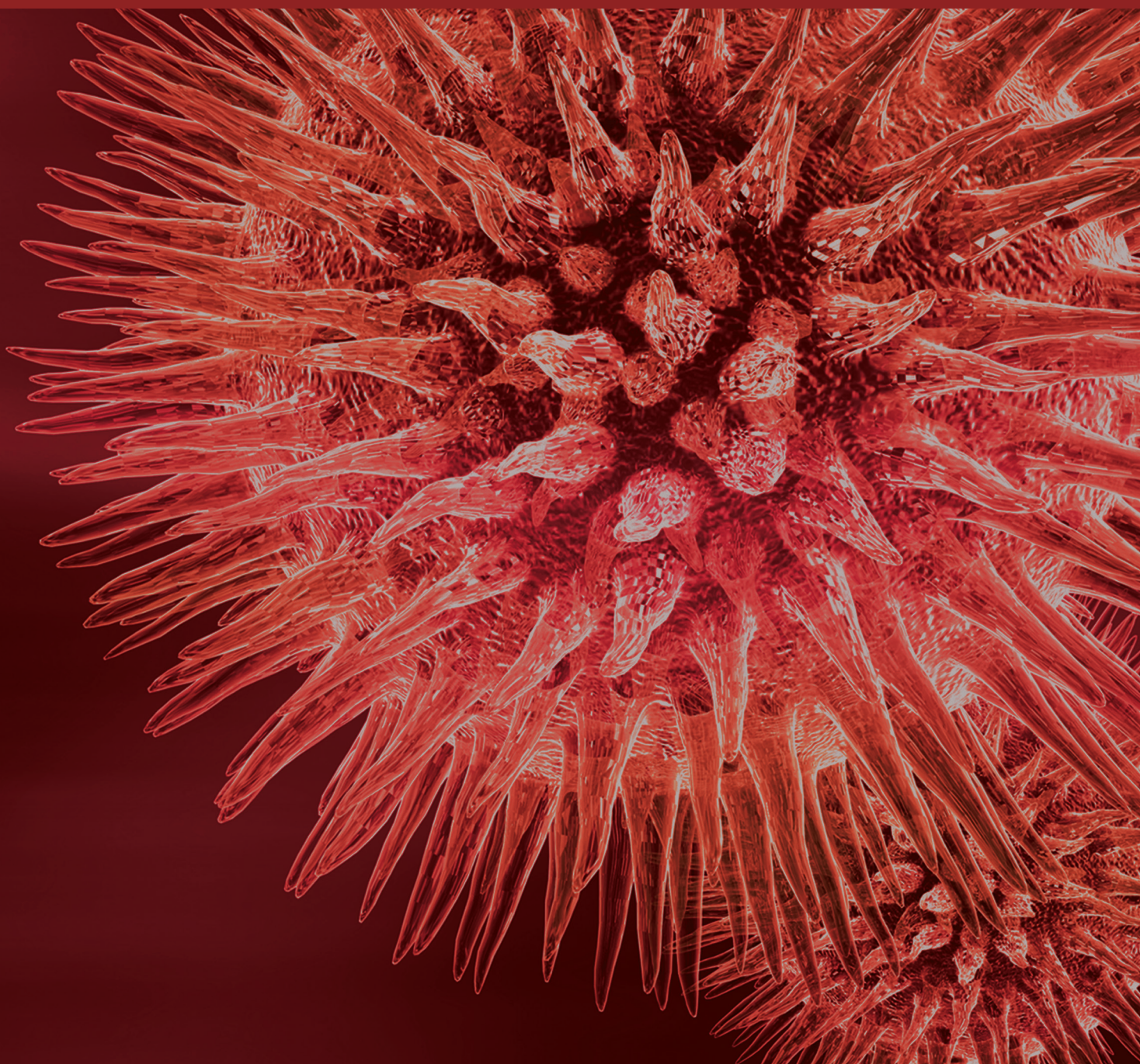


# Novel Computing Technologies for Bioinformatics and Cheminformatics

Guest Editors: Chuan Yi Tang, Che-Lun Hung, Ching-Hsien Hsu, Huiru Zheng, and Chun-Yuan Lin





---

# **Novel Computing Technologies for Bioinformatics and Cheminformatics**

BioMed Research International

---

## **Novel Computing Technologies for Bioinformatics and Cheminformatics**

Guest Editors: Chuan Yi Tang, Che-Lun Hung,  
Ching-Hsien Hsu, Huiru Zheng, and Chun-Yuan Lin



---

Copyright © 2014 Hindawi Publishing Corporation. All rights reserved.

This is a special issue published in “BioMed Research International.” All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Contents

**Novel Computing Technologies for Bioinformatics and Cheminformatics**, Chuan Yi Tang, Che-Lun Hung, Ching-Hsien Hsu, Huiru Zheng, and Chun-Yuan Lin  
Volume 2014, Article ID 392150, 3 pages

**A Priori Knowledge and Probability Density Based Segmentation Method for Medical CT Image Sequences**, Huiyan Jiang, Hanqing Tan, and Benqiang Yang  
Volume 2014, Article ID 769751, 11 pages

**Local Alignment Tool Based on Hadoop Framework and GPU Architecture**, Che-Lun Hung and Guan-Jie Hua  
Volume 2014, Article ID 541490, 7 pages

**Double-Bottom Chaotic Map Particle Swarm Optimization Based on Chi-Square Test to Determine Gene-Gene Interactions**, Cheng-Hong Yang, Yu-Da Lin, Li-Yeh Chuang, and Hsueh-Wei Chang  
Volume 2014, Article ID 172049, 10 pages

**Novel Design Strategy for Checkpoint Kinase 2 Inhibitors Using Pharmacophore Modeling, Combinatorial Fusion, and Virtual Screening**, Chun-Yuan Lin and Yen-Ling Wang  
Volume 2014, Article ID 359494, 13 pages

**New Strategies for Evaluation and Analysis of SELEX Experiments**, Rico Beier, Elke Boschke, and Dirk Labudde  
Volume 2014, Article ID 849743, 12 pages

**A Novel Approach for Discovering Condition-Specific Correlations of Gene Expressions within Biological Pathways by Using Cloud Computing Technology**, Tzu-Hao Chang, Shih-Lin Wu, Wei-Jen Wang, Jorng-Tzong Horng, and Cheng-Wei Chang  
Volume 2014, Article ID 763237, 8 pages

**Gene Prioritization of Resistant Rice Gene against *Xanthomas oryzae pv. oryzae* by Using Text Mining Technologies**, Jingbo Xia, Xing Zhang, Daojun Yuan, Lingling Chen, Jonathan Webster, and Alex Chengyu Fang  
Volume 2013, Article ID 853043, 9 pages

**Enabling Large-Scale Biomedical Analysis in the Cloud**, Ying-Chih Lin, Chin-Sheng Yu, and Yen-Jen Lin  
Volume 2013, Article ID 185679, 6 pages

**A Novel Framework for the Identification and Analysis of Duplicons between Human and Chimpanzee**, Trees-Juen Chuang, Shian-Zu Wu, and Yao-Ting Huang  
Volume 2013, Article ID 264532, 12 pages

## Editorial

# Novel Computing Technologies for Bioinformatics and Cheminformatics

Chuan Yi Tang,<sup>1</sup> Che-Lun Hung,<sup>2</sup> Ching-Hsien Hsu,<sup>3</sup> Huiru Zheng,<sup>4</sup> and Chun-Yuan Lin<sup>5</sup>

<sup>1</sup>Department of Computer Science and Information Engineering, Providence University, Taichung 433, Taiwan

<sup>2</sup>Department of Computer Science and Communication Engineering, Providence University, Taichung 433, Taiwan

<sup>3</sup>Department of Computer Science and Information Engineering, Chung Hua University, Hsinchu 30012, Taiwan

<sup>4</sup>School of Computing and Mathematics, Computer Science Research Institute, University of Ulster, Jordanstown BT370QB, UK

<sup>5</sup>Department of Computer Science and Information Engineering, Chang Gung University, Taoyuan 333, Taiwan

Correspondence should be addressed to Chuan Yi Tang; [cytang@cs.nthu.edu.tw](mailto:cytang@cs.nthu.edu.tw)

Received 18 September 2014; Accepted 18 September 2014; Published 28 December 2014

Copyright © 2014 Chuan Yi Tang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## 1. Introduction

In the past, many computing technologies have been proposed and utilized to accelerate biologists/chemists to analyze biological and chemical data, such as homology detection, evolutionary analysis, function prediction, computer-aided drug design, and cheminformatics. Leveraging a power of these technologies, a lot of tools and services are valuable for biologists/chemists to efficiently analyze large-scale and complicated data. However, today's data is being generated and collected at an incredible scale, the buzzword "big data." For instance, an individual laboratory can generate terabase scales of DNA and RNA sequencing data within a day by next-generation sequencing technologies. It is difficult to manage and process big biological and chemical data using conventional methods due to not only their size but also their complexity. It requires entirely different thoughts, while the major obstacle could be the complexity, size, or integration of various data sources. These barriers spur the revolutions of both storage and computing technologies whereby the developed tool and service can be highly scalable, totally reliable, more elastic, and so on.

Therefore, the computing technologies required to maintain, process, and integrate the large amounts of data are beyond the reach of small laboratories and introduce serious challenges even for large institutes. Success at the bioinformatics and cheminformatics fields will heavily rely on an ability to explain these large-scale and great diversification data,

which encourage biologists/chemists to adopt novel computing technologies. The research papers selected for this special issue represent recent progresses in the aspects, including theoretical studies, practical applications, novel strategies and framework, high performance computing technologies, method and algorithm improvement, and review. All of these papers not only provide novel ideas and state-of-the-art technologies in the field but also stimulate future research for Bioinformatics and Cheminformatics.

## 2. Large-Scale Biomedical Analysis

Recent progress in high-throughput instrumentations has led to an astonishing growth in both volume and complexity of biomedical data collected from various sources. The planet-size data brings serious challenges to the storage and computing technologies. The paper by Y.-C. Lin et al. entitled "Enabling large-scale biomedical analysis in the cloud" indicates the coming age of sharp data growth and increasing data diversification is a major challenge for biomedical research in the postgenome era. Cloud computing is an alternative to crack the nut because it gives concurrent consideration to enable storage and massive computing on large-scale data. Developing cloud-based biomedical applications can integrate the vast amount of diversification data in one place and analyze them on a continuous basis. This would make a significant breakthrough to launch a high quality healthcare. This

review paper briefly introduces the data intensive computing system and summarizes existing cloud-based resources in bioinformatics. These developments and applications would facilitate biomedical research to make the vast amount of diversification data meaningful and usable.

With the rapid growth of next generation sequencing technologies, more and more data have been discovered and published. To analyze such huge data, the computational performance is an important issue. The paper by C.-L. Hung and G.-J. Hua entitled “*Local alignment tool based on hadoop framework and GPU architecture*” combines two different heterogeneous architectures, software architecture-Hadoop framework and hardware architecture-GPU, to develop a high performance cloud computing service, called Cloud-BLASTP, for protein sequence alignment. Cloud-BLASTP takes advantage of high performance, availability, reliability, and scalability. Cloud-BLASTP guarantees that all submitted jobs are properly completed, even when running job on an individual node or mapper experience failure. The performance experiment shows that Cloud-BLASTP is faster than GPU-BLASTP and is desirable for biologists to investigate the protein structure and function analysis by comparing large protein database under reasonable time constraints.

Organ segmentation is a crucial step prior to computer-aided diagnosis, since it is fundamental for further medical image processing such as cancer detection, lesion recognition, and three-dimensional visualization. However, organ extraction is considered as a challenge task due to huge shape variations, heterogeneous intensity distribution, and low contrast of CT image. The paper by H. Jiang et al. entitled “*A priori knowledge and probability density based segmentation method for medical CT image sequences*” briefly introduces a novel segmentation strategy for CT images sequences. In their strategy, a priori knowledge is effectively used to guide the determination of objects and a modified distance regularization level set method can accurately extract actual contour of object in a short time. Their proposed method is compared to other seven state-of-the-art medical image segmentation methods, GAC, C-V, SPLS, HLS, SDLS, CCRG, and IVLS, on abdominal CT image sequences datasets. The evaluated results demonstrate their method performs better and has the potential for segmentation in CT image sequences.

### 3. Novel Strategies for Drug Design

Quantitative structure-activity relationships (QSAR) is a widely adapted computational method that correlates the structural properties of compounds with their biological activities, such as the affinity between the ligand and protein and the toxicity of existing/hypothetical molecules. Recently, the prediction quality using the QSAR method was improved by considering the three-dimensional structure (3D-QSAR) of targeted inhibitors. The paper by C.-Y. Lin and Y.-L. Wang entitled “*Novel design strategy for checkpoint kinase 2 inhibitors using pharmacophore modeling, combinatorial fusion, and virtual screening*” proposes a novel design strategy for drug design by applying combinatorial fusion into pharmacophore hypotheses and virtual screening techniques. They first used 3D-QSAR study to build pharmacophore

hypotheses for Chk2 inhibitors by HypoGen Best, Fast, and Caesar algorithms, respectively. Then, they used the combinatorial fusion to select and combine prediction results for improving the predictive accuracy in biological activities of inhibitors. Finally, all of feasible compounds in NCI database were selected by using ligand-based virtual screening.

Aptamers are an interesting alternative to antibodies in pharmaceuticals and biosensorics, because they are able to bind to a multitude of possible target molecules with high affinity. Therefore, the process of finding such aptamers, which is commonly a SELEX screening process, becomes crucial. The standard SELEX procedure schedules the validation of certain found aptamers via binding experiments, which is not leading to any detailed specification of the aptamer enrichment during the screening. The paper by R. Beier et al. entitled “*New strategies for evaluation and analysis of SELEX experiments*” uses sequence information gathered by next generation sequencing techniques on SELEX experiments. They propose a motif search algorithm which helps to describe the aptamers enrichment in more detail. The extensive characterization of target and binding aptamers may later reveal a functional connection between these molecules, which can be modeled and used to optimize future SELEX runs in case of the generation of target-specific starting libraries.

### 4. Computational Genomics

Human and other primate genomes consist of segmental duplications due to fixation of copy number variations. Structure of these duplications within the human genome has been shown to be a complex mosaic composed of juxtaposed subunits, called duplicons. These duplicons are difficult to be uncovered from the mosaic repeat structure. In addition, the distribution and evolution of duplicons among primates are still poorly investigated. The paper by T.-J. Chuang et al. entitled “*A novel framework for the identification and analysis of duplicons between human and chimpanzee*” develops a statistical framework for discovering duplicons via integration of a Hidden Markov Model (HMM) and a permutation test. Their experimental results indicate that the mosaic structure composed of duplicons is common in copy number variations and segmental duplications of both human and chimpanzee. Gene ontology analysis, hierarchical clustering, and phylogenetic analysis of duplicons also were used in their work and then suggested that most copy number variations/segmental duplications share common duplication ancestry.

Due to the availability of abundant genomic resources, rice has become a model species for the genomic study. Bacterial blight, caused by *Xanthomonas oryzae* pv. *oryzae* (Xoo), is a worldwide devastating disease, and bacterial blight resistance genes have been cloned by a map-based cloning approach. It is important to find a more effective way to locate vital resistant genes. The text mining strategy represents another effective way to improve the efficiency of gene discovery. The paper by J. Xia et al. entitled “*Gene prioritization of resistant rice gene against Xanthomonas oryzae pv. oryzae by using text mining technologies*” proposes a hybrid strategy to enhance gene prioritization by combining text

mining technologies with a sequence-based approach. Their scheme consists of two sieves, the text-mining sieve and the classifier sieve. The text-mining sieve is to screen candidate gene according to the important phrase evaluation through  $TF * IDF$  and voting scheme. The classifier sieve is a built-in classifier based on chaos games representation. Their experiment results show that the hybrid strategy achieves enhanced gene prioritization.

## 5. Computational Systems Biology

Genome-wide association studies for the analysis of gene-gene interaction are important fields for detecting the effects of cancer and disease. Such studies usually entail the collection of a vast number of samples and SNPs selected from several related genes of disease in order to identify the association amongst genes. A method for searching high-order interactions is needed to determine the potential association between several loci. Statistical methods are widely used to search for a good model of gene-gene interaction for disease analysis; however, the huge numbers of potential combinations of SNP genotypes limit the use of statistical methods for analysing high-order interaction. It remains a challenge to find an available high-order model of gene-gene interaction. The paper by C.-H. Yang et al. entitled “*Double-bottom chaotic map particle swarm optimization based on chi-square test to determine gene-gene interactions*” presents an improved particle swarm optimization with double-bottom chaotic maps (DBM-PSO) to assist statistical methods in the analysis of associated variations to disease susceptibility. Analysis results supported that the DBM-PSO is a robust and precise algorithm, and it can identify good models and provide higher chi-square values than conventional PSO.

