. 11, no. 4, pp. 285–296, 2010.

[22] S. Chantalat, A. Depaux, P. Héry et al., "Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin," *Genome Research*, vol. 21, no. 9, pp. 1426–1437, 2011.

[23] R. Karlić, H.-R. Chung, J. Lasserre, K. Vlahoviček, and M. Vingron, "Histone modification levels are predictive for gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 7, pp. 2926–2931, 2010.

[24] G. Gill, "Regulation of the initiation of eukaryotic transcription," *Essays in Biochemistry*, vol. 37, pp. 33–43, 2001.

[25] J. L. Rinn, M. Kertesz, J. K. Wang et al., "Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs," *Cell*, vol. 129, no. 7, pp. 1311–1323, 2007.

[26] K. Yano, N. Ueki, T. Oda, N. Seki, Y. Masuho, and M.-A. Muramatsu, "Identification and characterization of human ZNF274 cDNA, which encodes a novel Kruppel-type zinc-finger protein having nucleolar targeting ability," *Genomics*, vol. 65, no. 1, pp. 75–80, 2000.

[27] K. L. Dunn and J. R. Davie, "The many roles of the transcriptional regulator CTCF," *Biochemistry and Cell Biology*, vol. 81, no. 3, pp. 161–167, 2003.

[28] B. M. Turner, "Histone acetylation and an epigenetic code," *BioEssays*, vol. 22, no. 9, pp. 836–845, 2000.

[29] S. M. Fuchs, R. N. Laribee, and B. D. Strahl, "Protein modifications in transcription elongation," *Biochimica et Biophysica Acta*, vol. 1789, no. 1, pp. 26–36, 2009.

[30] J. Marx, "Molecular biology. Protein tail modification opens way for gene activity," *Science*, vol. 311, no. 5762, p. 757, 2006.

[31] O. Bell, V. K. Tiwari, N. H. Thomä, and D. Schübeler, "Determinants and dynamics of genome accessibility," *Nature Reviews*, vol. 12, pp. 554–564, 2011.

[32] L. A. Cirillo, F. R. Lin, I. Cuesta, D. Friedman, M. Jarnik, and K. S. Zaret, "Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4," *Molecular Cell*, vol. 9, no. 2, pp. 279–289, 2002.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 304029, 8 pages
http://dx.doi.org/10.1155/2013/304029

Research Article

Prediction and Analysis of Retinoblastoma Related Genes through Gene Ontology and KEGG

Zhen Li,[1] Bi-Qing Li,[2] Min Jiang,[3] Lei Chen,[4] Jian Zhang,[5] Lin Liu,[1] and Tao Huang[6]

[1] *Department of Ophthalmology, Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China*
[2] *Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, 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 200031, 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 Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York City, NY 10029, USA*

Correspondence should be addressed to Lin Liu; linliurjsh@gmail.com and Tao Huang; tohuangtao@126.com

Received 17 June 2013; Accepted 16 July 2013

Academic Editor: Yudong Cai

One of the most important and challenging problems in biomedicine is how to predict the cancer related genes. Retinoblastoma (RB) is the most common primary intraocular malignancy usually occurring in childhood. Early detection of RB could reduce the morbidity and promote the probability of disease-free survival. Therefore, it is of great importance to identify RB genes. In this study, we developed a computational method to predict RB related genes based on Dagging, with the maximum relevance minimum redundancy (mRMR) method followed by incremental feature selection (IFS). 119 RB genes were compiled from two previous RB related studies, while 5,500 non-RB genes were randomly selected from Ensemble genes. Ten datasets were constructed based on all these RB and non-RB genes. Each gene was encoded with a 13,126-dimensional vector including 12,887 Gene Ontology enrichment scores and 239 KEGG enrichment scores. Finally, an optimal feature set including 1061 GO terms and 8 KEGG pathways was obtained. Analysis showed that these features were closely related to RB. It is anticipated that the method can be applied to predict the other cancer related genes as well.

1. Introduction

Retinoblastoma (Rb) is a rapidly developing cancer in infants that develops in the cells of retina, the light-detecting tissue of the eye [1], which can be heritable or nonheritable. The most common and obvious sign of retinoblastoma is an abnormal appearance of the pupil, leukocoria, also known as amaurotic cat's eye reflex [2]. Retinoblastoma is rare and affects approximately 1 in 15,000 live births, but it is the most common inherited childhood malignancy. In China, around 1100 new cases are diagnosed each year, just second to that of India. Patients without diagnosis and being treated untimely would undergo enucleation or even die. In about two-thirds of cases, only one eye is affected (unilateral retinoblastoma); in the other third, tumours develop in both eyes (bilateral retinoblastoma). The number and size of tumours on each eye may vary [2].