Using microarray technology combined with computational analysis is one of the most efficient and cost-effective methods for studying cancer. Most studies focus primarily on identifying differential gene expressions between conditions, for discovering the major factors that cause diseases. Previous studies have not identified the correlations of differential gene expression between conditions; crucial but abnormal regulations that cause diseases might have been disregarded. The paper by T.-H. Chang et al. entitled “*A novel approach for discovering condition-specific correlations of gene expressions within biological pathways by using cloud computing technology*” proposes a novel approach for discovering the condition-specific correlations of gene expressions within biological pathways. An Apache Hadoop cloud computing platform was implemented to reduce the time for analyzing gene expression correlations. The experimental results showed that breast cancer recurrence might be highly associated with the abnormal regulations of these gene pairs, rather than with their individual expression levels. Their proposed method was computationally efficient and reliable for identifying meaningful biological regulation patterns between conditions.

## 6. Conclusions

All of the above papers address either cloud computing service or novel strategies for large-scale biomedical analysis and

drug design. They also develop related method and approach improvements in applications of computational genomics and systems biology. Honorably, this special issue serves as a landmark source for education, information, and reference to professors, researchers, and graduate students interested in updating their knowledge about or active in biomedical analysis, drug design, computational genomics, and systems biology.

## Acknowledgments

The guest editors would like to express sincere gratitude to numerous reviewers for their professional effort, insight, and hard work put into commenting on the selected articles which reflect the essence of this special issue. We are grateful to all authors for their contributions and for undertaking two-cycle revisions of their manuscripts, without which this special issue could not have been produced.

Chuan Yi Tang  
Che-Lun Hung  
Ching-Hsien Hsu  
Huiru Zheng  
Chun-Yuan Lin



## Research Article

# A Priori Knowledge and Probability Density Based Segmentation Method for Medical CT Image Sequences

Huiyan Jiang,<sup>1</sup> Hanqing Tan,<sup>1</sup> and Benqiang Yang<sup>2</sup>

<sup>1</sup> Software College, Northeastern University, Shenyang 110819, China

<sup>2</sup> Radiology Department, PLA General Hospital, Shenyang 110016, China

Correspondence should be addressed to Huiyan Jiang; [hyjiang@mail.neu.edu.cn](mailto:hyjiang@mail.neu.edu.cn)

Received 4 October 2013; Accepted 28 April 2014; Published 19 May 2014

Academic Editor: Huiru Zheng

Copyright © 2014 Huiyan Jiang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper briefly introduces a novel segmentation strategy for CT images sequences. As first step of our strategy, we extract a priori intensity statistical information from object region which is manually segmented by radiologists. Then we define a search scope for object and calculate probability density for each pixel in the scope using a voting mechanism. Moreover, we generate an optimal initial level set contour based on a priori shape of object of previous slice. Finally the modified distance regularity level set method utilizes boundaries feature and probability density to conform final object. The main contributions of this paper are as follows: a priori knowledge is effectively used to guide the determination of objects and a modified distance regularization level set method can accurately extract actual contour of object in a short time. The proposed method is compared to other seven state-of-the-art medical image segmentation methods on abdominal CT image sequences datasets. The evaluated results demonstrate our method performs better and has the potential for segmentation in CT image sequences.

## 1. Introduction

Organ segmentation is a crucial step prior to computer-aided diagnosis, since it is fundamental for further medical image processing such as cancer detection, lesion recognition, and three-dimensional visualization. However, organ extraction is considered as a challenge task due to huge shape variations, heterogeneous intensity distribution, and low contrast of CT image [1]. Especially complicated surrounding and weak edge cause serious impediment to accurately segment pancreas.

Various methods are proposed to solve the medical image segmentation problem. The main categories of these methods can be classified as statistical shape model (SSM) [2], level set [3–8], probabilistic atlases [9], histogram-based approaches [10], and region growing method [11, 12].

The statistical shape model and probabilistic atlases seriously depend on the shape and intensity distribution of objects in training dataset, so that they suffer from large variations of shape and intensity. The histogram-based approaches always use a classification system to differentiate target object

from other tissues; the leakage problem exists in these systems.

Level set methods can represent complex topology of contours and handle topological changes in a natural and effective way, such that various level set methods are proposed to solve the medical image segmentation problem. The shape detection level set method [3] applies a shape modeling scheme in level set evolution. The geodesic active contour (GAC) [4] model employs edge feature to guide segmentation. However these edge-based level set methods easily cause leakage in weak boundaries of objects. The C-V model [5] which seeks global optimization is not suitable for local optimization segmentation. A hybrid level set method [6] combines both boundary and region information to achieve segmentation results. It utilizes a predefined parameter to indicate the lower bound of the gray level of the target object in region term. Its boundary term is similar to the one in GAC method. However its predefined parameter is not easy to be accurately defined and reinitialization of zero level set is needed. A priori shape based level set method [7] uses a priori

shape knowledge to guide the segmentation, but it suffers from large variations of shape and intensity distribution. Moreover, level set methods have a high requirement to locate initial zero level set near final contour. The similarity between nearby slices in CT image sequences is ignored in level set methods. The problem of leakage easily happens in weak boundary area.

In order to solve these problems, this paper proposes a novel segmentation strategy that regards similarity of intensity distribution, shape, and location between nearby slices as a priori knowledge to guide the segmentation of image sequences. The kernel of this paper is that a probability density map which is generated using the novel application strategy of a priori knowledge is used to modify a distance regularization level set method. The proposed method is compared to geodesic active contour model, C-V model, shape detection level set method, the hybrid level set method, and confident connected region growing method. Finally the novel method is compared to our previous improved variational level set method [8]. The evaluated results prove that our method is effective to segment organs from abdominal CT image sequences. The rest of this paper is arranged as follows. The proposed method is explained in Section 2. Evaluation and discussion of our method are presented in Section 3, and Section 4 concludes this paper.

## 2. Materials and Methods

**2.1. Distance Regularity Level Set.** A distance regularity level set method is proposed in [13]. This method inherently maintains a signed distance profile near the zero level set, such that it eliminates the requirement of reinitialization of level set function. It is able to provide accurate numerical calculation in level set evolution.

The energy function of level set is define by

$$E(\phi) = \mu R_p(\phi) + \beta \eta(\phi), \quad (1)$$

where  $\mu > 0$  is a constant,  $R_p(\phi)$  is level set distance regularization term, and  $\eta(\phi)$  is external force term.

$R_p(\phi)$  is defined in [11] by

$$R_p(\phi) \triangleq \int_{\Omega} p(|\nabla\phi|) dx, \quad (2)$$

where  $p$  is a double-well potential function for the distance regularization term  $R_p$  and is constructed as

$$p(s) = \begin{cases} \frac{1}{(2\pi)^2} (1 - \cos(2\pi s)), & \text{if } s \leq 1 \\ \frac{1}{2}(s-1)^2, & \text{if } s > 1. \end{cases} \quad (3)$$

$\delta_\varepsilon$  and  $H_\varepsilon$  are smooth functions in level set methods proposed in [14, 15]. Moreover,  $H'_\varepsilon = \delta_\varepsilon$  and  $\varepsilon$  is set to 1.5.

$$\delta_\varepsilon(x) = \begin{cases} \frac{1}{2\varepsilon} \left[ 1 + \cos\left(\frac{\pi x}{\varepsilon}\right) \right] & |x| \leq \varepsilon \\ 0 & |x| > \varepsilon, \end{cases}$$

$$H_\varepsilon(x) = \begin{cases} \frac{1}{2} \left( 1 + \frac{x}{\varepsilon} + \frac{1}{\pi} \sin\left(\frac{\pi x}{\varepsilon}\right) \right) & |x| \leq \varepsilon \\ 1 & x > \varepsilon \\ 0 & x < -\varepsilon. \end{cases} \quad (4)$$

The  $R_p(\phi)$  makes the level set evolution have a unique forward-and-backward diffusion effect, which eliminates the need for reinitialization, such that its induced numerical errors are avoided. Therefore level set evolution is more stable and robust.

**2.2. A Priori Information Extraction.** The traditional a priori knowledge such as shape and intensity distribution is always extracted from training dataset, which represents the commonness of object but cannot directly represent the individual characteristics of the current object in medical image. The differences between commonness and individuality usually cause errors in final segmentation results. Moreover, the large variation of shape and intensity distribution of organs bring a great difficulties in using traditional commonness to guide the segmentation.

In order to overcome these problems, a new scheme is proposed to extract the individuality feature of object as a priori knowledge which is then employed to optimize the segmentation process of level set method. As the first step of processing, we check through the input abdominal CT volume to find out a slice in which object organs have a largest cross-section. A radiologist defines the boundary of organs in this slice. The shape of boundary and the intensity distribution parameters of this object organ region are used as a priori knowledge in the next step of segmentation.

Though variation of shape and intensity is obvious between different volumes or slices that have a large imaging distance in the same volume, these features in neighbor slices which belong to the same volume are similar. Thus, we follow the a priori shape of previous slice to segment next slice. The statistics dataset is initial as the manually segmented slice. Subsequent segmented results will be added into the statistics dataset as statistical sample.

Each segmented sample in the training dataset is regarded as a scope of statistics. Mean intensity and intensity variance for each sample are calculated:

$$u = \frac{1}{n} \sum_{i=1}^n p_i \quad p_i \in R_s,$$

$$\sigma = \frac{1}{n} \sqrt{\sum_{i=1}^n (p_i - u)^2} \quad p_i \in R_s. \quad (5)$$

$p_i$  is the intensity value of pixels in samples. All the pairs of parameters make a statistical feature set

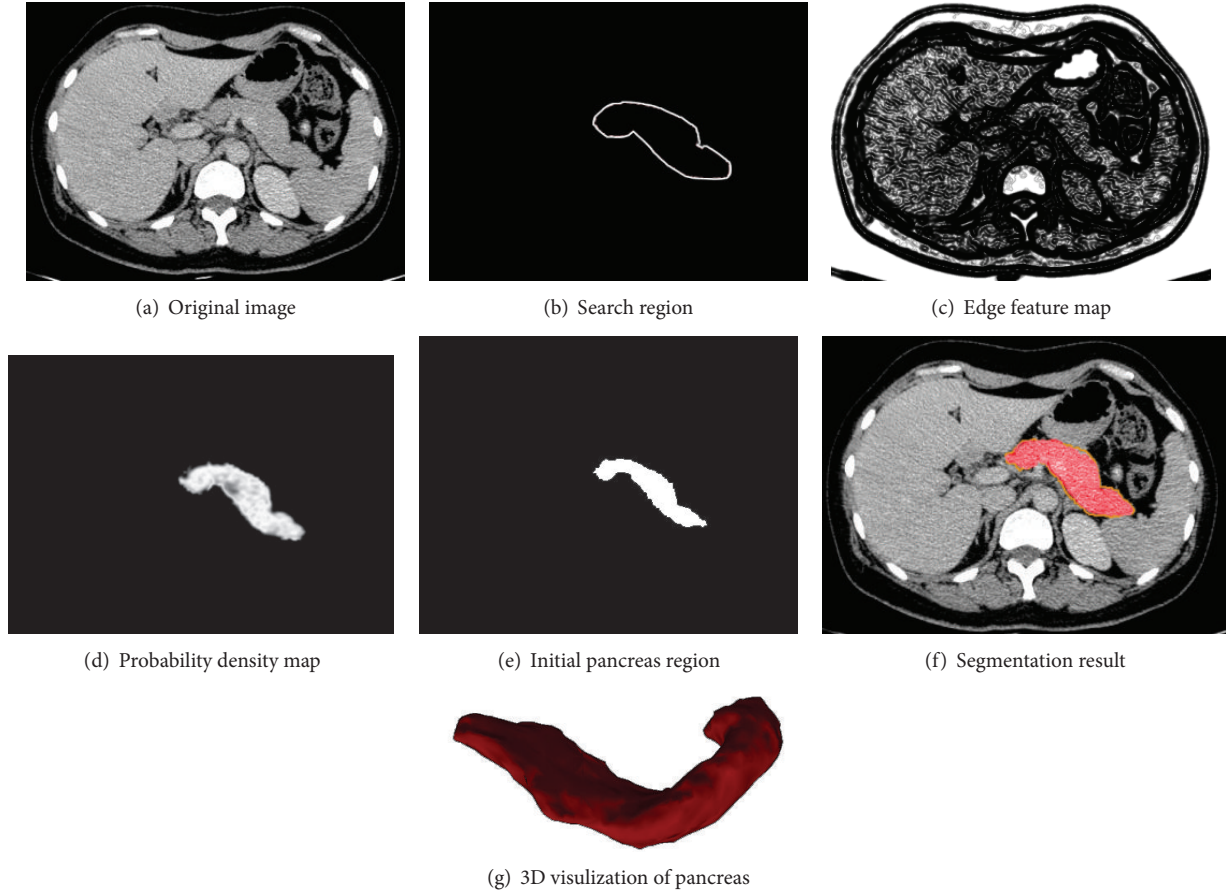


FIGURE 1: Segmentation example: necessary reprocessing for pancreas segmentation. (a) Denoised CT image. (b) The mask of previous slice. (c) The search region of pancreas based on mask of previous slice. (d) Edge feature map. (e) Probability density map. (f) Initial pancreas region used in our level set method.

$F = \{(u_i, \sigma_i) \mid i = 1, 2, \dots, n\}$ , which plays an important role in processing of segmentation. Through amount of statistical experiments, the statistics indicate that about 92.4% of pixels in object region is located in  $[\mu - 2\sigma, \mu + 2\sigma]$ .

**2.3. A Priori Based Distance Regularity Level Set.** Some defects exist in the original distance regularization level set method. It is sensitive to initial position of the zero level set contour. The initial zero level set is required to locate near the final contour. Otherwise, the curve evolution needs amount of iterative calculation to pull curve toward object contour. Moreover, original distance regularization level set method has oversegmentation problem of leakages into nearby tissue in weak boundary area. Especially object is always connected to neighbor organs and boundary usually is fuzzy in CT image; the original method cannot get satisfactory results in most case.

In order to solve these problems, we employ a priori statistical feature to modify distance regularity level set as well as confirming an optimal initial level set. Then the modified method is used to extract the object organ from CT images.

The statistical information which comes from statistical dataset is added into the external energy term of energy

function of level set, such that new energy function is defined as

$$E(\phi) = \mu R_p(\phi) + \alpha S(\phi) + \lambda L_g(\phi), \quad (6)$$

where the first term is distance regularization term, the second and third terms are external energy terms, which are used to pull the initial curve toward the final object curve in evolution.  $\lambda > 0$  and  $\alpha \in \mathfrak{R}$  are coefficients to control the weight of external energy.  $L_g(\phi)$  depends on image gradient information and  $S(\phi)$  relays on a priori statistical feature. They correspond to  $\eta(\phi)$  in function (1).

$S(\phi)$  is defined as

$$S(\phi) = \int_{\Omega} s(I_m) H(-\phi) dx dy, \quad (7)$$

where  $I_m = MI$  is a search area which contains all pixels of current object region.  $M$  is a mask function used to define a search domain which includes object organ in the CT slice  $I$ . The mask  $M$  derives from the extracted object region of previous slice of current slice  $I$ . The previous object region extends outward  $n$  pixel along its shape to generate the mask scope (see Figure 1). The pixels inside the scope are set to 1 and those outside the scope set to 0. Since the location and shape

are similar between two contiguous slices,  $s(I_m)$  is a similarity measure function. It estimates the probability of belonging to object tissue of each pixel in search area.

In order to measure the similarity, first a probability density formula is defined as

$$p(x) = \begin{cases} e^{-(x-\mu)^2/2\sigma^2}, & x \in [\mu - 2\sigma, \mu + 2\sigma] \\ -\frac{|x - \mu|}{2\sigma}, & \text{otherwise,} \end{cases} \quad (8)$$

where  $p(x)$  is probability density.  $x$  is an intensity value of pixel within search area.  $\mu$  is mean intensity, and  $\sigma$  is intensity variance. They come from statistical feature set  $F$ . For each pixel within search area, a set of probability density  $P = \{p_1, p_2, \dots, p_n\}$  is calculated based on all statistical features  $\{(u_i, \sigma_i) \mid i = 1, 2, \dots, n\}$ .