As a kind of neural ectoderm tumor, heritable Rb is mainly caused by the mutation of Rb gene and dysfunction of tumor suppressor genes [3]. In these years, a rise in the number of cases was found, which was partly blue to the environmental pollution. The defective RB1 gene can be inherited from either parents; in some children, however, the mutation occurs in the early stages of fetal development [4]. Somatic amplification of the MYCN oncogene is responsible for some cases of non-hereditary, early onset, aggressive, and unilateral Rb. Although MYCN amplification accounted for only 1.4% of Rb cases, researchers identified it in 18% of infants diagnosed at less than 6 months of age. Median age at diagnosis for MYCN Rb was 4.5 months, compared with 24 months for those who had nonfamilial unilateral disease with two RB1 gene mutations [5]. Bilaterally affected individuals and 13%–15% of unilaterally affected individuals are expected

to show an RB1 mutation in blood [6, 7]; the rest 85% of unilaterally affected patients were found not to carry either of their eye tumor RB1 mutations in blood; neither molecular testing nor clinical surveillance of siblings is required [8]. So to find more molecular markers or more effective prediction method is crucial for Rb diagnosis.

System biology approaches for discovering cancer related genes have been reported [9–11]. The Gene Ontology (GO) is a major bioinformatics tool to unify the representation of gene and gene product attributes across all species [12]. GO terms have been used previously to characterize protein function and to elucidate trends in protein datasets [13]. In addition, it has been shown that GO annotations are good predictors of cancer genes [14]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database is a widely used comprehensive inference for pathway mapping of genes.

Here, we developed a new systems biological measure to effectively and deficiently identify RB genes and their pathways. First, we identified 119 RB genes from the overlap of two gene expression studies of retinoblastoma. In order to identify GO terms and KEGG pathways that are distinct between RB and non-RB genes, 5,500 non-RB genes were randomly selected from the Ensembl genes. Then all the genes were encoded with 12,887 Gene Ontology enrichment scores and 239 KEGG enrichment scores. mRMR and IFS was used to rank these features. Dagging was employed as the prediction engine. Finally, 1061 GO terms and 8 KEGG pathways were obtained as the optimal features to discriminate an RB and non-RB gene, which has been shown to be closely related to RB.

2. Materials and Methods

2.1. Dataset. The 119 consistently differentially expressed genes between retinoblastoma and normal retina were obtained from the overlap between differentially expressed genes discussed in two gene expression studies of retinoblastoma [15, 16] (see Supplementary Material available at http://dx.doi.org/10.1155/2013/304029). In Chakraborty et al.'s study [15], there was a total of 10 RB samples and three normal retina samples. Human 19K cDNA microarray which was interrogating 19,000 human genes was used to get the expression profiling and the raw data were normalized by grid-wise normalization. In Ganguly and Shields's study [16], they investigated the gene expression of six matched RB tissues and normal retinal tissues with GeneChip Human U133 V2.0 microarray. Both the Affymetrix standard protocols and the standard model-based methods of robust multichip average were used. The values are background adjusted, normalized, and log transformed. There were 110 proteins corresponding to these 119 RB genes, which were regarded as positive samples in this study. The gene symbols were mapped to Ensembl proteins ID with the tool BioMart [17]. We randomly selected $110 \times 50 = 5{,}500$ non-RB genes from Ensembl as the negative samples. We refer the reader to [18] to deal with imbalanced data; all the negative samples were randomly split into 10 parts to comprise 10 datasets with the 110 positive samples. All the RB related genes and non-RB related genes are given in Supplementary S1.

2.2. Gene Ontology and KEGG Enrichment Scores. The Gene ontology enrichment score of a protein is defined as the −log10 of the hypergeometric test P value [19–21] of its direct neighbors in STRING network [22]. The higher the enrichment score of a certain Gene Ontology term, the more overrepresented it is. There were 12,887 Gene Ontology enrichment score features. In the same way, the KEGG enrichment score of a protein is defined as the −log10 of the hypergeometric test P value [19, 20] of its direct neighbors in STRING network [22]. The higher the enrichment score of one pathway, the more overrepresented the pathway is. There were 239 KEGG enrichment score features.

2.3. Feature Reduction. We calculated the Cramer's V coefficient [23, 24] between features and target variables. Cramer's V coefficient is a statistical measurement derived from the Pearson Chi-square test [25]. It ranges from 0 to 1. The smaller Cramer's V coefficient indicates weaker association. The features with Cramer's V coefficient small than 0.1 were removed.

2.4. mRMR Method and Dagging. We used the minimum redundancy maximal relevance (mRMR) method to rank the importance of the features [26]. The mRMR method ranks features based on both their relevance to the target and the redundancy between features. A smaller index of a feature denotes that it has a better tradeoff between maximum relevance to the target and minimum redundancy. For detail, please refer to our previous works [21, 27–31].

Dagging is a metaclassifier that employs majority vote to combine multiple models derived from a single learning algorithm using disjoint samples [32]. For a training dataset $\mathfrak{T} = \{s_1, s_2, \ldots, s_n\}$, k disjoint subsets of size n' are constructed by randomly taking samples in $\mathfrak{T}$ without replacement, where $kn' \leq n$. Use a basic classifier to derive k classification models $M_1, M_2, \ldots, M_k$ from the constructed k disjoint subsets of $\mathfrak{T}$. For a query sample, each of these models provides an output. The final predicted result is the class with most votes. In Weka 3.6.4 [33], the classifier "Dagging" implements the dagging classifier described above. In this study, it was employed as the classification model. For convenience, it was run with its default parameters. In detail, SMO is used as a basic classifier, and k is set to 10. In recent years, Dagging has been employed to deal with some biological problems [34–37]. Its performances in these studies show that it can be superior to some classic classifiers in some cases.