A voting mechanism is employed to determine the actual probability density of a pixel. The voting mechanism is defined as

$$\text{Votes}(F) \triangleq \begin{cases} V_{x,a} + 1 & x \in [\mu_i - 2\sigma_i, \mu_i + 2\sigma_i] \\ V_{x,o} + 1 & \text{otherwise,} \end{cases} \quad (9)$$

where  $V_{x,a}$  represents affirmative vote and  $V_{x,o}$  represents negative vote. If intensity of pixel is located in  $[\mu - 2\sigma, \mu + 2\sigma]$ , the  $V_{x,a}$  increases by one. Otherwise,  $V_{x,o}$  increase by one. The total votes are equal to the number of statistical features  $V_{x,a} + V_{x,o} = n$ .

Based on the votes and probability density set, the actual probability density of a pixel within search area is confirmed as

$$s(x) = \begin{cases} P_{x,\max} & V_{x,a} > V_{x,o} \\ P_{x,\min} & V_{x,a} \leq V_{x,o}, \end{cases} \quad (10)$$

where  $P_{x,\max}$  is the maximal value in probability density set and  $P_{x,\min}$  is the minimum value. If affirmative votes are more than negative votes, the probability density of a pixel is set to maximum in probability density set. On the contrary, it is set to minimum in probability density set.

A probability density map  $s(I_m)$  is generated after probability density of all pixels within search region is ascertained using voting mechanism. It is used to limit oversegmentation. The  $S(\phi)$  term can speed up the propagation motion of zero level set when the initial contour is far away from the desired object boundaries.

Moreover, the second energy term  $L_g(\phi)$  represents edge force which pushes the initial curve towards the boundaries of the object. It is defined as

$$L_g(\phi) = \int_{\Omega} g(I) \delta_{\varepsilon}(\phi) |\nabla\phi| dx, \quad (11)$$

where  $g(I)$  is an edge detection function which is defined as

$$g(I) = \frac{1}{1 + |\nabla[G * I]|^2}, \quad (12)$$

where  $G$  is Gaussian filtering operator.  $*$  means convolution.  $I$  is the CT image. Edge force is minimized when the contour

of zero level set is located at boundaries of object, because edge detection function takes small value at boundaries.

In order to generate an optimal initial level set, which can satisfy the location requirement of initial zero level set, we apply a mask of previous slice to define the initial contour of zero level set. Since the shape variation is not obvious between two adjacent slices, the extracted object region of previous slice is regarded as a priori shape mark. The binary mask shrinks  $k$  pixel along its shape to generate an initial contour (See Figure 1(e)). The initial contour is located in the object region of current slice, because location of object organ in adjacent slices is similar.

The initial level set function (LSF)  $\phi_0$  is defined as a binary step function:

$$\phi_0(x) = \begin{cases} -c, & \text{if } x \in R_0 \\ c, & \text{otherwise,} \end{cases} \quad (13)$$

where the  $R_0$  is the initial contour region.  $c$  is a constant set to 2.

The level set evolution equation in a priori based distance regularity level set is finally defined by

$$\begin{aligned} \frac{\partial\phi}{\partial t} = & \mu \operatorname{div} \left( d_p (|\nabla\phi| \nabla\phi) + \alpha s(I_m) \delta_{\varepsilon}(\phi) \right. \\ & \left. + \lambda \delta_{\varepsilon}(\phi) \operatorname{div} \left( g(I) \frac{\nabla\phi}{|\nabla\phi|} \right) \right), \end{aligned} \quad (14)$$

where  $\operatorname{div}(\cdot)$  is the divergence operator and  $d_p$  is a function defined in [11]:

$$d_p(s) \triangleq \frac{p'(s)}{s}. \quad (15)$$

**2.4. Object Organ Segmentation.** A priori based distance regularity level set method is applied to extract the object organ in CT images. Since the intensity distribution of the object organ is irregular due to the noise caused in the image formation stage, a Gaussian blur filter is used to reduce the noise in preprocess. The steps of segmentation process are shown in Figure 3.

- (1) Initialize the training dataset by manually segmenting a slice in which object organ has a largest cross-section in input abdominal CT volume. Its next slice is the first one to segment.
- (2) Based on training dataset, generate the statistical feature set which is regarded as a priori knowledge and used to guide segmentation of pancreas.
- (3) Reduce the noise in CT slice using a Gaussian blur filter.
- (4) Generate a search region based on mask of previous slice and then calculate the probability density map using voting mechanism.

- (5) Generate an optimal initial zero level set based on mask of previous slice.
- (6) Based on optimal initial zero level set, extract the object using a priori based distance regularity level set method.

The extracted object will be added into training dataset as a priori knowledge to guide the segmentation of its next slice.

In practical process of object segmentation, a two-phase segmentation scheme is employed to get a better result. The first phase can be seen as a high speed level set evolution and the second phase can be seen as a high accurate level set evolution. In the first phase, the zero level set is initialized as a binary step function using function (13). The level set evolution follows function (14). After the first phase, the zero level set contour is closed to the object boundary. In the second phase, the main purpose is to accurately extract the object region. The level set evolution equation is reset as

$$\frac{\partial \phi}{\partial t} = \mu \operatorname{div} \left( d_p (|\nabla \phi| \nabla \phi) + \lambda \delta_\varepsilon(\phi) \operatorname{div} \left( g(I) \frac{\nabla \phi}{|\nabla \phi|} \right) \right). \quad (16)$$

Because the energy term  $S(\phi)$  pushes the initial contour toward the final boundary in a high speed, it is likely to make the contour across the object boundary and then cause oversegmentation. Thus, it is abolished in the second phase.

Through amount of experiment, we empirically define some values of parameters of great significance to optimize the segmentation result. In this configuration of parameters, the average similarity index of all segmentation results can get a high rate (SI = 0.922, introduced in Section 3.1).

In the first phase,  $u = 0.2$ ,  $\lambda = 3$ , and  $\alpha = -1$  are employed in (14). A small coefficient  $\alpha$  for the energy term  $S(\phi)$  is to restrict contour expanding too rapidly and preserve the zero level set contour from crossing the boundary of object region. The iterator time in first phase is set between 5 and 10.

In the second phase, the zero level set contour is closed to the boundary of object, such that  $u = 0.2$ ,  $\lambda = 2$ , and  $\alpha = 0$  are employed. Level set evolution is dominated by edge force. A large weight is assigned to energy term  $L_g(\phi)$ , which means a stronger constraint force of boundary pushes zero level set curve towards final boundary while limiting the oversegmentation of object region. The iterator time is set between 3 and 5 in this phase.

The segmentation results of different shape and acreage of object are controlled by adjusting the iteration time. Moreover, the parameters can be fine-tuned to adapt with different CT volume to get an optimal result.

### 3. Results and Discussion

The proposed method is compared to geodesic active contour method (GAC), geodesic active without edge method (C-V), shape a priori based level set method (SPLS), a hybrid level set method (HLS), a shape detection level set method (SDLS), confident connected region growing method (CCRG), and improved variational level set method (IVLS). Our method is referred to as PBDR. Our method, GAC method, shape

detection level set method, shape a priori based level set method, and improved variational level set method are implemented using C/C++ language. C-V method and HLS method are implemented in Matlab code. All methods run on a desktop PC with 8 GB RAM and 2.4 GHz Intel Core i7 processor. The same preprocess are applied to all methods.

The trade-off between number of manual labelling and algorithm efficiency of proposed method is also evaluated. Based on a volume with 161 CT abdominal images, different numbers of manual labelling are applied as a priori knowledge to guide the segmentation.

**3.1. Performance Measure Standard.** For evaluation of efficiency and accuracy, three measures, (1) false positive error (FPE), (2) false negative error (FNE), and (3) the similarity index (SI), are used to measure the performance of methods.

False positive error [16] is defined as the ratio of the total number of extracted object region pixels outside the golden standard region to the total number of golden standard of object region:

$$\text{FPE} = \frac{N(O) \cap N(B)}{N(G)} \times 100\%, \quad (17)$$

where  $O$  represents the pixels of extracted object region.  $G$  represents the golden standard of object organ.  $B$  represents the remaining areas except the region of golden standard in the CT image.  $N(O) \cap N(B)$  represents the total number of extracted object region pixels outside the golden standard region.  $N(G)$  represents the total number of golden standard of object region.

False negative error [16] is defined as the ratio of the total number of golden standard of object outside the extracted object region to the total number of pixels of golden standard of object region:

$$\text{FNE} = \frac{N(G) - (N(O) \cap N(G))}{N(G)} \times 100\%, \quad (18)$$

where  $N(O) \cap N(G)$  is total number of pixels in intersection of extracted object region and golden standard of object.  $N(G) - (N(O) \cap N(G))$  is the total number of golden standard of object outside the extracted object region.

Similarity index [17] is defined as the percentage of pixels in intersection of extracted object region and golden standard of object:

$$\text{SI} = \frac{2(N(O) \cap N(G))}{N(O) + N(G)} \times 100\%, \quad (19)$$

where  $N(O)$  is the total number of extracted object region.

**3.2. Experimental Datasets.** Three medical image datasets including pancreas dataset, liver dataset, and spleen dataset are used in evaluation. Pancreas dataset contains 10 volumes of CT image. Liver dataset contains 9 volumes of abdominal CT images. Spleen dataset contains 5 volumes of abdominal CT images. All datasets are provided by PLA General Hospital, Shenyang, China. CT images in datasets have a resolution of  $515 \times 512$  pixels with a thickness

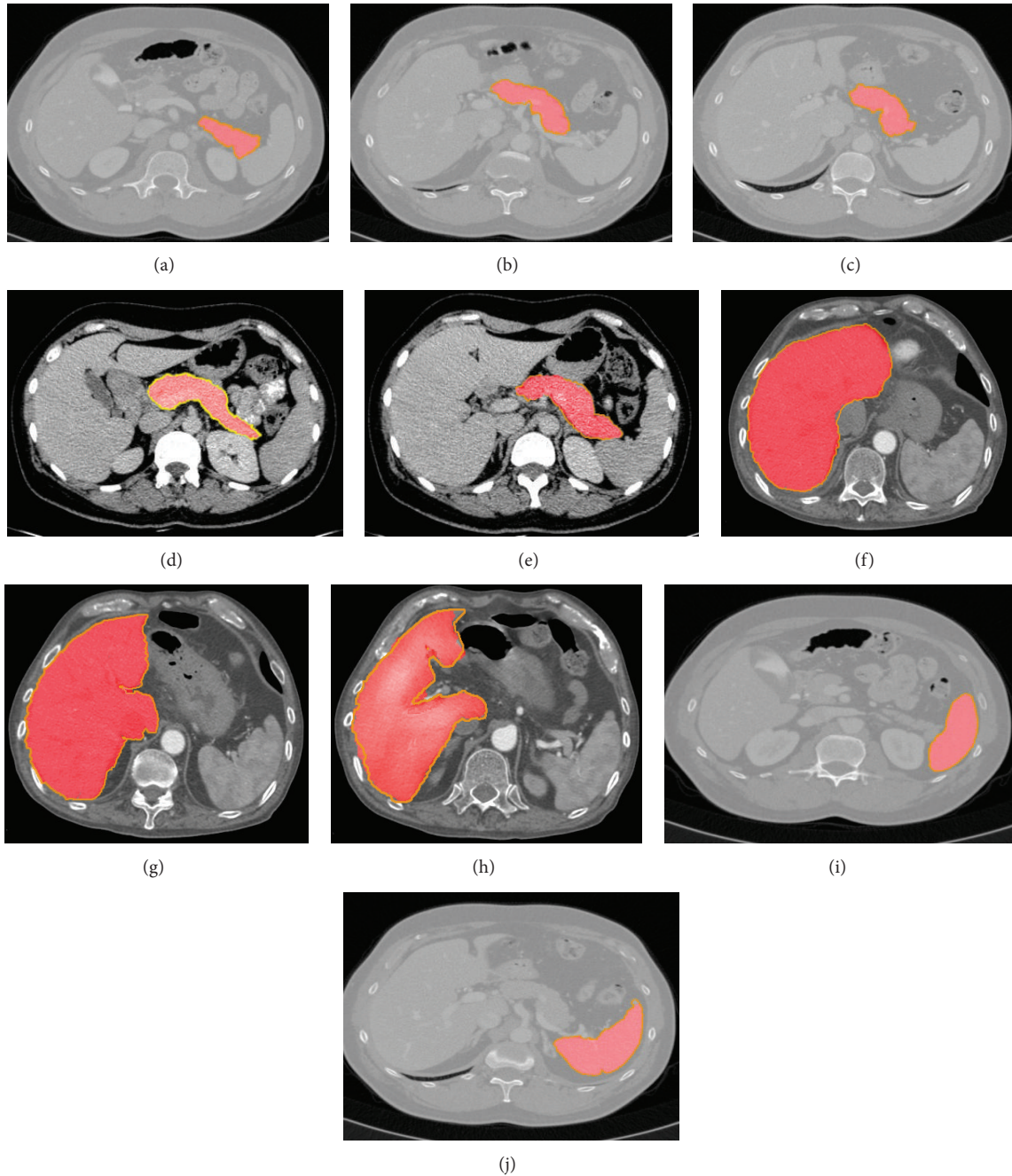


FIGURE 2: Exemplary segmentation results of our proposed method based on pancreas, liver, and spleen datasets. Red regions are segmentation results using proposed method and yellow outline marks the golden standard.

varied between 0.6 mm and 0.7 mm. Each image in the datasets is provided corresponding golden standard manually delineated by experienced radiologists.

**3.3. Segmentation Results and Evaluation.** All the state-of-the-art medical image segmentation methods and the proposed method are applied to extract object region from the CT volume in all the medical image datasets. Average false positive error, false negative error, and similarity index are, respectively, computed for each compared method based on

all segmentation results of all slices. First we calculate false positive error, false negative error, and similarity index for each segmentation results of all methods. Then average values of the three measure standard of each method are calculated based on their respective segmentation results.

Figure 2 shows some examples of segmentation results of our method. The extracted object regions are complete and the edges are smooth.

Figure 3 shows examples of pancreas extraction results based on all evaluated method.

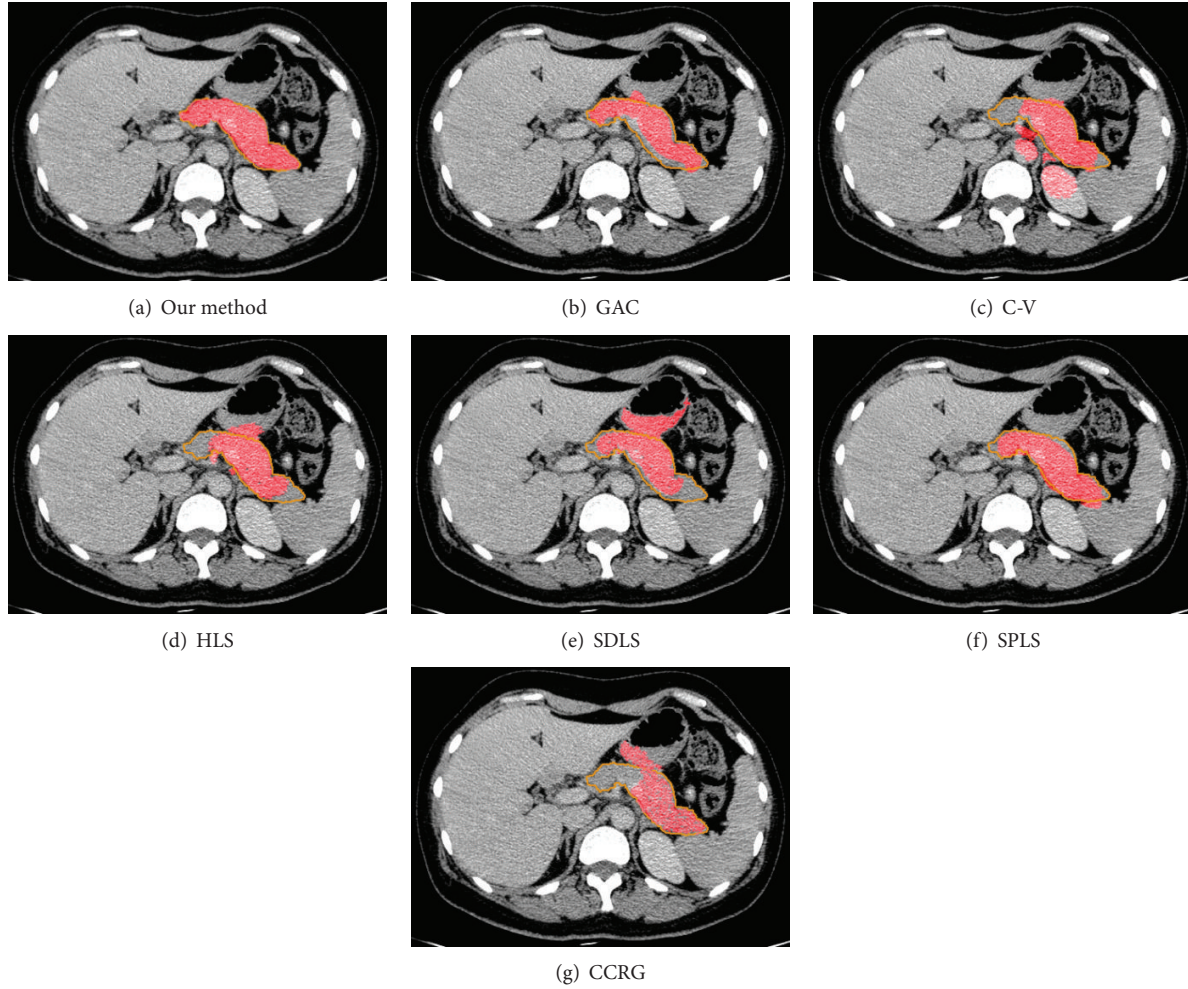


FIGURE 3: The examples of pancreas extraction result based on different methods. (a) Our method. (b) Geodesic active contour method. (c) Shape a priori based level set method. (d) Geodesic active without edge method. (e) Hybrid level set method. (f) Shape detection level set method. (g) Confident connected region growing method.

Figure 4 shows comparison of segmentation results of our proposed method and the improved variational level set method.

Figure 5 shows 3D view of the extracted object organ using our proposed a priori based level set method.

Figures 6, 7, and 8 show histogram of average value of each measure standard for all compared methods. Table 1 contains accurate value of measure standards of all the compared methods. A lower false positive error value means less pixels of background are segmented as object region, and a lower false negative error value means less golden standard of object has not been extracted. Moreover, a higher similarity index means the segmentation results are more accurate. In summary, false positive error and false negative error are lower; the segmentation result is better. Oppositely, similarity index is higher; the segmentation result is better.

Table 2 shows time efficiency of each evaluated method. Table 3 shows trade-off between number of initial manual labelling and algorithm efficiency of proposed method.

TABLE 1: Accurate evaluation value of FPE, FNE, and SI for each method.

Method	FNE	FPE	SI
PBDR	0.093458	0.100255	0.922897
HLS	0.257118	0.408528	0.696948
C-V	0.307937	0.503982	0.669372
SPLS	0.231851	0.201315	0.814745
GAC	0.263718	0.321395	0.744463
SDLS	0.286753	0.353512	0.718136
CCRG	0.495482	1.136335	0.478246
IVLS	0.185473	0.194282	0.8521478

TABLE 2: Quantitative measure of time efficiency for each method.

Method	PBDR	HLS	C-V	SPLS
Time (sec/slice)	0.34	3.66	2.87	1.12
Method	SDLS	GAC	CCRG	IVSL
Time (sec/slice)	0.47	0.51	0.081	0.78

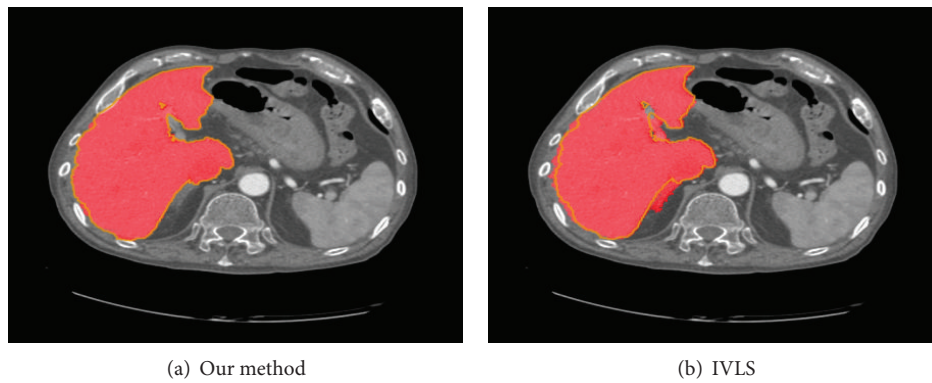


FIGURE 4: Comparison of segmentation results of our proposed method and improved variational level set method.

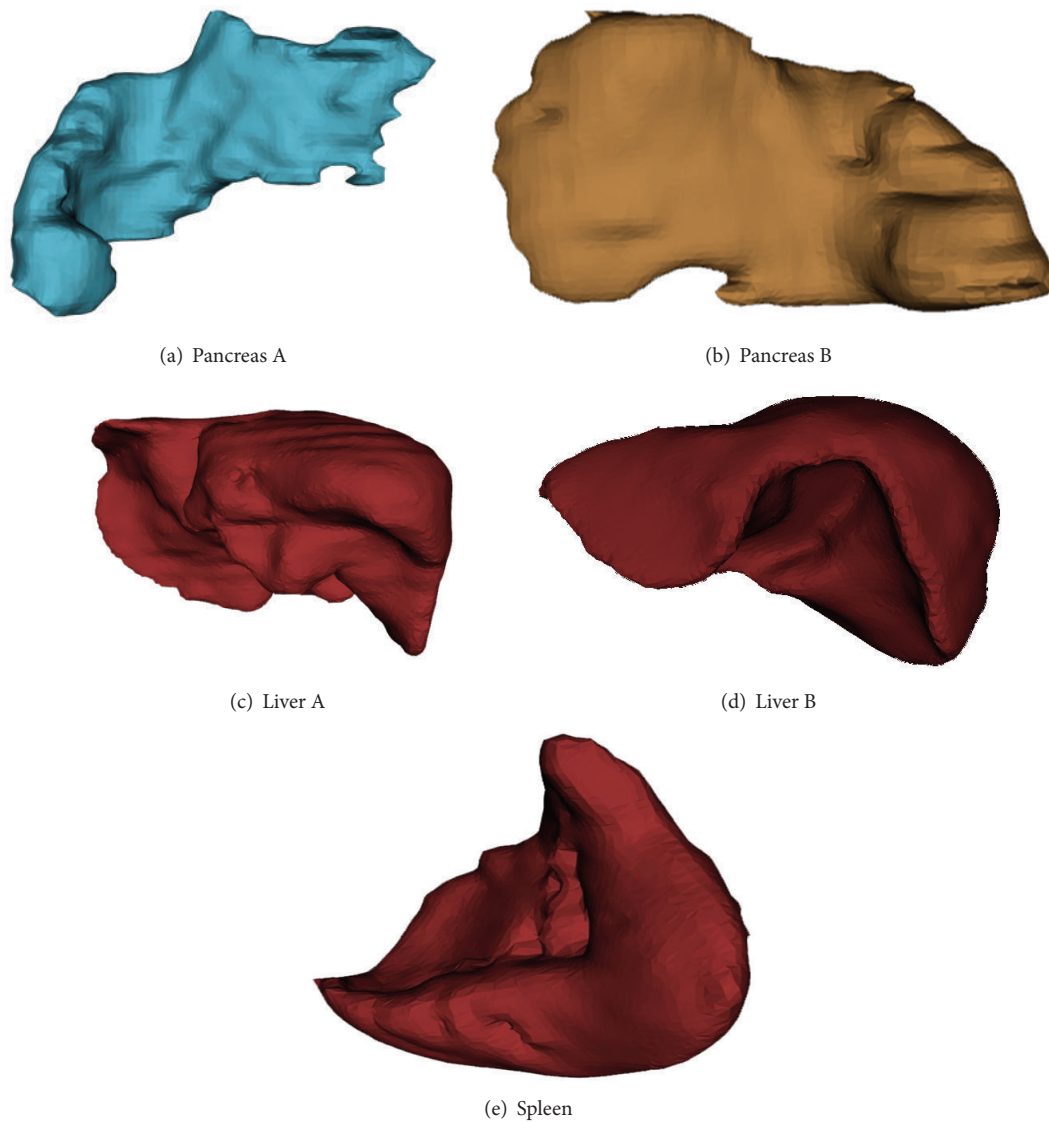


FIGURE 5: 3D view of extracted organs based on our proposed method. (a), (b) 3D view of different pancreas. (c), (d) 3D view of different liver. (e) 3D view of spleen. They are reconstructed using the sequence of segmentation results based on proposed method.



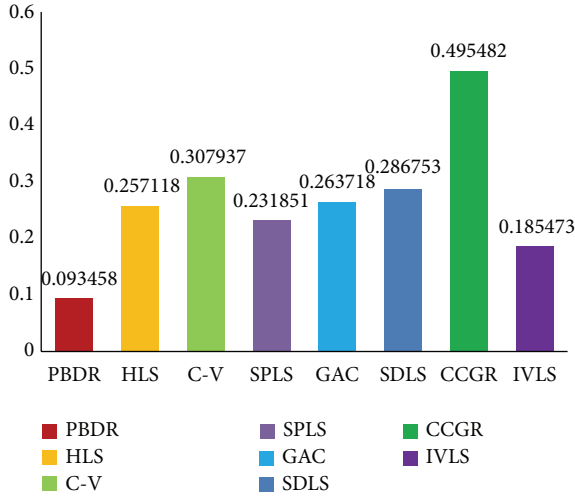


FIGURE 6: False negative error evaluation results of our method (PBDR), hybrid level set method (HLS), C-V method (C-V), shape a priori based level set method (SPLS), geodesic active contour method (GAC), shape detection level set (SDLS), confident connected region growing method (CCRG), and improved variational level set method (IVLS).

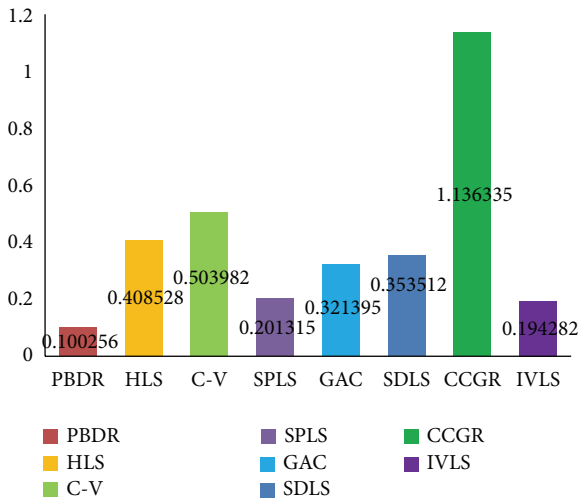


FIGURE 7: False positive error evaluation results of our method (PBDR), hybrid level set method (HLS), C-V method (C-V), shape a priori based level set method (SPLS), geodesic active contour method (GAC), shape detection level set (SDLS) method, confident connected region growing method (CCRG), and improved variational level set method (IVLS).

3.4. Discussion. Evaluated results indicate that the proposed a priori based level set methods (FNE = 0.093, FPE = 0.100, and SI = 0.922) outperform other state-of-art methods in object organ extraction. The a priori based and edge-based level set methods are more suitable for single organ segmentation from a medical image which contains many other organs. The C-V method (FNE = 0.307, FPE = 0.503, and SI = 0.669) abandons edge constraints and intends to achieve global optimal segmentation result, but not local optimal organ.

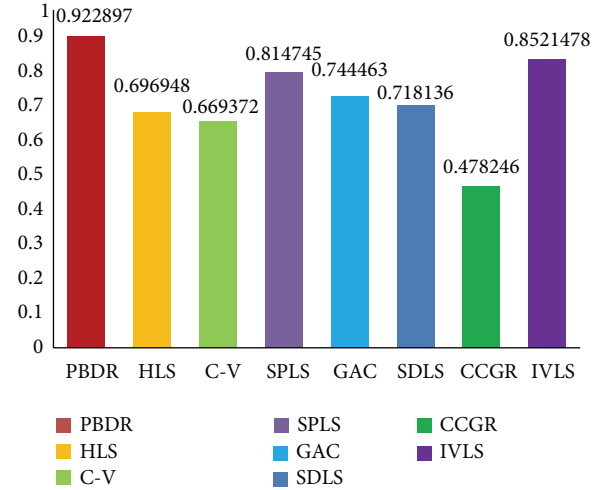


FIGURE 8: Similarity index evaluation results of our method (PBDR), hybrid level set method (HLS), C-V method (C-V), shape a priori based level set method (SPLS), geodesic active contour method (GAC), shape detection level set (SDLS) method, confident connected region growing method (CCRG), and improved variational level set method (IVLS).

TABLE 3: Trade-off between number of initial manual labelling and algorithm efficiency of proposed method in one volume.

Number of labelling	SI
1	0.726
3	0.819
5	0.923
7	0.924
9	0.924
11	0.924
13	0.925
15	0.925

The HLS method (FNE = 0.257, FPE = 0.408, and SI = 0.696) utilizes both edge and region information to segment object. It performs better than C-V method due to the edge constraints. The GAC method (FNE = 0.263, FPE = 0.321, and SI = 0.744) and SDLS method (FNE = 0.286, FPE = 0.201, and SI = 0.718) perform better than region-based level set method, but it is easy to cause oversegmentation at week boundary.

The a priori based level set methods perform better than edge-based level set method; especially our method gets highest accuracy and makes less false segmentation. The SPLS employs a mean statistical shape model to guide the segmentation. But the mean shape cannot adapt to the huge shape variance of object organs, such that leakage problem still exists in results.

The CCRG method and IVLS method both apply statistical feature, average intensity value, and the standard deviation to guide segmentation. In CCRG method, the mean and standard deviation of intensity value are used to define a value range. Neighbor pixels whose intensity values fall inside the range are included in the object region. This rule makes

the neighbor pixels whose intensity is similar with object are easily classified into object region. This causes serious oversegmentation which is difficult to control.

IVLS method uses average intensity value and the standard deviation as a constraint parameter to optimize the evolution of level set. But the statistical information is fixed and not changed through the whole segmentation process; it cannot reflect the gradual change of intensity in image sequence. This method also applied a region growing method to generate an initial object region, but the initial region is not good enough in some cases. This causes error in segmentation.

The proposed method employs a priori statistical feature set and the shape of extracted object in previous slice to guide the segmentation. A probability density map is generated based on feature set. The probability density map is used in energy term of level set evolution function to overcome problem of leakage in segmentation results. New segmented results are added into training set to update the statistical feature. A voting mechanism is used to support the update and it can reduce the effect of singular value to the statistical features. The initial contour which is product based on shape mask of previous slice can satisfy the requirement of locating initial zero level set closed to the final contour. Therefore, our a priori based distance regularization level set method outperforms other evaluated methods in object organs extraction. On the time efficiency comparison, our method is fastest and needs least time to process a slice.

In the time efficiency comparison, among all evaluated level set methods, the proposed method is the fastest ( $0.34 \pm 0.02$  sec/slice). Because the initial zero level set is closed to the final contour and probability density map makes the contour propagate of level set has a high speed. The shape detection level set method costs  $0.47 \pm 0.02$  sec/slice and GAC method costs  $0.51 \pm 0.05$  sec/slice. They both just need to calculate the edge feature, but not depend on region information. C-V method and HLS method need more execution time, because they depend on the global information whose calculation is time consuming.

Evaluation of trade-off between number of initial manual labelling and algorithm efficiency of proposed method indicates that equilibrium exists. Assuming that  $N$  big shape variations exist in a volume, the volume is divided into  $N + 1$  segment. In each segment, the slice in which object organ has a largest cross section is found out and manually labelled. Such that total  $N + 1$  samples are applied to guide the extraction. Under this strategy, good algorithm efficiency can be achieved while the manual labelling is marked as little as possible.

#### 4. Conclusion and Future Work

The proposed method effectively incorporates a priori statistical feature of intensity distribution and a modified distance regularized level set (MDRLS) method to extract object organs from CT image. Our main contribution is coming up with a novel application strategy of a priori knowledge for segmentation and achieving better accuracy and time

efficiency in object organ extraction. Our method needs fewer and simple human-computer interaction.

Based on a priori shape of previous slice, an optimal level set contour is generated for the modified distance regularized level set. A probability density map is employed in MDRLS for further preventing the oversegmentation in object region of nonideal edges. Moreover, the proposed method is simultaneously time efficient due to high speed propagation and less iteration time.

#### Conflict of Interests

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

#### Acknowledgments

The research is supported by the National Natural Science Foundation of China (no. 61272176). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the paper.