2.5. Ten-Fold Cross Validation and Incremental Feature Selection (IFS). Ten-fold cross validation was often used to evaluate the performance of a classifier [38]. To evaluate the performance of the predictor, the prediction accuracy, specificity, sensitivity, and MCC (Matthews's correlation coefficient) were calculated as follows:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})\,(\text{TP} + \text{FN})\,(\text{TN} + \text{FP})\,(\text{TN} + \text{FN})}}, \quad (1)$$

where TP denotes true positive. TN denotes true negative. FP denotes false positive and FN denotes false negative.

Based on the features ranked by mRMR, we used incremental feature selection (IFS) [21, 28, 39, 40] to determine the optimal number of features. During IFS procedure, features in the ranked feature set are added one by one from higher to lower rank. A new feature set is composed when one feature is added. For each of the feature sets, a Dagging classifier is constructed and tested using ten-fold cross-validation test. Thus, an IFS table is obtained with one column being the index of the feature set and the other columns being the prediction accuracies, sensitivities, specificities, and MCCs. We then can get the optimal feature set, using the predictor that achieves the best prediction performance.

3. Results and Discussion

3.1. The mRMR Result. After running the mRMR software, we obtained two tables for each of the ten datasets (see Supplementary S2): one is called MaxRel feature table that ranks the features according to their relevance to the class of samples and the other is called mRMR feature table that lists the ranked features by the maximum relevance and minimum redundancy to the class of samples. In the mRMR feature table, a feature with a smaller index implies that it is more important for discriminating RB and non-RB genes. Such list of ranked feature was to be used in the following IFS procedure for the optimal feature set selection.

3.2. IFS Result. By adding the ranked features one by one, we built 500 individual predictors based on 500 subfeature sets to predict RB genes for each of the ten datasets. We then tested the prediction performance for each of the 500 predictors and obtained the IFS results (see Supplementary S3). The IFS curves plotted based on the data of Supplementary S3 are shown in Figure 1. The IFS curve of dataset 1 is shown in Figure 1, we can see that the maximal MCC was 0.5174 when 156 features as given in Supplementary S3 were used. Such 156 features were regarded as the optimal feature set for dataset 1. Based on these 156 features, the prediction sensitivity, specificity, and accuracy were 0.5727, 0.9291, and 0.8697, respectively (Table 1). For the other nine datasets, the IFS results can be found in Supplementary S3 and corresponding IFS curves can be found in Supplementary S4. Finally, we took the union of optimal features for all the ten datasets as the final optimal feature set, which included 1061 GO terms and 8 KEGG pathways (see Supplementary S5). Hereafter, the further analysis was based on this final optimal feature set.

3.3. 119 RB Genes Enrichment Analysis. To compare the enrichment result of only positive sample and the selected GO and KEGG terms, we conducted the enrichment analysis for the 119 RB genes. The results showed that 12 GO terms were enriched significantly (Benjamini adjusted P value < 0.05; see Supplementary S6). Among them, two GO terms (GO:0007049: cell cycle and GO:0000087: M phase of mitotic cell cycle) were in our optimal feature set. For KEGG pathways, only hsa04110 (cell cycle) was significantly enriched (see Supplementary S6) and it has been included in our optimal feature set for distinguishing RB genes and non-RB genes, which suggested that these three enriched terms including GO:0007049: cell cycle, GO:0000087: M phase of mitotic cell cycle, and hsa04110: cell cycle are critical discriminators for RB genes and non-RB genes.

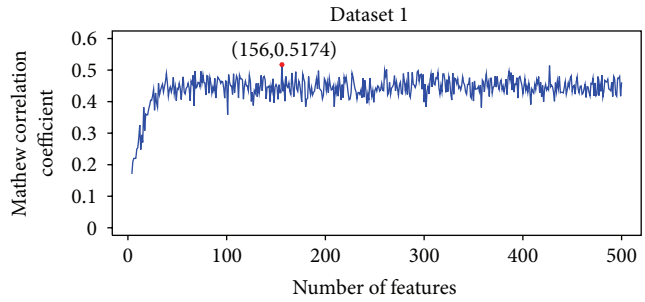


Figure 1: IFS curve for the first datasets. The maximal MCC was 0.5174 when 156 features were used.

3.4. Analysis of the Optimal Feature Set

3.4.1. GO Number and Percentage. To illustrate the biological meanings of the selected optimal feature subset, we firstly tried to classify GO terms in the optimal set into the three kinds: the biological process, cellular component, and molecular function GO terms. And the GO terms of the feature obtained by mRMR method were mapped to the children of the three root GO terms. The figures show the frequency of each GO term in the feature subset and display the ratio of the number of each GO term to the scale of the number of its children terms.

(1) Biological Process GO Terms. From Figure 2, it can be seen that in the frequency of BP terms, the top five GO biological process terms are GO:0009987: cellular process (662), GO:0008152: metabolic process (425), GO:0065007: biological regulation (386), GO:0050789: regulation of biological process (364), and GO:0019740: nitrogen utilization (210). The inclusion of cellular process (GO:0009987), biological regulation (GO:0065007), and regulation of biological process (GO:0050789) within the top five frequencies of GO terms may suggest that these biological functions performed by certain proteins at the cellular level are very important in normal persons and may be dysfunctional in Rb patients.

For the percentage of BP terms, the top five GO biological processes are GO:0006794: phosphorus utilization (4.99%), GO:0022610: biological adhesion (4.85%), GO:0008283: cell proliferation (4.81%), GO:0071840: cellular component organization or biogenesis (4.26%), and GO:0019740: nitrogen utilization (4.08%). Phosphorus utilization provides cells

Table 1: The predicted results for ten datasets.

Dataset	Optimal feature number	Sn	Sp	Acc	MCC
1	156	0.5727	0.9291	0.8697	0.5174
2	141	0.6273	0.9218	0.8727	0.5452
3	337	0.7364	0.8691	0.8470	0.5347
4	140	0.6000	0.9327	0.8773	0.5471
5	126	0.5636	0.9436	0.8803	0.5434
6	489	0.6273	0.9255	0.8758	0.5527
7	78	0.5545	0.9527	0.8864	0.5588
8	222	0.6364	0.9345	0.8848	0.5795
9	319	0.6545	0.9218	0.8773	0.5663
10	235	0.5545	0.9491	0.8833	0.5495
Mean (standard deviation)		0.6127 (0.0567)	0.928 (0.0234)	0.8755 (0.0113)	0.5494 (0.017)

Sn: sensitivity; Sp: specificity; Acc: accuracy; MCC: Matthews's correlation coefficient.

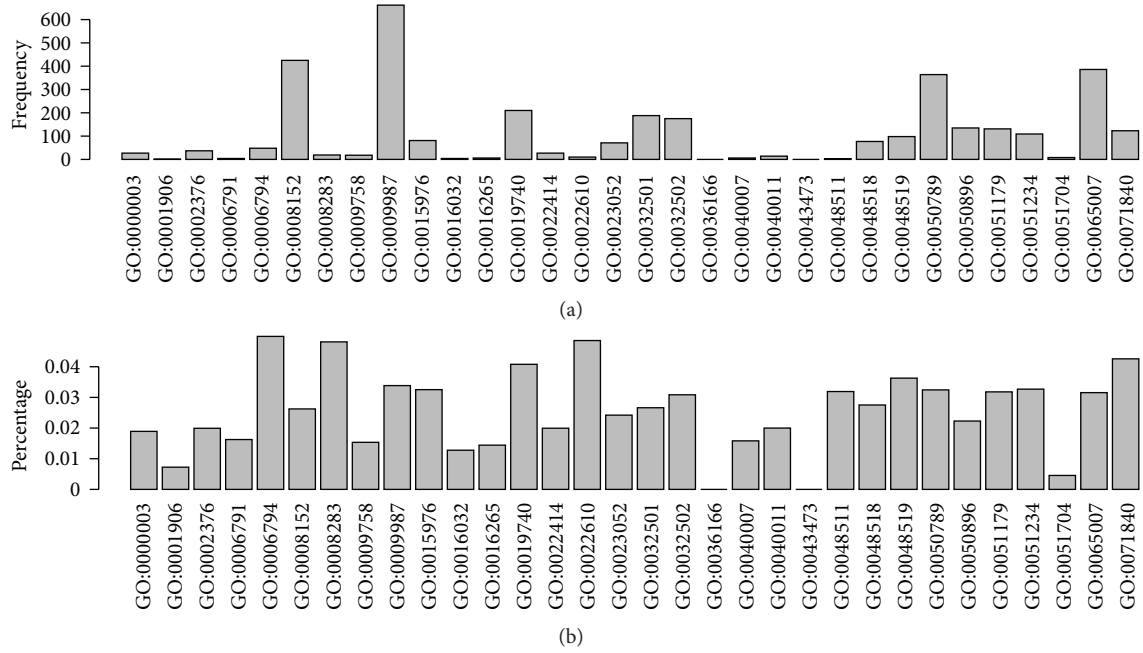


Figure 2: Illustrating the distribution of GO terms of biological process in the optimal feature set. (a) The frequency of GO terms of biological process. (b) The percentage of GO terms of biological process.

phosphorylation sources and ensures regular cellular activities. From the GO biological process term percentage distribution, it can be seen that GO terms related with cell proliferation and biological adhesion are also highlighted, although their term numbers are less than those of the others. This indicates that proteins assigned with these two GO terms have relatively high influence on RB. For example, RB1 is a key regulator of cell proliferation and fate in retinoblastoma, phosphorylation of which can lead to conformational alterations and inactivates the capability of RB1 to bind partner proteins [41]. Cell adhesion also contributes to normal cells' exchange and communication. Epithelial cell adhesion molecule (EpCAM) can regulate expression of the oncogenic miR17-92 cluster in RB and thereby controls Rb cell proliferation and invasion [42].