#### References

- [1] P. Campadelli and E. Casiraghi, "Liver segmentation from CT scans: a survey," in *Applications of Fuzzy Sets Theory*, vol. 4578 of *Lecture Notes in Computer Science*, pp. 520–528, 2007.
- [2] B. van Ginneken, A. F. Frangi, J. J. Staal, B. M. Ter Haar Romeny, and M. A. Viergever, "Active shape model segmentation with optimal features," *IEEE Transactions on Medical Imaging*, vol. 21, no. 8, pp. 924–933, 2002.
- [3] R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape modeling with front propagation: a level set approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 2, pp. 158–175, 1995.
- [4] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International Journal of Computer Vision*, vol. 22, no. 1, pp. 61–79, 1997.
- [5] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.
- [6] Y. Zhang, B. J. Matuszewski, L. K. Shark, and C. J. Moore, "Medical image segmentation using new hybrid level-set method," in *Proceedings of the IEEE International Conference on Biomedical Visualisation (MEDi08VIS)*, pp. 71–76, London, UK, July 2008.
- [7] M. E. Leventon, W. E. L. Grimson, and O. Faugeras, "Statistical shape influence in geodesic active contours," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, vol. 1, pp. 316–323, June 2000.
- [8] H. Y. Jiang and R. J. Feng, "Image segmentation method research based on improved variational level set and region growth," *Chinese Journal of Electronics*, vol. 40, no. 8, pp. 1659–1664, 2012.
- [9] M. G. Linguraru, J. K. Sandberg, Z. Li, F. Shah, and R. M. Summers, "Automated segmentation and quantification of liver and spleen from CT images using normalized probabilistic atlases and enhancement estimation," *Medical Physics*, vol. 37, no. 2, pp. 771–783, 2010.
- [10] A. H. Foruzan, R. A. Zoroofi, M. Hori, and Y. Sato, "A knowledge-based technique for liver segmentation in CT data,"

*Computerized Medical Imaging and Graphics*, vol. 33, no. 8, pp. 567–587, 2009.

- [11] A. Barthod-Malat, V. Kopylova, G. I. Podoprigora et al., “Development of Multi-compartment Model of the Liver Using Image-based Meshing Software,” in *Proceedings of the 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '12)*, San Diego, Calif, USA, September 2012.
- [12] A. Barthod-Malat, V. Kopylova, G. I. Podoprigora et al., “Development of multi-compartment model of the liver using image-based meshing software,” in *Proceedings of the 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '12)*, San Diego, Calif, USA, September 2012.
- [13] C. M. Li, C. Y. Xu, C. F. Gui, and M. D. Fox, “Distance regularized level set evolution and its application to image segmentation,” *IEEE Transactions on Image Processing*, vol. 19, no. 12, pp. 3243–3254, 2010.
- [14] S. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Springer, New York, NY, USA, 2002.
- [15] H. Zhao, T. Chan, B. Merriman, and S. Osher, “A variational level set approach to multiphase motion,” *Journal of Computational Physics*, vol. 127, no. 1, pp. 179–195, 1996.
- [16] A. Klein, J. Andersson, B. A. Ardekani et al., “Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration,” *NeuroImage*, vol. 46, no. 3, pp. 786–802, 2009.
- [17] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer, “Morphometric analysis of white matter lesions in MR images: method and validation,” *IEEE Transactions on Medical Imaging*, vol. 13, no. 4, pp. 716–724, 1994.

## Research Article

# Local Alignment Tool Based on Hadoop Framework and GPU Architecture

Che-Lun Hung<sup>1</sup> and Guan-Jie Hua<sup>2</sup>

<sup>1</sup> Department of Computer Science and Communication Engineering, Providence University, No. 200, Section 7, Taiwan Boulevard, Shalu District, Taichung 43301, Taiwan

<sup>2</sup> Department of Computer Science and Information Engineering, Providence University, No. 200, Section 7, Taiwan Boulevard, Shalu District, Taichung 43301, Taiwan

Correspondence should be addressed to Che-Lun Hung; [clhung@pu.edu.tw](mailto:clhung@pu.edu.tw)

Received 30 January 2014; Accepted 14 April 2014; Published 14 May 2014

Academic Editor: Chun-Yuan Lin

Copyright © 2014 C.-L. Hung and G.-J. Hua. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the rapid growth of next generation sequencing technologies, such as Slex, more and more data have been discovered and published. To analyze such huge data the computational performance is an important issue. Recently, many tools, such as SOAP, have been implemented on Hadoop and GPU parallel computing architectures. BLASTP is an important tool, implemented on GPU architectures, for biologists to compare protein sequences. To deal with the big biology data, it is hard to rely on single GPU. Therefore, we implement a distributed BLASTP by combining Hadoop and multi-GPUs. The experimental results present that the proposed method can improve the performance of BLASTP on single GPU, and also it can achieve high availability and fault tolerance.

## 1. Introduction

In the past decade, the sequencing technologies have been improved dramatically. An entirely new technology was developed, next generation sequencing (NGS), a fundamentally different approach to sequencing DNA and RNA much more cheaply and quickly than traditional Sanger sequencing. Meanwhile, NGS is well known as a high-throughput sequencing technology. The number of output data produced by NGS data has increased more than double each year since it was invented. In 2007, a single sequencing run could produce around one gigabase (Gb) of sequence data. By 2011, it approaches a terabase (Tb) of data produced in a single sequencing run—nearly a 1000× increase in four years. With the ability to rapidly generate large amount of sequencing data, NGS has enabled the researches in the field of biology and other closely related fields can be done at a large-scale level and also can move quickly from an idea to full data sets in a matter of hours or days [1]. As NGS becomes key player in modern biological research, the analysis of the vast amount of produced data is not an easy task and a great challenge in the

field of bioinformatics. Therefore, efficient tools to cope with these big data to provide the knowledge easier and faster are essential.

With the rapid development of multicore hardware, graphics processing units (GPUs) are being used in numerous applications to enhance computational performance. GPUs have a low design cost and the increased programmability of GPUs allows them to be more flexible than FPGAs. General-purpose graphics processing units (GPGPU) programming has been successfully used in scientific computing domains, which involve a high level of numeric computation. The greatest benefit of GPUs is that the number of processing units is immense compared to those of CPUs (CPU, approximately 2–16; GPU, approximately 128–512). In 2006, Nvidia proposed the compute unified device architecture (CUDA). CUDA uses a new computing architecture named single instruction multiple threads (SIMT). This architecture allows threads to execute independent and divergent instruction streams, thus facilitating decision-based execution, which is not provided by the more common single instruction multiple data (SIMD). Many well-known tools have been

reimplemented based on GPU architecture [2–4]. One of the wild-use alignment tools, BLASTP, is a heuristic algorithm to produce a local alignment for protein. BLASTP has three implementations on GPU, GPU-NCBI-BLASTP [5], CUDA-BLASTP [6], and GPU-BLASTP [7]. All three implementations achieve 4x~40x speedup over a single-thread CPU implementation of NCBI-BLAST.

Meanwhile, the software architectures of distribution computing have been developed rapidly as well. The cloud computing as a new distribution computing service concept has become popular for providing services with availability, reliability, and on-demand computation to users. The cloud computing environment can be a distributed system that has massively scalable IT-related capabilities, providing multiple external customers many services on Internet. In addition, cloud computing can be used to copy with big data and maintain high availability and fault tolerance. Hadoop [8] is one of the commonly used open source software frameworks intended to support data-intensive distributed applications. Hadoop adopts Map/Reduce programming model to process petabytes of data with thousands of nodes. Map/Reduce programming model is useful to develop parallel computing applications on cloud computing environment. In Map/Reduce model, mappers and reducers complete a task. Each mapper performs a map operation and each map operation is independent of the others. A task is split into many subtasks, and each mapper processes its subtask. Similarly, a set of reducers can perform a set of reduce operations. Reducers deal with the data produced by mappers. An important benefit of using Hadoop to develop the application is fault tolerance. Hadoop can guide jobs toward a successful completion even when individual nodes experience failure in computation. In these situations, Hadoop platform is considered as a much better solution for these real-world applications. Currently, Hadoop has been applied in various domains in bioinformatics [9–13]. Cloud-PLBS [14] is a cloud service that combines the SMAP [15–17] and Hadoop frameworks for 3D ligand binding site comparison and similarity searching of a structural proteome. This platform is computationally more efficient than standard SMAP. Hung and Lin [12] proposed a parallel protein structure alignment service based on the Hadoop distribution framework. This service includes a protein structure alignment algorithm, a refinement algorithm, and a Map/Reduce programming model. The computational performance of their service is proportional to the number of processors used in their cloud platform.

In this paper, we combine these two different heterogeneous architectures, software architecture-Hadoop framework and hardware architecture-GPU, to develop a high performance cloud computing service for protein sequence alignment. In this cloud service, each mapper performs BLASTP and a reducer collects all resulting alignments produced by mappers. The mappers work simultaneously. By using Hadoop, the proposed GPU based bioinformatics cloud service can recover the comparison job from a crashed GPU host by assigning this job to other health GPU hosts. This cloud platform can achieve high performance, scalability, and availability. The experimental results present that the

computational performance of the proposed service can be enhanced by using Hadoop and GPU architecture.

## 2. Method

In the work, we integrate BLASTP with Hadoop. Hadoop framework works with mappers and reducer. Mappers perform BLASTP on GPU, and reducer collects all alignment results produced by mappers. Despite Hadoop distribution computing framework, performance of BLASTP can be enhanced by multiple mappers. Hadoop guarantees that all of BLASTP computational jobs on mappers can be completed, even if some of the mappers stop.

*2.1. GPU Programming.* As the GPU has become increasingly more powerful and ubiquitous, researchers have begun developing various nongraphics or general-purpose applications [18]. Traditionally, the GPUs are organized in a streaming, data-parallel model in which the coprocessors execute the same instructions on multiple data streams simultaneously. Modern GPUs include several (tens to hundreds) of each type of stream processor; both of graphical and general-purpose applications thus are faced with parallelization challenges [19].

Nvidia released the compute unified device architecture (CUDA) SDK to assist developers in creating nongraphics applications that run on GPUs. CUDA programs typically consist of a component that runs on the CPU, or host, and a smaller but computationally intensive component called the kernel that runs in parallel on the GPU. Input data for the kernel must be copied to the GPU's on-board memory from CPU's main memory through the PCI-E bus prior to invoking the kernel, and output data also should be written to the GPU's memory first. All memory used by the kernel should be preallocated.

Kernel executes a collection of threads that computes a result for a small segment of data. To manage multiple threads, kernel is partitioned into thread blocks, with each thread block being limited to a maximum of 512 threads. The thread blocks are usually positioned within a one- or two-dimensional grid. Each thread can be positioned within a given block where it belongs, and this given block can be positioned within the grid. Therefore, each thread can calculate which elements of data to operate on and which regions of memory to write output to by an algebraic formula. Each block is executed by a single multiprocessor, which allows all threads within the block to communicate through on-chip shared memory. The parallelism architecture of GPGPU is illustrated in Figure 1.

*2.2. Hadoop Framework.* Hadoop is a software framework to copy with distributed data in parallel by communicating computing nodes. Hadoop runs data-intensive applications through the Map/Reduce parallel processing technique. This framework has been used in many cloud industry companies, such as Yahoo, Amazon EC2, IBM, and Google. The example of computation of Map/Reduce framework is illustrated in Figure 2. In the mapper stage, the input data is split into

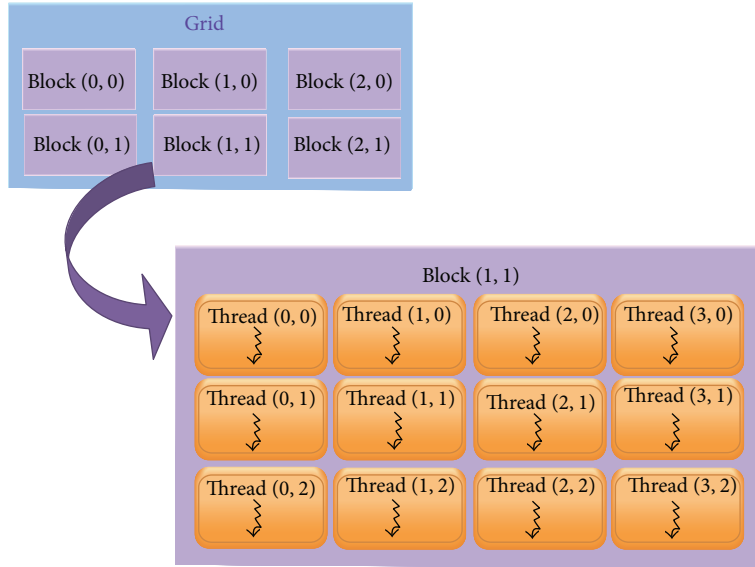


FIGURE 1: The parallelism architecture of GPGPU.

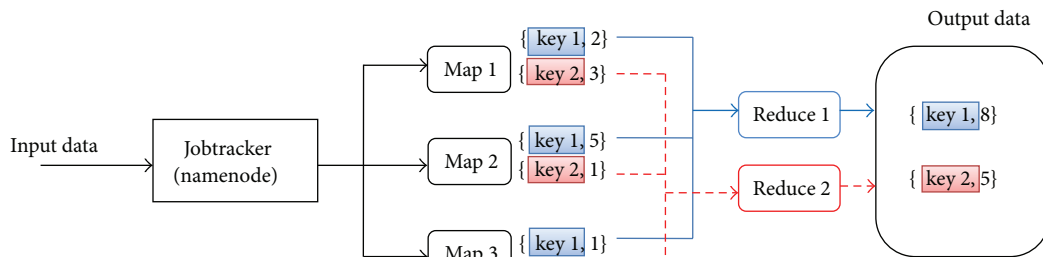


FIGURE 2: Computation of Map/Reduce framework of Hadoop.

smaller chunks corresponding to the number of mappers, and each mapper performs the operation on the data chunk. Output of each mapper has the format of  $\langle \text{key}, \text{value} \rangle$  pairs. Outputs from all mappers,  $\langle \text{key}, \text{value} \rangle$  pairs, are classified by key before being distributed to reducer. Reducer adds values by the same key. Outputs of reducers are  $\langle \text{key}, \text{value} \rangle$  pairs where each key is unique.

Hadoop cluster consists of a single master and multiple slave nodes. The role of the master node is a jobtracker, tasktracker, namenode, and datanode. A slave node, as computing node, is a datanode and tasktracker. The jobtracker is the service within Hadoop that manages Map/Reduce tasks that can be completed on computing nodes in the cluster, the nodes that already have the data. A tasktracker is a node in the cluster that accepts tasks and maps, reduces, and shuffles operations from a jobtracker. The architecture of Hadoop cluster is shown in Figure 3.

Hadoop distributed file system (HDFS) is the distribution file system used by Hadoop framework in default. Each input data file is split into data blocks that are distributed on datanodes by HDFS. HDFS can create multiple replicas of data blocks and distributes them on datanodes usually in the same rack as the source datanode throughout a cluster to enable

reliable, extremely rapid computations. The namenode serves as both a directory namespace manager and a node metadata manager for the HDFS. There is a single namenode running in the HDFS architecture. The architecture of HDFS is shown in Figure 3.

2.3. *BLASTP*. The basic local alignment search tool (BLAST) [20], as it is commonly referred to, is a database search tool, developed and maintained by the National Center for Biotechnology Information (NCBI). The web-based tool for BLAST searches is available at <http://blast.ncbi.nlm.nih.gov/Blast.cgi>.

The BLAST suite of programs has been designed to find high scoring local alignments between sequences, without compromising the speed of such searches. BLAST uses a heuristic algorithm which seeks local as opposed to global alignments and is therefore able to detect relationships among sequences which share only isolated regions of similarity [20]. The first version of BLAST was released in 1990 and allowed users to perform ungapped searches only. The second version of BLAST, released in 1997, allowed gapped searches [21]. BLASTP is used for both identifying a query amino acid sequence and finding similar sequences in protein databases.