(2) Cellular Component GO Terms. In Figure 3(a), for frequency of CC terms, the top five GO cellular component terms are GO:0005623: cell (158), GO:0044464: cell part (145), GO:0043226: organelle (76), GO:0044422: organelle part (63), and GO:003299: macromolecular complex (60), mostly because of their large base numbers. In the percentage of CC terms, the top five GO cellular component terms also include GO:0044420: extracellular matrix part (9.09%), GO:0031012: extracellular matrix (7.07%), GO:0031974: membrane-enclosed lumen (5.41%), GO:0044422: organelle part (5.15%), and GO:0043226: organelle (4.73%).

Extracellular matrix is associated with cell adhesion mentioned in the last section. Inadhesive cells having destroyed extracellular matrix and no natural protections tend to be

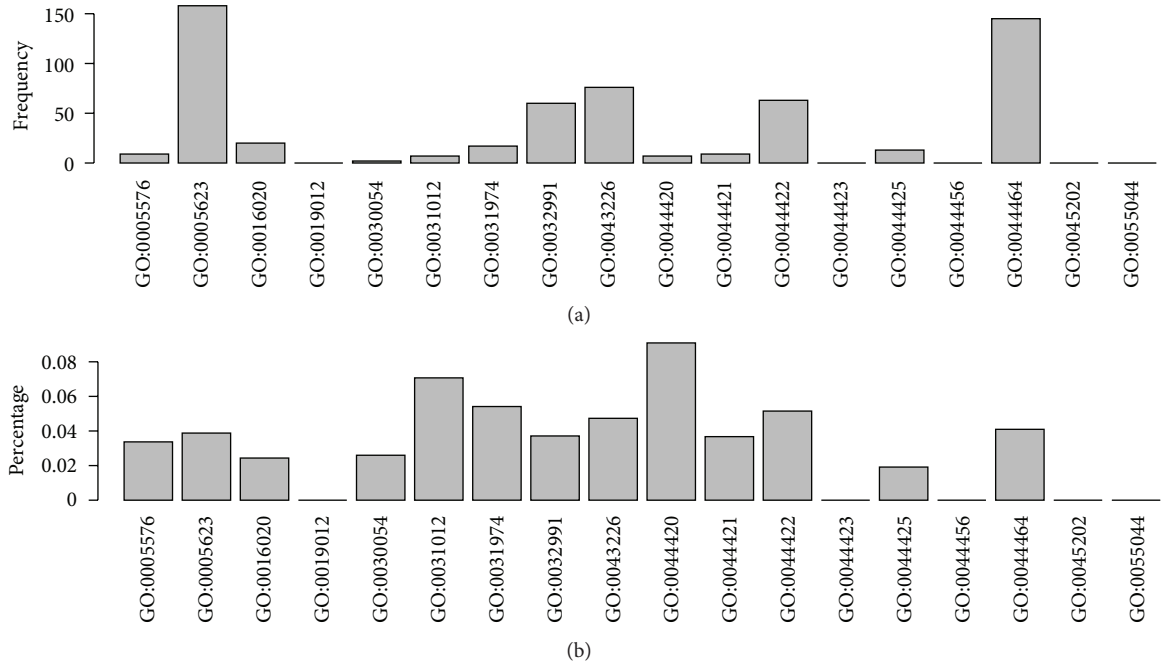


Figure 3: Illustrating the distribution of GO terms of cellular component in the optimal feature set. (a) The frequency of GO terms of cellular component. (b) The percentage of GO terms of cellular component.

tumor cells under outside pressures. Here, from the percentage distribution, it is suggested that extracellular matrix was highly related with RB. Additionally, the inclusion of membrane-enclosed lumen, organelle, and organelle part indicated that cell organelles (with or without membrane) may involve in Rb too.

(3) Molecular Function GO Terms. In Figure 4, the top five GO molecular function terms of frequency are GO:0003824: catalytic activity (152), GO:0005488: binding (85), GO:0000988: protein binding transcription factor activity (51), GO:0065009: regulation of molecular function (38), and GO:0005215: transporter activity (30). Because of large base numbers, protein GO terms related to RB are relatively more enriched in the top five molecular function GO terms, especially in catalytic activity (GO:0003824) and binding (GO:0005488). Proteins assigned to these GO terms required interaction to carry out their structural or functional activities. This suggests that dysfunction of proteins assigned to these GO terms contributed profoundly to Rb tumorogenesis. The highlight of catalytic activity (GO:0003824) may be attributed to the fact that many Rb related proteins are involved in catalytic activities such as enzymes. The highlight of binding (GO:0005488) may be ascribed to the fact that proteins expressing specific function should regulate or interact with others through binding each other. Biological progresses such as phosphorylation and acetylation are critical in disease and both of them need certain enzyme to catalyze; for example, phosphorylated p53 can intiatie cell cycle arrest of abnormal cells and acetylated ones can cause apoptosis of injured cells [43, 44], and all these processes need binding and catalysis to execute function.

In Figure 4(b), the top five GO molecular function terms are GO:0045182: translation regulator activity (14.3%), GO:0030234: enzyme regulator activity (7.78%), GO:0001071: nucleic acid binding transcription factor activity (6.31%), GO:0044093: positive regulation of molecular function (6%), and GO:0005198: structural molecule activity (5.88%). Because of the large base number of the top five GO terms in frequency, they have relatively lower enrichment than the top five GO terms in percentage. But, the top five GO terms in MF percentage are all interrelated with these in BP percentage and CC percentage. For example, ribosome as a kind of organelle serves as translation vehicle in cells, which may somehow take part in the translation regulation. RB protein phosphorylation also needs enzyme to catalyze [45].

3.4.2. The KEGG Pathways in the Optimal Set. We got eight KEGG pathway terms in the optimal set of features (see Supplementary S5), which are hsa00520 (amino sugar and nucleotide sugar metabolism) and has 00563 (glycosylphosphatidylinositol- (GPI-) anchor biosynthesis), hsa03015 (mRNA surveillance pathway), hsa03440 (homologous recombination), hsa03450 (nonhomologous end joining), hsa04110 (cell cycle), hsa04114 (oocyte meiosis), and hsa04330 (notch signaling pathway). Among them, amino sugar and nucleotide sugar metabolism (hsa00520) emphasize the sugar metabolism in eye cancer. Glycosylphosphatidylinositol- (GPI-) anchor biosynthesis (hsa00563) pathway is related with anchoring of proteins outside of membrane. The next three are all included in genetic information processing pathway. The mRNA surveillance pathway (hsa03015) involved in translation and the other two deal with replication and repair. Cell cycle (hsa04110) and oocyte meiosis

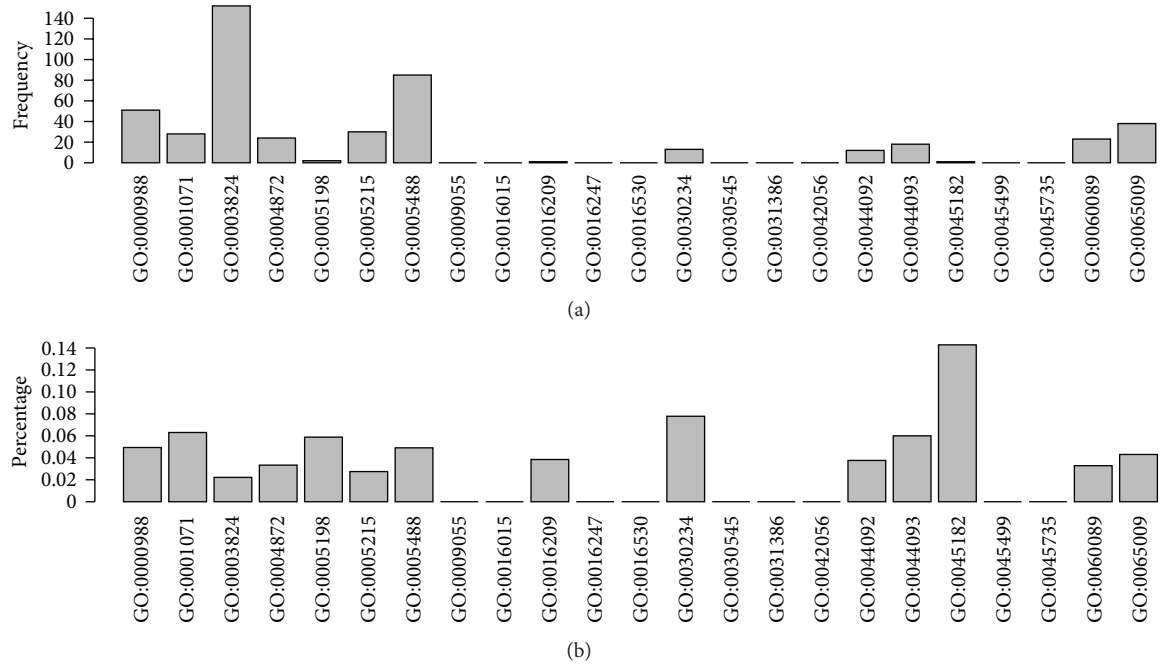


Figure 4: Illustrating the distribution of GO terms of molecular function in the optimal feature set. (a) The frequency of GO terms of molecular function. (b) The percentage of GO terms of molecular function.

(hsa04114) are related to cell growth and death, and notch signaling pathway (hsa04330) is involved in signal transduction.