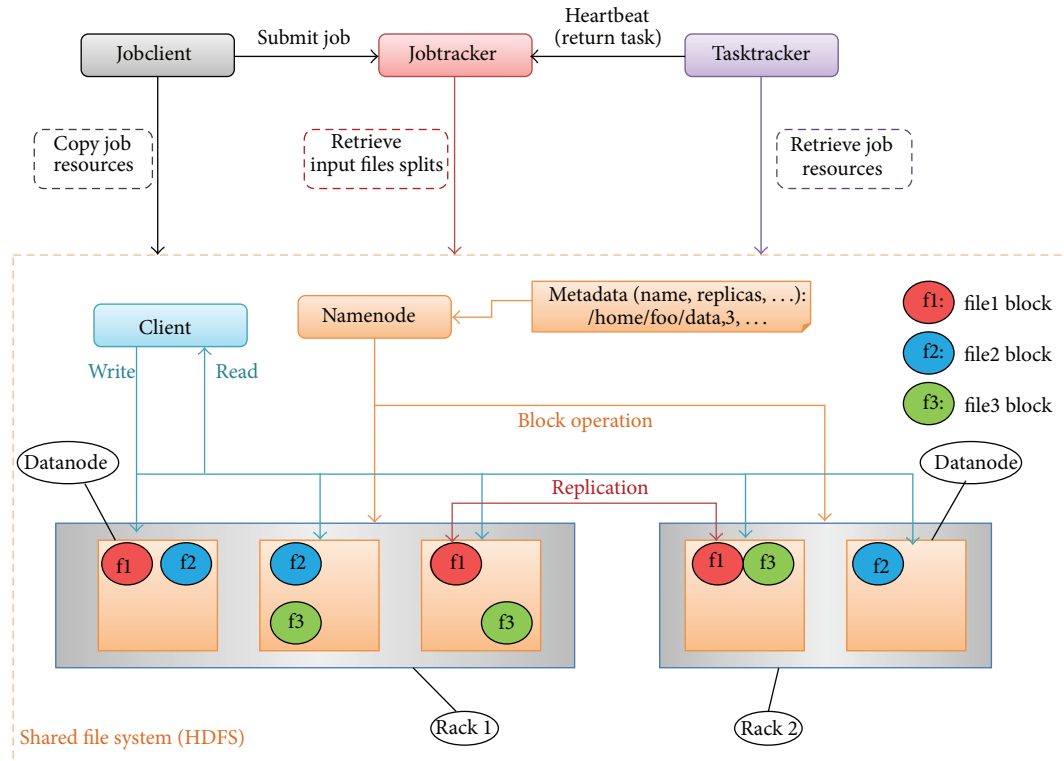


FIGURE 3: The architecture of Hadoop cluster and HDFS.

BLASTP has three implementations on GPU, GPU-NCBI-BLAST, CUDA-BLASTP, and GPU-BLASTP. All three implementations achieve 4x~40x speedup over a single-thread CPU implementation of NCBI-BLAST.

**2.4. Cloud-BLASTP.** To enhance the performance of CUDA-BLASTP on single GPU is to scale with multiple GPUs. In the proposed distributed GPU system, we utilized Hadoop framework to manage multiple GPUs. The Cloud-BLASTP architecture is demonstrated in Figure 4. Each single GPU server has a GPU card. To derive these distributed GPU cards, Hadoop is suitable for managing these cards. Every mapper in a node performs BLASTP and a reducer collects all the results produced by mappers. In this architecture, the sequence database is separated into several parts and uploaded to servers by HDFS. The features of Hadoop BLASTP are high performance, availability and reliability, and scalability.

**2.4.1. High Performance.** In Hadoop BLASTP, the BLASTP jobs are performed in parallel by Map/Reduce framework. The number of the BLASTP jobs can be performed simultaneously which is the same as the number of the mappers. If the number of the BLASTP jobs is greater than the number of the mappers, then the number of mappers will assign the rest of unperformed BLASTP jobs to available mappers immediately.

**2.4.2. Availability and Reliability.** Hadoop BLASTP is able to avoid the BLASTP jobs stop when mappers are down. By

using Hadoop fault tolerance mechanism, when a datanode (mapper) is down during BLASTP computation, its BLASTP job will be reassigned to another slave node (mapper) by namenode. Therefore, all of the submitted BLASTP jobs never stop because one of the datanodes fails in Hadoop BLASTP. A hardware failure on the physical server causes a disastrous failure as all mappers running on it die. One way is that all of these jobs can be reassigned, and another way is that several new mappers are created on available hosts and then these jobs are reassigned to these new mappers. Thus, Hadoop BLASTP has high availability.

**2.4.3. Scalability.** By Hadoop framework, Hadoop BLASTP can create new slave mappers as datanodes according to the number of submitted BLASTP jobs. When large amounts of the BLASTP jobs are submitted, Hadoop BLASTP can create more mappers to copy with more BLASTP jobs to enhance the performance.

### 3. Cloud-BLASTP Platform

Cloud-BLASTP is a protein alignment cloud service under Hadoop framework, BLASTP, and GPU architecture. The cloud computing platform is composed of one NFS server and 4 GPU servers in the Providence University Cloud Computation Laboratory. Each server is equipped with an Intel i7 3930 3.2 GHz CPU, 16 G RAM, and Nvidia GeForceGTS 480 graphics card (Fermi architecture). Each server is running under the O.S. Ubuntu version 10.4 with Hadoop version 0.2

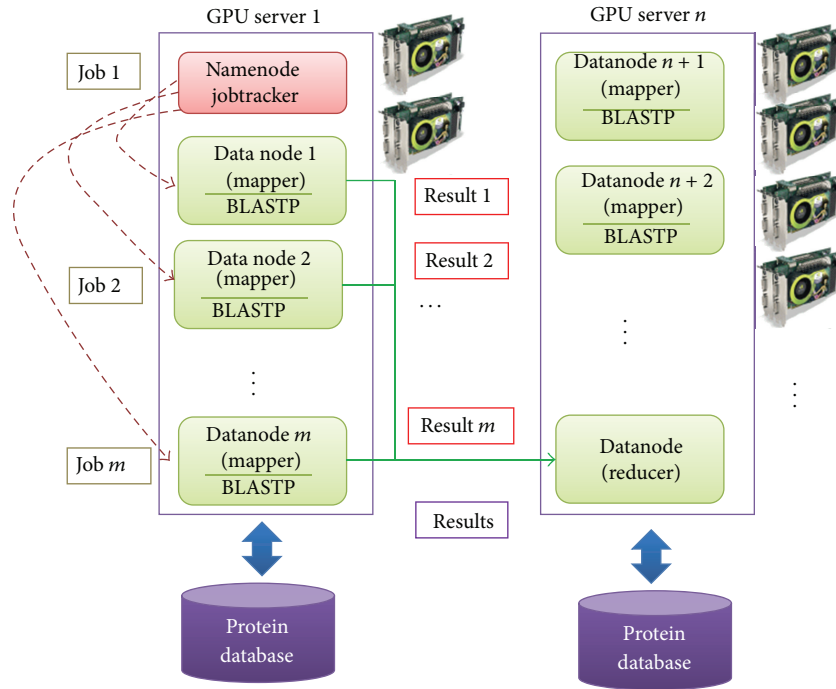


FIGURE 4: The architecture of Cloud-BLASTP.

Map/Reduce framework. Each server is responsible for a map operation and a reduce operation. The total number of the Map/Reduce operations is up to 4, respectively.

#### 4. Performance Evaluation

To assess the performance of the proposed cloud service, we compared the execution time between stand-alone BLASTP and Cloud-BLASTP. The performance of both programs depends upon the amount of data set and the number of computing mappers. Therefore, the performance between the programs is tested with respect to these two factors. In the first experiment, the data size of protein database is 841 MB, and the numbers of query sequences are 102, 204, and 408. The number of query sequences processed by each mapper is the number of query sequences divided by the number of mappers. For example, suppose there are two mappers, and mapper 1 has to process 26 sequences and mapper 2 has to process 25 sequences. The results are shown in Figure 5. As shown in the figure, the execution time of comparing 102 sequences can be reduced from 318 seconds (consumed by the single GPU-BLASTP) to 187 seconds and 88 seconds by executing Cloud-BLASTP with 2 and 4 mappers, respectively. Also, the execution time of comparing 204 sequences can be reduced from 622 seconds to 318 seconds and 164 seconds by executing Cloud-BLASTP with 2 and 4 mappers, respectively. For querying 408 sequences, the execution time can be reduced from 1236 seconds to 622 seconds and 318 sequences by executing Cloud-BLASTP with 2 and 4 mappers, respectively. It is obvious that with less mappers (GPU servers) the performance is much worse. Clearly, the execution time is effectively reduced when more

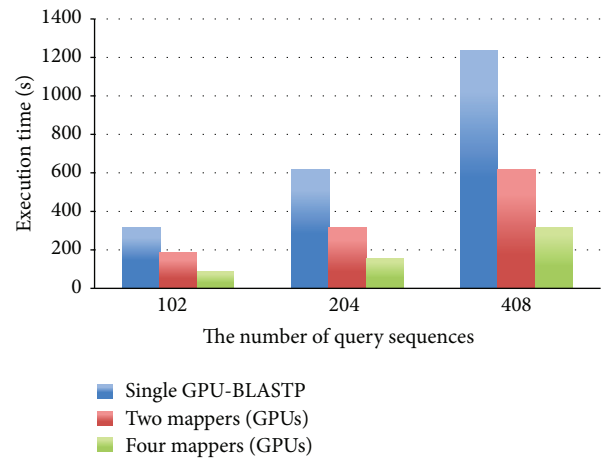


FIGURE 5: The performance of Cloud-BLASTP in the various numbers of mappers.

mappers are involved. In general, more mappers achieve a faster processing speed.

In Cloud-BLASTP, the important features are reliability and availability. The computing process at the failed node is able to continue at another node that has the replica of data of the failed node. Therefore, we performed a simulation to evaluate the reliability and availability of the proposed cloud service when mappers fail. In this simulation, we make half of the mappers fail in the duration of executing BLASTP. In this simulation, the heartbeat time is set to one minute, and the number of replicas is set to three as default. Therefore, all of jobs can be completed even when some of the nodes fail. Figures 6(a) and 6(b) demonstrate the



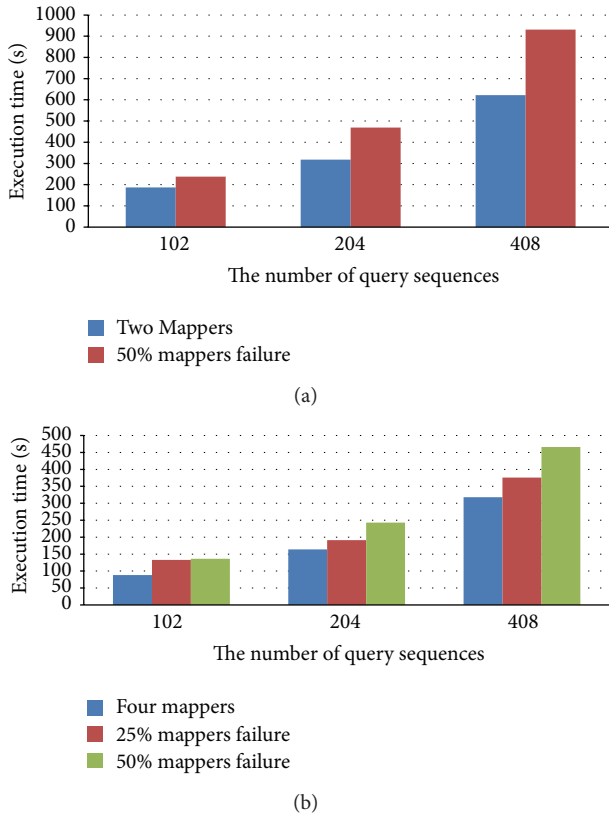


FIGURE 6: Execution time of node failure at half of execution duration of Cloud-BLASTP. (a) Two mappers; (b) four mappers.

performance of the proposed method meeting corresponding half of mappers fail and quarter of mappers fail for querying 102, 204 and 408 sequences when failures happen at duration of 50% execution, respectively. The execution time with no failure is shown as the blue bar, and the execution time with failure in a half of mappers is shown as red bar. From the experiment results, it shows that the jobs can be completed when mappers fail, but the execution time is more than normal execution time because the failed jobs have to be assigned to other health mappers. Figures 7(a) and 7(b) demonstrate the performance when the failures happen at the duration of 25% execution. Although the mappers fail, the execution time of redundancy is related to the number of mappers too. Thereby, Cloud-BLASTP is mapper failure-free.

## 5. Conclusion

In the past few years, sequencing technologies have grown rapidly. The amount of produced sequence data is from gigabase increased to terabase, and the duration is from months decreased to days. Therefore, the performance of the bioinformatics tools is important to analyze data efficiently. Sequence alignment is the basic and common analysis step for biologists to practice further experiment. BLASTP is one of the wild-used local alignment tools for protein sequences. It is now provided on NCBI organization. BLASTP has also been implemented on GPU to enhance the alignment performance. Although BLASTP outperforms most existing

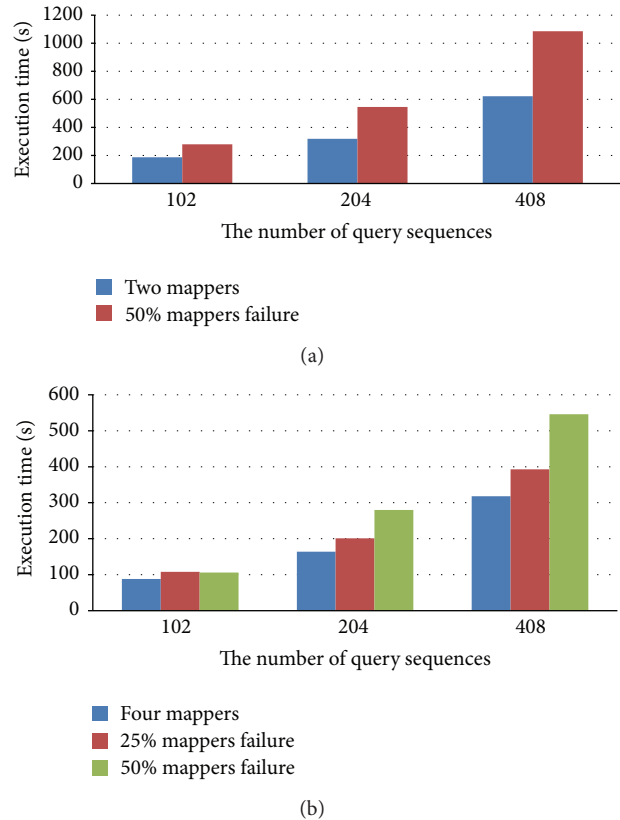


FIGURE 7: Execution time of node failure at 25% of execution duration of Cloud-BLASTP. (a) Two mappers; (b) four mappers.

local sequence alignment tools, it does not satisfy the need of high scalability and high availability for searching huge protein database.

Hadoop framework has become popular for providing efficient and available distributed computation to users. In this paper, we propose a cloud computing tool, called Cloud-BLASTP, for protein local alignment by integrating Hadoop framework and BLASTP tool. Cloud-BLASTP takes advantage of high performance, availability, reliability, and scalability. Cloud-BLASTP guarantees that all submitted jobs are properly completed, even when running job on an individual node or mapper experience failure. The performance experiment shows that it is desirable for biologists to investigate the protein structure and function analysis by comparing large protein database under reasonable time constraints.

## Conflict of Interests

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

## Acknowledgments

This research was partially supported by the National Science Council under Grants 100-2221-E-126-007-MY3. Yu-Lin Tsai and Hsiu-Ping Hou are appreciated for their assistance in system construction and experiment analysis.

## References

- [1] E. R. Mardis, "Next-generation DNA sequencing methods," *Annual Review of Genomics and Human Genetics*, vol. 9, pp. 387–402, 2008.
- [2] S. T. Lee, C. Y. Lin, and C. L. Hung, "GPU-based cloud service for smith-waterman algorithm using frequency distance filtration scheme," *Biomed Research International*, vol. 2013, Article ID 72173, 8 pages, 2013.
- [3] C. Y. Lin, C. L. Hung, and Y. C. Hu, "A Re-sequencing Tool for High-throughput Long Reads Based on UNImarker with non-Overlapping iNterval indexing strategy," *Information—an International Interdisciplinary Journal*, vol. 16, no. 1, pp. 827–832, 2013.
- [4] R. Li, C. Yu, Y. Li et al., "SOAP2: an improved ultrafast tool for short read alignment," *Bioinformatics*, vol. 25, no. 15, pp. 1966–1967, 2009.
- [5] P. D. Vouzis and N. V. Sahinidis, "GPU-BLAST: using graphics processors to accelerate protein sequence alignment," *Bioinformatics*, vol. 27, no. 2, Article ID btq644, pp. 182–188, 2011.
- [6] W. Liu, B. Schmidt, and W. Müller-Wittig, "CUDA-BLASTP: accelerating BLASTP on CUDA-enabled graphics hardware," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 6, pp. 1678–1684, 2011.
- [7] S. Xiao, H. Lin, and W. C. Feng, "Accelerating protein sequence search in a heterogeneous computing system," in *Proceedings of the 25th IEEE International Parallel and Distributed Processing Symposium (IPDPS '11)*, pp. 1212–1222, May 2011.
- [8] Hadoop—Apache Software Foundation project home page, <http://hadoop.apache.org/>.
- [9] R. C. Taylor, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics," *BMC Bioinformatics*, vol. 11, no. 12, article S1, 2010.
- [10] M. C. Schatz, "CloudBurst: highly sensitive read mapping with MapReduce," *Bioinformatics*, vol. 25, no. 11, pp. 1363–1369, 2009.
- [11] B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg, "Searching for SNPs with cloud computing," *Genome Biology*, vol. 10, no. 11, article R134, 2009.
- [12] C. L. Hung and Y. L. Lin, "Implementation of a parallel protein structure alignment service on cloud," *International Journal of Genomics*, vol. 2013, Article ID 439681, 8 pages, 2013.
- [13] C. L. Hung and C. Y. Lin, "Open reading frame phylogenetic analysis on the cloud," *International Journal of Genomics*, vol. 2013, Article ID 614923, 9 pages, 2013.
- [14] C. L. Hung and G. J. Hua, "Cloud computing for protein-ligand binding site comparison," *Biomed Research International*, vol. 2013, Article ID 170356, 7 pages, 2013.
- [15] L. Xie and P. E. Bourne, "A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites," *BMC Bioinformatics*, vol. 8, no. 4, article S9, 2007.
- [16] L. Xie and P. E. Bourne, "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 14, pp. 5441–5446, 2008.
- [17] L. Xie, L. Xie, and P. E. Bourne, "A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery," *Bioinformatics*, vol. 25, no. 12, pp. i305–i312, 2009.
- [18] J. D. Owens, D. Luebke, N. Govindaraju et al., "A survey of general-purpose computation on graphics hardware," *Computer Graphics Forum*, vol. 26, no. 1, pp. 80–113, 2007.
- [19] N. K. Govindaraju, S. Larsen, J. Gray, and D. Manocha, "A memory model for scientific algorithms on graphics processors," in *Proceedings of the ACM/IEEE Supercomputing Conference (SC '06)*, p. 6, 2006.
- [20] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.
- [21] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

## Research Article

# Double-Bottom Chaotic Map Particle Swarm Optimization Based on Chi-Square Test to Determine Gene-Gene Interactions

Cheng-Hong Yang,<sup>1</sup> Yu-Da Lin,<sup>1</sup> Li-Yeh Chuang,<sup>2</sup> and Hsueh-Wei Chang<sup>3,4</sup>

<sup>1</sup> Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung 80778, Taiwan

<sup>2</sup> Department of Chemical Engineering and Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung 84001, Taiwan

<sup>3</sup> Institute of Medical Science and Technology, National Sun Yat-sen University, Kaohsiung, 80708, Taiwan

<sup>4</sup> Department of Biomedical Science and Environmental Biology, Translational Research Center, Cancer Center, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung, Taiwan

Correspondence should be addressed to Li-Yeh Chuang; [chuang@isu.edu.tw](mailto:chuang@isu.edu.tw) and Hsueh-Wei Chang; [changhw@kmu.edu.tw](mailto:changhw@kmu.edu.tw)

Received 20 October 2013; Accepted 16 April 2014; Published 7 May 2014

Academic Editor: Huiru Zheng

Copyright © 2014 Cheng-Hong Yang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Gene-gene interaction studies focus on the investigation of the association between the single nucleotide polymorphisms (SNPs) of genes for disease susceptibility. Statistical methods are widely used to search for a good model of gene-gene interaction for disease analysis, and the previously determined models have successfully explained the effects between SNPs and diseases. However, the huge numbers of potential combinations of SNP genotypes limit the use of statistical methods for analysing high-order interaction, and finding an available high-order model of gene-gene interaction remains a challenge. In this study, an improved particle swarm optimization with double-bottom chaotic maps (DBM-PSO) was applied to assist statistical methods in the analysis of associated variations to disease susceptibility. A big data set was simulated using the published genotype frequencies of 26 SNPs amongst eight genes for breast cancer. Results showed that the proposed DBM-PSO successfully determined two- to six-order models of gene-gene interaction for the risk association with breast cancer (odds ratio > 1.0;  $P$  value < 0.05). Analysis results supported that the proposed DBM-PSO can identify good models and provide higher chi-square values than conventional PSO. This study indicates that DBM-PSO is a robust and precise algorithm for determination of gene-gene interaction models for breast cancer.

## 1. Introduction

Genome-wide association studies (GWAS) for the analysis of gene-gene interaction are important fields for detecting the effects of cancer and disease [1–4]. Such studies usually entail the collection of a vast number of samples and SNPs selected from several related genes of disease in order to identify the association amongst genes. Disease effect, in general, is influenced by the best association between SNPs from several genes; these SNPs could have a potential association to provide information for disease analysis. Therefore, a method for searching high-order interactions is needed to determine the potential association between several loci.

Good models of the association between SNPs from several genes are usually hidden in the large number of

possible models. The sum of all possible models of association between case data and control data can be computed by  $C(n, m) \times g^m$ , where  $n$  represents a total number of SNPs,  $m$  is a selected number of SNPs, and  $g$  is the number of genotypes. Data mining and machine learning methods have been proposed for use in GWAS data analysis. These computational approaches were developed to examine epistasis in family-based and case-control association studies [5–12]. The genetic algorithm (GA), particle swarm optimization (PSO), and chaotic particle swarm (CPSO) methods were proposed to identify the models of gene-gene interaction. However, the ability to determine the relative model quality needs to be improved. In mathematics, the problem space for identifying good models is not linear and the algorithm converges easily to a local optima, since no better models are

found near the best model in that region. PSO often leads to premature convergence, especially in complex multipeak search problems. Therefore, the use of chaotic sequences to improve the PSO has been proposed to identify models of gene-gene interaction [7]. An improved PSO using a double-bottom chaotic maps (DBM-PSO) [13] has been shown to overcome the respective disadvantages of PSO and CPSO. In this study, DBM-PSO is applied to assist statistical methods in the analysis of associated variations to disease susceptibility.

A total of 26 SNPs obtained from eight related genes of breast cancer (EGF, IGF1, IGF1R, IGF2, IGF1BP3, IL10, TGFBI, and VEGF) were used to test the various methods for comparison of the association models. It is proposed that the interactions between polymorphisms of breast cancer-related genes may have synergistic effects on the pathogenesis of cancer and disease; this would explain differences in disease susceptibility. The quality of a model of gene-gene interaction can be assessed by determining its odds ratio (OR), confidence intervals, and  $P$  value. We systematically evaluate the model effects from two- to five-order interactions to compare the DBM-PSO with other PSOs methods.

## 2. Methods

**2.1. Problem Description.** To identify the quality of the models of gene-gene interaction problem, the model includes SNPs and their corresponding genotypes. The set  $X = \{x_1, x_2, x_3, \dots, x_D\}$  represents a possible model as a solution in the problem space; each parameter  $x$  is a real number. The chi-square test is used to design the PSO and DBM-PSO fitness functions. The objective is to search for a vector  $X^*$  which has its own best fitness value according to the evaluation of fitness function  $f(X)$  ( $f : \delta \subseteq R^D \rightarrow R$ ); that is,  $f(X^*) > f(X)$ , for all  $X \in \delta$ , where  $\delta$  is a nonempty large finite set serving as the search space and  $\delta = R^D$ .

**2.2. Particle Swarm Optimization.** Particle swarm optimization (PSO) is a population-based stochastic optimization technique [14]. The conception of PSO is based on a robust theory of swarm intelligence to search for an optimal resolution of complex problems. Swarm intelligence describes an automatically evolving system based on simulating the social behaviour of organisms, for example, knowledge sharing. Therefore, valuable information can be shared amongst swarm members to suggest a common objective which leads individuals toward an optimal direction. PSO has been used to solve several types of optimization problems [15], including function optimization and parameter optimization [16] and shows promise for nonlinear function optimization [17–22]. In PSO, possible solutions are represented as the particles. During generation, particle positions are adjusted according to the updated velocity toward a significant objective. The objective of each particle is defined based on the particle's previous experience ( $pbest$ ) and knowledge commonly held by the population ( $gbest$ ). Thus, particles can effectively converge into a solution-rich area to find the better solution. Finally, the particles follow the current best particle in the search space until a predefined number of generations are

reached. The PSO procedure entails (1) population initialization, (2) objective function evaluation, (3) identification of  $pbest$  and  $gbest$ , (4) particle updating, and (5) the termination condition. These steps are described in detail in the following section.

**2.3. Double-Bottom Map Particle Swarm Optimization.** Double-bottom map particle swarm optimization (DBM-PSO) was proposed by Yang et al. in 2012 [13]. While PSO is easily complicated by the existence of nonlinear fitness function with multiple local optima, this is not an issue for DBM-PSO. A local optima,  $f_i = f(X_i)$ , can be described as  $\exists \epsilon > 0 \forall X \in \delta : \|X - X_i\| < \epsilon \Rightarrow f(X) \leq f_i \leq f(X^*)$ , where  $\|\cdot\|$  represents any  $p$ -norm distance measure. In PSO, the flexibilities of given constraints and vector space in the problem influence the determination of the best solution. Generally speaking,  $r_1$  and  $r_2$  independently influence search exploitation and exploration, and the effect of  $r_1$  and  $r_2$  on the convergence behaviour is very important in PSO. Recently, chaos approaches have been proposed to overcome the inherent disadvantages of PSO. Chaotic maps are easily applied in PSO to prevent entrapment of the population in a local optima [23]. DBM-PSO proposes a new type of chaotic map, called double-bottom maps, to improve the search ability of PSO. Double-bottom maps are used to design an updating function to balance the exploration and exploitation for PSO search capability. The superiority of the double-bottom map over other chaotic maps lies in the fact that it provides high frequencies in the three regions over time, that is, 0.0, 0.5, and 1.0. Ideally, the distribution ratios of 0.0, 0.5, and 1.0 can be effective in balancing the search behaviour; however, the double-bottom map is designed to satisfy this PSO property.

Algorithm 1 shows the DBM-PSO pseudocode and explains all processes in DBM-PSO to identify the best model of gene-gene interaction. The difference between PSO and DBM-PSO is that the proposed double-bottom map is applied in the updating function of the PSO process (symbol 14 of Algorithm 1). All steps in DBM-PSO for identifying the models of gene-gene interaction problems are explained below.

**2.4. Initializing Particles and DBMr.** In DBM-PSO, a point in the search space is a set which includes the real element  $x$ ,  $x \in R$ . Each particle is a possible solution to the corresponding problem. The subsequent iteration is denoted by  $i = 0, 1, \dots, \text{Iteration}_{\max}$ . Since the elements in a set are likely to change over a sequence of iterations, (1) represents the  $j$ th particle in the population of  $i$ th iteration as

$$X_{j,i} = \{x_{j,i,1}, x_{j,i,2}, \dots, x_{j,i,D} \mid x \in R\}. \quad (1)$$

In this study, a particle in the population represents a solution, that is, a model of gene-gene interaction. A particle contains two separate sets: a set of selected SNPs and a set of genotypes. For each element in  $X_j$ , a certain range within the value is restricted. The values are related to physical components or measurement, that is, natural bounds. The initial population (at  $i = 0$ ) process covers a certain range as much

```

(01) begin
(02) Randomly initialize particles swarm and DBMr
(03) for  $i = 1$  to the number of iteration
(04) Evaluate fitness values of particles by  $FITNESS(X_j, P, N)$ 
(05) for  $j = 1$  to number of particles
(06) Find  $pbest$  by (13)
(07) Find  $gbest$  by (14)
(08) for  $d = 1$  to the number of dimension of particle
(09) Update the velocities of particles by (15)
(10) Update the positions of particles by (16)
(11) next  $d$ 
(12) next  $n$ 
(13) Update the inertia weight value by (17)
(14) Update the value of DBMr by (18)
(15) next  $i$ 
(16) end
    
```

ALGORITHM 1: DBMPSO pseudocode.

```

(01)  $FITNESS(X_j, P, N)$ 
(02) Compute  $a$  using (4)
(03) Compute  $b$  using (5)
(04) Compute  $c$  using (6)
(05) Compute  $d$  using (7)
(06) Compute  $RorP$  using (9)
(07) if the objective is search of risk association model
(08) Compute fitness_value using (10)
(09) else if the objective is search of protection association model
(10) Compute fitness_value using (11)
(11) End if
(12) Return fitness_value
(13) End
    
```

ALGORITHM 2: Fitness value computation pseudocode.

as possible by uniformly randomizing individuals within the search space constrained according to the minimum and maximum bounds, which are represented by  $SNP_{min}$  and  $SNP_{max}$  and  $Genotype_{min}$  and  $Genotype_{max}$ , respectively. Equation (2) shows all genotypes. The homozygous reference genotype is represented as 1, while the heterozygous genotype is represented as 2, and the homozygous variant genotype is represented as 3:

$$Genotype = \begin{cases} 1, & \text{AA type,} \\ 2, & \text{Aa type,} \\ 3, & \text{aa type.} \end{cases} \quad (2)$$

The particles are generated by (3). Particles are initialized by generating the random set in a particle:

$$x_{j,d} = \begin{cases} \text{Random}(SNP_{min}, SNP_{max}), & d \leq \frac{D}{2} \\ \text{Random}(Genotype_{min}, Genotype_{max}), & d > \frac{D}{2} \end{cases} \quad (3)$$

where  $SNP_{max}$  and  $SNP_{min}$  represent a limited SNP, while  $Genotype_{max}$  and  $Genotype_{min}$  represent the limited possible genotypes. For example, let  $X_{j,0} = (1, 3, 4, 2, 1, 2)$ ; thus  $X_{j,0}$  represents the  $j$ th  $X$  in the first generation (at  $i = 0$ ) of selected SNPs (1, 3, 4) and genotypes (2, 1, 2) and can be described by the SNPs associated with the genotypes as follows: (1, 2), (3, 1), and (4, 2).

All random values (DBMr) in the particles are generated with a random value between 0.0 and 1.0 for each independent run.

**2.5. Evaluating the Qualities of Particles Using Fitness Function.** In the DBM-PSO process, the fitness function measures the quality of particles in the population. The studies of gene-gene interaction focus on the combinations of SNP genotypes to identify the highest chi-square ( $\chi^2$ ) value between breast cancer cases and noncancer cases; the value is called the fitness value in DBM-PSO. Algorithm 2 shows the fitness value computation pseudocode. In (4) and (5), symbols  $p$  and  $n$  are, respectively, the sizes of case data and control data, while in (4), (5), (6), and (7),  $P$  and  $N$  are, respectively, the sets

of case data and control data. The  $a$  in (4) is used to count the number of  $P$  including the  $X_j$ ; that is,  $X_j \subseteq P_k$ . The  $b$  in (5) is used to count the number of  $N$  including the  $X_j$ ; that is,  $X_j \subseteq N_k$ . The  $c$  in (6) represents the total number of unmatched  $X_j$  in the  $P$ ; that is,  $X_j \notin P_k$ . The  $d$  in (7) represents the total number of unmatched  $X_j$  in the  $N$ ; that is,  $X_j \notin N_k$ . Equation (9) computes the difference between case data and control data and is used to determine whether the model is associated with risk or protection. Equation (10) is used to compute the fitness value if the objective is to search the risk association model. Equation (11) is used to compute the fitness value if the objective is to search the protection association model. Equation (12) is the chi-square ( $\chi^2$ ) function and is used to compute the  $\chi^2$  value between breast cancer cases and noncancer cases in this study. Consider

$$a = f(X_j) = \sum_{k=1}^p u(X_j, P_k), \quad (4)$$

$$b = f(X_j) = \sum_{k=1}^n u(X_j, N_k), \quad (5)$$

$$c = p - a, \quad (6)$$

$$d = n - b, \quad (7)$$

where

$$u(X_j, A) = \begin{cases} 1, & \forall x \subseteq A, \\ 0, & \forall x \notin A, \end{cases} \quad \forall x \in X_j \quad (8)$$

$$RorP = \frac{100}{(p \times n)(n \times a - p \times b)}, \quad (9)$$

$$\text{fitness\_risk} = \begin{cases} 0, & RorP < 1, \\ \chi^2, & RorP > 1, \end{cases} \quad (10)$$

$$\text{fitness\_protection} = \begin{cases} 0, & RorP > 1, \\ \chi^2, & RorP < 1, \end{cases} \quad (11)$$

$$\chi^2 = \frac{(a + b + c + d)(a \times d - b \times c)^2}{(a + b)(c + d)(a + c)(b + d)}. \quad (12)$$

**2.6. Updating the  $pbest$ s of Particles and  $gbest$  of Population.** Each particle can be improved according to the two objectives,  $pbest$  and  $gbest$ , to search for a better solution.  $pbest_j$  indicates the best value of a position previously visited by the  $j$ th particle, and its position is denoted by  $P_j = (p_{j,1}, p_{j,2}, \dots, p_{j,d})$ . Equations (13) are the updating functions for a particle's  $pbest$  position and  $pbest$  value, respectively, as follows:

$$P_j = \begin{cases} X_j, & f(X_j) \geq pbest_j, \\ P_j, & f(X_j) < pbest_j, \end{cases} \quad (13)$$

$$pbest_j = \begin{cases} f(X_j), & f(X_j) \geq pbest_j, \\ pbest_j, & f(X_j) < pbest_j, \end{cases}$$

where  $gbest$  indicates the best value of all  $pbest$  values for a particle and its position is denoted by  $G = (g_1, g_2, \dots, g_d)$ . Equations (14) provide the updating function for  $gbest$  position and  $gbest$  value, respectively, as follows:

$$G = \begin{cases} P_j, & pbest_j \geq gbest, \\ G, & pbest_j < gbest, \end{cases} \quad (14)$$

$$gbest = \begin{cases} pbest_j, & pbest_j \geq gbest, \\ gbest, & pbest_j < gbest. \end{cases}$$

**2.7. Updating Particle Velocities and Positions.** DBM-PSO executes a search for optimal solutions by continuously updating particle positions in all iterations. Equations (15) and (16) are used to update the velocity and a position of the  $j$ th particle, respectively, as follows:

$$v_{j,d}^{\text{new}} = w \times v_{j,d}^{\text{old}} + c_1 \times \text{DBMr}_{j,1} \times (p_{j,d} - x_{j,d}^{\text{old}}) \quad (15)$$

$$+ c_2 \times \text{DBMr}_{j,2} \times (g_d - x_{j,d}^{\text{old}}),$$

$$x_{j,d}^{\text{new}} = x_{j,d}^{\text{old}} + v_{j,d}^{\text{new}}, \quad (16)$$

where  $c_1$  and  $c_2$  are acceleration constants that control how far a particle moves in a given iteration. Random values,  $\text{DBMr}_{j,1}$  and  $\text{DBMr}_{j,2}$ , in (15) are generated by a function based on the results of the double-bottom map with values between 0.0 and 1.0; they are described in the following section. Velocities  $v_{j,d}^{\text{new}}$  and  $v_{j,d}^{\text{old}}$  are a particle's new and old velocities, respectively. Positions  $x_{j,d}^{\text{old}}$  and  $x_{j,d}^{\text{new}}$  are the particle's current and updated positions, respectively. Variable  $w$  is the inertia weight and is described in the following section.