The canonical pathway that links tumor suppressor gene Rb to human cancers details its interaction with the E2F transcription factors and cell-cycle progression [46]; recent studies have shown a significant role for RB-1 in the suppression of glycolytic and glutaminolytic metabolism [47, 48]. So the RB-E2F axis and the up- and down-stream genes should be very important in finding new potent antitumor target for Rb treatment.

4. Conclusion

We proposed a computational method to identify cancer related genes taking GO enrichment scores and KEGG enrichment scores as features. We applied this method to RB. An optimal feature set including 1061 GO terms and 8 KEGG pathways was revealed by our method, which has been shown to be closely related to RB. We believe this method is efficient and effective in prediction of novel cancer related genes and has universal applicability in the cancer research.

Authors' Contribution

Zhen Li and Bi-Qing Li contributed equally to this paper.

Acknowledgments

This work was supported by grants from the National Basic Research Program of China (2011CB510101 and 2011CB510102), Innovation Program of Shanghai Municipal Education Commission (12ZZ087), and the grant of "The First-class Discipline of Universities in Shanghai" and Medical Introductory Project of Science and Technology Commission of Shanghai Municipality (124119a9500).

References

[1] F. Di Nicolantonio, M. Neale, Z. Onadim, J. L. Hungerford, J. L. Kingston, and I. A. Cree, "The chemosensitivity profile of retinoblastoma," *Recent Results in Cancer Research*, vol. 161, pp. 73–80, 2003.

[2] A. MacCarthy, J. M. Birch, G. J. Draper et al., "Retinoblastoma in Great Britain 1963–2002," *British Journal of Ophthalmology*, vol. 93, no. 1, pp. 33–37, 2009.

[3] G. M. Gordon and W. Du, "Conserved RB functions in development and tumor suppression," *Protein and Cell*, vol. 2, no. 11, pp. 864–878, 2011.

[4] D. Lohmann, "Retinoblastoma," *Advances in Experimental Medicine and Biology*, vol. 685, pp. 220–227, 2010.

[5] D. W. Felsher, "Role of MYCN in retinoblastoma," *The Lancet Oncology*, vol. 14, pp. 270–271, 2013.

[6] A. Schüler, S. Weber, M. Neuhäuser et al., "Age at diagnosis of isolated unilateral retinoblastoma does not distinguish patients with and without a constitutional RB1 gene mutation but is influenced by a parent-of-origin effect," *European Journal of Cancer*, vol. 41, no. 5, pp. 735–740, 2005.

[7] D. Rushlow, B. Piovesan, K. Zhang et al., "Detection of mosaic RB1 mutations in families with retinoblastoma," *Human Mutation*, vol. 30, no. 5, pp. 842–851, 2009.

[8] S. Richter, K. Vandezande, N. Chen et al., "Sensitive and efficient detection of RB1 gene mutations enhances care for families with retinoblastoma," *American Journal of Human Genetics*, vol. 72, no. 2, pp. 253–269, 2003.

[9] L. Hood, J. R. Heath, M. E. Phelps, and B. Lin, "Systems biology and new technologies enable predictive and preventative medicine," *Science*, vol. 306, no. 5696, pp. 640–643, 2004.

[10] S. S. Knox, "From 'omics' to complex disease: a systems biology approach to gene-environment interactions in cancer," *Cancer Cell International*, vol. 10, article 11, 2010.

[11] J. J. Hornberg, F. J. Bruggeman, H. V. Westerhoff, and J. Lankelma, "Cancer: a systems biology disease," *BioSystems*, vol. 83, no. 2-3, pp. 81–90, 2006.

[12] D. Altshuler, M. J. Daly, and E. S. Lander, "Genetic mapping in human disease," *Science*, vol. 322, no. 5903, pp. 881–888, 2008.

[13] E. Camon, D. Barrell, V. Lee, E. Dimmer, and R. Apweiler, "The Gene Ontology Annotation (GOA) database—an integrated resource of GO annotations to the UniProt knowledgebase," *In Silico Biology*, vol. 4, no. 1, pp. 5–6, 2004.

[14] L. Li, K. Zhang, J. Lee, S. Cordes, D. P. Davis, and Z. Tang, "Discovering cancer genes by integrating network and functional properties," *BMC Medical Genomics*, vol. 2, article 61, 2009.

[15] S. Chakraborty, S. Khare, S. K. Dorairaj, V. C. Prabhakaran, D. R. Prakash, and A. Kumar, "Identification of genes associated with tumorigenesis of retinoblastoma by microarray analysis," *Genomics*, vol. 90, no. 3, pp. 344–353, 2007.

[16] A. Ganguly and C. L. Shields, "Differential gene expression profile of retinoblastoma compared to normal retina," *Molecular Vision*, vol. 16, pp. 1292–1303, 2010.

[17] R. J. Kinsella, A. Kähäri, S. Haider et al., "Ensembl BioMarts: a hub for data retrieval across taxonomic space," *Database*, vol. 2011, article bar030, 2011.

[18] Z. He, T. Huang, X. Shi et al., "Computational analysis of protein tyrosine nitration," in *Proceedings of the 4th International Conference on Computational Systems Biology (ISB '10)*, pp. 35–42, 2010.

[19] P. Carmona-Saez, M. Chagoyen, F. Tirado, J. M. Carazo, and A. Pascual-Montano, "GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists," *Genome Biology*, vol. 8, no. 1, article R3, 2007.

[20] 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.

[21] 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.

[22] 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.

[23] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ, USA, 1946.

[24] M. Kendall and A. Stuart, *The Advanced Theory of Statistics, vol. 2, Inference and Relationship*, Macmillan, New York, NY, USA, 1979.

[25] K. M. 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.

[26] 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.

[27] L.-L. Zheng, S. Niu, P. Hao, K. Feng, Y.-D. Cai, and Y. Li, "Prediction of protein modification sites of pyrrolidone carboxylic acid using mRMR feature selection and analysis," *PLoS One*, vol. 6, no. 12, Article ID e28221, 2011.

[28] Y.-F. Gao, B. Li -Q, Y.-D. Cai, K.-Y. Feng, Z.-D. Li, and Y. Jiang, "Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection," *Molecular BioSystems*, vol. 9, pp. 61–69, 2013.

[29] B.-Q. Li, Y.-D. Cai, K.-Y. Feng, and G.-J. Zhao, "Prediction of protein cleavage site with feature selection by random forest," *PLoS One*, vol. 7, Article ID e45854, 2012.

[30] N. Zhang, B.-Q. Li, S. Gao, J.-S. Ruan, and Y.-D. Cai, "Computational prediction and analysis of protein [gamma]-carboxylation sites based on a random forest method," *Molecular BioSystems*, vol. 8, pp. 2946–2955, 2012.

[31] B.-Q. Li, K.-Y. Feng, L. Chen, T. Huang, and Y.-D. Cai, "Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS," *PLoS One*, vol. 7, Article ID e43927, 2012.

[32] K. M. Ting and I. H. Witten, "Stacking bagged and dagged models," in *Proceedings of the 14th international Conference on Machine Learning*, pp. 367–375, San Francisco, Calif, USA, 1997.

[33] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2005.

[34] P. K. Srimani and M. S. Koti, "A Comparison of different learning models used in data mining for medical data," in *Proceedings of the 2nd International Conference on Methods and Models in Science and Technology (ICM2ST '11)*, pp. 51–55, India, November 2011.

[35] 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.

[36] Y.-D. Cai, L. Lu, L. Chen, and J.-F. He, "Predicting subcellular location of proteins using integrated-algorithm method," *Molecular Diversity*, vol. 14, no. 3, pp. 551–558, 2010.

[37] C. Peng, L. Liu, B. Niu et al., "Prediction of RNA-binding proteins by voting systems," *BioMed Research International*, vol. 2011, Article ID 506205, 8 pages, 2011.

[38] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI '95)*, vol. 2, pp. 1137–1145, 1995.

[39] B.-Q. Li, T. Huang, J. Zhang et al., "An ensemble prognostic model for colorectal cancer," *PLoS One*, vol. 8, Article ID e63494, 2013.

[40] 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, no. 4, Article ID e33393, 2012.

[41] E. P. Lamber, F. Beuron, E. P. Morris, D. I. Svergun, and S. Mittnacht, "Structural insights into the mechanism of phosphoregulation of the retinoblastoma protein," *PLoS One*, vol. 8, Article ID e58463, 2013.

[42] M. M. Kandalam, M. Beta, U. K. Maheswari, S. Swaminathan, and S. Krishnakumar, "Oncogenic microRNA 17-92 cluster is regulated by epithelial cell adhesion molecule and could be a potential therapeutic target in retinoblastoma," *Molecular Vision*, vol. 18, pp. 2279–2287, 2012.

[43] A. Loewer, E. Batchelor, G. Gaglia, and G. Lahav, "Basal dynamics of p53 reveal transcriptionally attenuated pulses in cycling cells," *Cell*, vol. 142, no. 1, pp. 89–100, 2010.

[44] C. Dai and W. Gu, "P53 post-translational modification: deregulated in tumorigenesis," *Trends in Molecular Medicine*, vol. 16, no. 11, pp. 528–536, 2010.

[45] S. Huang, Z. Zhu, Y. Wang et al., "Tet1 is required for Rb phosphorylation during G1/S phase transition," *Biochemical and Biophysical Research Communications*, vol. 434, no. 2, pp. 241–244, 2013.

[46] B. F. Clem and J. Chesney, "Molecular pathways: regulation of metabolism by RB," *Clinical Cancer Research*, vol. 18, pp. 6096–6100, 2012.

[47] G. Ciavarra and E. Zacksenhaus, "Rescue of myogenic defects in Rb-deficient cells by inhibition of autophagy or by hypoxia-induced glycolytic shift," *Journal of Cell Biology*, vol. 191, no. 2, pp. 291–301, 2010.

[48] G. Ciavarra and E. Zacksenhaus, "Multiple pathways counteract cell death induced by RB1 loss: implications for cancer," *Cell Cycle*, vol. 10, no. 10, pp. 1533–1539, 2011.