**2.8. Updating Particle Inertia Weight Values.** Variable  $w$  in DBM-PSO is called the inertia weight which is used to control the impact of a particle's previous velocity. Throughout all iterations,  $w$  decreases linearly from 0.9 to 0.4 [24], and the equation can be written as

$$w = (w_{\max} - w_{\min}) \times \frac{\text{Iteration}_{\max} - \text{Iteration}_i}{\text{Iteration}_{\max}} + w_{\min}, \quad (17)$$

where  $\text{Iteration}_i$  represents the  $i$ th iteration and  $\text{Iteration}_{\max}$  represents the iteration size. Values  $w_{\max}$  and  $w_{\min}$  represent the maximal and minimal values of  $w$ , respectively.

**2.9. Updating Particle DBMr Values.** In DBM-PSO, two random values in the updating function are generated by the following double-bottom map function:

$$\text{DBMr}_{j,t+1} = \frac{[\sin(4\pi \text{DBMr}_{j,t}) + 1]}{2}. \quad (18)$$

**2.10. Parameter Settings.** In this study, all methods used the same parameters to test the search ability for the identification of the models of gene-gene interaction. The population size is 100 and the maximal iteration is 100. The value of

```

(01) begin
(02) for  $n = 1$  to the number of SNP
(03) compute size of "AA" genotype in  $n$ -SNP
(04) compute size of "Aa" genotype in  $n$ -SNP
(05) compute size of "aa" genotype in  $n$ -SNP
(06) generate three genotypes into a set  $A_n$  according each size
(07) randomly sort the elements of  $A_n$ 
(08) next  $n$ 
(09) set dataset =  $\{A_1, A_2, \dots, A_m/m$  is the number of SNP $\}$ 
(10) end

```

ALGORITHM 3: Genotype generator pseudocode.

inertia weight  $w$  is set from 0.9 to 0.4 [25]. Both learning factors,  $c_1$  and  $c_2$ , are equal to 2 [26]. All tests are implemented in Java as a single thread in a PC environment running 32-bit Windows 7 with an Intel coreTM2 Quad CPU Q6600 at 2.4 GHz and 4 GB of RAM.

**2.11. Statistical Analysis.** The model of associations between SNPs can be evaluated by odds ratio (OR) and its 95% CI and  $P$  value [27]. OR can evaluate the models to quantitatively measure the risk of disease;  $P$  value can evaluate whether the results are statistically significant for the difference between the case data and control data. All statistical analyses are implemented using SPSS version 19.0 (SPSS Inc., Chicago, IL).

### 3. Results and Discussion

**3.1. Data Set.** The growth factor-related genes of breast cancer, including genes of EGF, IGF1, IGF1R, IGF2, IGFBP3, IL10, TGFB1, and VEGF with 26 SNPs, were tested in this study. A genotype generator is used to generate a large simulated data set according to the genotype frequencies. Algorithm 3 shows the genotype generator pseudocode to explain how the data set was generated. The genotype frequencies of SNPs are collected from Pharoah et al.'s breast cancer association study [39], which explains the significance of these SNPs of genes in breast cancer.

**3.2. Evaluation of Breast Cancer Susceptibility Using 26 SNPs from Eight Growth Factor-Related Genes.** Table 1 shows the performance (OR and 95% CI) for estimating the effect of a single SNP from eight growth factor-related genes (EGF, IGF1, IGF1R, IGF2, IGFBP3, IL10, TGFB1, and VEGF). Amongst the 26 SNPs in the eight genes, eight SNPs in four genes display a statistically significant OR ( $P < 0.05$ ) for breast cancer. Six SNPs have a risk ( $OR > 1.0$ ) association for breast cancer, including rs5742678-GG, rs1549593-AA, rs6220-GG, IGF1R-10-aa, rs2132572-GA and -AA, and rs1800470-CC. The highest and lowest OR values are 1.33 and 1.09, respectively. Two SNPs have a protection ( $OR < 1.0$ ) association for breast cancer, including rs2229765-AA and rs2854744-CC. The highest and lowest OR values are 0.88 and 0.82, respectively. The other SNPs show no statistically significant OR for breast cancer.

**3.3. Analysis of Models for Gene-Gene Interaction with Risk Association between the Case and Control Data Sets Using PSO, CPSO, and DBM-PSO.** Table 2 shows the 2- to 7-order risk association models for gene-gene interaction. The results are compared with the  $\chi^2$  value, with a high value indicating a good result. The model of 2-SNPs with their corresponding genotypes, SNPs (1, 7) with genotypes 1-3, [rs5742678-CC]-[IGF1R-10-aa], is identified as having 9.451  $\chi^2$  value to explain the difference between the case and control data sets for three methods. However, the results of 3- to 7-SNPs clearly indicate that the DBM-PSO algorithm exhibited an improved search ability over PSO and CPSO in terms of the comparison with the  $\chi^2$  value. For example, in 3-SNPs, DBM-PSO is identified as having a  $\chi^2$  value of 8.772, but those of PSO and CPSO are 3.364 and 3.997, respectively. Table 2 shows the (OR) and its 95% CI, which estimate the impact of the risk association model on the occurrence of breast cancer. A bigger OR value ( $>1$ ) indicates a stronger risk association between the SNPs with combined genotypes and the disease. DBM-PSO shows high OR (1.346–10.018) values for models with a high association for the risk of breast cancer, and the  $P$  value ( $<0.05$ ) indicates that the models have a statistically significant difference between patients and nonpatients. Aside from a 3-SNP model of CPSO, the  $P$  values of models in 3- to 7-SNPs of PSO and CPSO show no statistical significance, indicating that PSO and CPSO have difficulty in identifying statistically significant models for risk association for breast cancer. However, DBM-PSO successfully identifies good models for risk association for breast cancer.

**3.4. Analysis of Models of Gene-Gene Interaction with Protection Association between Case and Control Data Sets Using PSO, CPSO, and DBMPSO.** Table 3 shows the 2- to 7-order protection association models. The OR values ( $<1$ ) estimate the impact of the protection association model on the occurrence of breast cancer. High  $\chi^2$  values in the models indicate good results, and the  $P$  value ( $<0.05$ ) indicates that the model has a statistically significant difference between patients and nonpatients. The results of 3- to 7-SNPs show that DBM-PSO possesses higher  $\chi^2$  values than PSO and CPSO, indicating that DBM-PSO is better to search for good

TABLE 1: Estimated effect (odds ratio and 95% CI) from individual SNPs of 26 growth factor-related genes on the occurrence of breast cancer patients.

SNP (Genes) <sup>a</sup>	SNP type	Case number/normal number <sup>a</sup>	Odds ratio	95% CI
1. rs2237054 (EGF)	1-TT	4408/4418		
	2-TA	570/569	1.00	0.89–1.14
	3-AA	22/13	1.70	0.85–3.37
2. rs5742678 (IGF1)	1-CC	2797/2866		
	2-CG	1844/1837	1.03	0.95–1.12
	3-GG	359/297	1.24	1.05–1.46
3. rs1549593 (IGF1)	1-CC	2924/2970		
	2-CA	1753/1771	1.01	0.93–1.09
	3-AA	323/259	1.27	1.07–1.50
4. rs6220 (IGF1)	1-AA	2643/2698		
	2-AG	1933/1951	1.01	0.93–1.10
	3-GG	424/351	1.23	1.06–1.44
5. rs2946834 (IGF1)	1-CC	2295/2336		
	2-CT	2171/2150	1.03	0.95–1.12
	3-TT	534/514	1.06	0.93–1.21
6. rs1568502 (IGF1R)	1-AA	2914/2955		
	2-AG	1840/1807	1.03	0.95–1.12
	3-GG	246/238	1.05	0.87–1.26
7. IGF1R-10 (IGF1R)	1-AA	3169/3201		
	2-Aa	1545/1582	0.99	0.91–1.08
	3-aa	286/217	1.33	1.11–1.60
8. rs2229765 (IGF1R)	1-GG	1523/1429		
	2-GA	2533/2489	0.96	0.87–1.05
	3-AA	944/1082	0.82	0.73–0.92
9. rs8030950 (IGF1R)	1-CC	2737/2745		
	2-CA	1902/1917	1.00	0.92–1.08
	3-AA	361/338	1.07	0.92–1.25
10. rs680 (IGF2)	1-GG	2538/2451		
	2-GA	2074/2183	0.92	0.85–1.00
	3-AA	388/366	1.02	0.88–1.19
11. rs3741211 (IGF2)	1-TT	1936/1971		
	2-TC	2367/2269	1.06	0.98–1.16
	3-CC	697/760	0.93	0.83–1.05
12. IGF2-05 (IGF2)	1-AA	2651/2694		
	2-Aa	1955/1952	1.02	0.94–1.11
	3-aa	394/354	1.13	0.97–1.32
13. IGF2-06 (IGF2)	1-AA	2160/2162		
	2-Aa	2237/2284	0.98	0.90–1.07
	3-aa	603/554	1.09	0.96–1.24
14. rs2132571 (IGFBP3)	1-GG	2415/2407		
	2-GA	2163/2157	1.00	0.92–1.09
	3-AA	422/436	0.97	0.83–1.12
15. rs2471551 (IGFBP3)	1-GG	3225/3284		
	2-GC	1591/1515	1.07	0.98–1.17
	3-CC	184/201	0.93	0.76–1.15

TABLE 1: Continued.

SNP (Genes) <sup>a</sup>	SNP type	Case number/normal number <sup>a</sup>	Odds ratio	95% CI
16. rs2854744 (IGFBP3)	1-AA	1538/1469		
	2-AC	2487/2475	0.96	0.88–1.05
	3-CC	975/1056	0.88	0.79–0.99
17. rs2132572 (IGFBP3)	1-GG	2908/3027		
	2-GA	1805/1728	1.09	1.00–1.18
	3-AA	287/245	1.22	1.02–1.46
18. rs3024496 (IL10)	1-TT	1218/1235		
	2-TC	2533/2549	1.01	0.92–1.11
	3-CC	1249/1216	1.04	0.93–1.17
19. rs1800872 (IL10)	1-CC	3059/3017		
	2-CA	1660/1722	0.95	0.87–1.03
	3-AA	281/261	1.06	0.89–1.27
20. rs1800890 (IL10)	1-TT	1703/1701		
	2-TA	2455/2508	0.98	0.90–1.07
	3-AA	842/791	1.06	0.95–1.20
21. rs1554286 (IL10)	1-CC	3400/3446		
	2-CT	1431/1410	1.03	0.94–1.12
	3-TT	169/144	1.19	0.95–1.49
22. rs1800470 (TGFB1)	1-TT	1850/1914		
	2-TC	2372/2399	1.02	0.94–1.11
	3-CC	778/687	1.17	1.04–1.32
23. rs699947 (VEGF)	1-CC	1236/1273		
	2-CA	2511/2463	1.05	0.95–1.16
	3-AA	1253/1264	1.02	0.91–1.14
24. rs1570360 (VEGF)	1-GG	2278/2341		
	2-GA	2214/2132	1.07	0.98–1.16
	3-AA	508/527	0.99	0.87–1.13
25. rs2010963 (VEGF)	1-GG	2354/2279		
	2-GC	2133/2157	0.96	0.88–1.04
	3-CC	513/564	0.88	0.77–1.01
26. rs3025039 (VEGF)	1-CC	3744/3741		
	2-CT	1160/1174	0.99	0.90–1.08
	3-TT	96/85	1.13	0.84–1.52

<sup>a</sup>Data collected from the literature [39].

protection association models than other methods. DBM-PSO has OR values ranging from 0.755 to 0.850, with a  $P$  value of  $<0.05$  for protection with breast cancer. The 2-SNP and 3-SNP models in PSO and CPSO show a statistically significant difference between patients and nonpatients ( $P < 0.05$ ), and the 4-SNP model in CPSO also shows a statistically significant difference. Although CPSO provides better OR values than DBM-PSO in the 5-, 6-, and 7-SNP models, the  $P$  values indicate that these models are not statistically significant. DBM-PSO successfully identifies good models for protection association for breast cancer.

3.5. Discussion. Effects between SNPs from several genes could contribute to disease development. Case-control



TABLE 2: Estimation of the best risk model of gene-gene interaction on the occurrence of breast cancer as determined by PSO, CPSO, and DBMPSO.

	Combined SNP	SNP genotypes	Cases number	Controls number	$\chi^2$ value	Odds ratio	95% CI	P value
2-SNP								
PSO	1, 7	1-3	259	195	9.451	1.346	1.11-1.63	0.002
		Other	4741	4805				
CPSO	1, 7	1-3	259	195	9.451	1.346	1.11-1.63	0.002
		Other	4741	4805				
DBMPSO	1, 7	1-3	259	195	9.451	1.346	1.11-1.63	0.002
		Other	4741	4805				
3-SNP								
PSO	2, 14, 25	3-1-1	84	62	3.364	1.361	0.98-1.89	0.068
		Other	4916	4938				
CPSO	1, 6, 7	1-1-3	148	116	3.997	1.285	1.00-1.64	0.046
		Other	4850	4884				
DBMPSO	7, 11, 21	3-2-1	93	57	<b>8.772</b>	<b>1.644</b>	<b>1.18-2.29</b>	<b>0.003</b>
		Other	4907	4943				
4-SNP								
PSO	1, 14, 20, 23	3-3-1-2	2	0	1.000	3.001	0.31-28.86	0.341
		Other	4998	5000				
CPSO	1, 4, 11, 14	1-3-2-1	86	67	2.396	1.289	0.93-1.78	0.123
		Other	4914	4933				
DBMPSO	1, 7, 11, 21	1-3-2-1	87	53	<b>8.374</b>	<b>1.653</b>	<b>1.17-2.33</b>	<b>0.004</b>
		Other	4913	4947				
5-SNP								
PSO	2, 7, 15, 18, 24	1-3-1-3-2	15	8	2.135	1.878	0.80-4.43	0.151
		Other	4985	4992				
CPSO	3, 10, 17, 24, 26	3-1-1-3-1	9	3	3.004	3.004	0.81-11.10	0.099
		Other	4991	4997				
DBMPSO	1, 2, 7, 11, 21	1-1-3-2-1	49	27	<b>6.417</b>	<b>1.823</b>	<b>1.14-2.92</b>	<b>0.013</b>
		Other	4951	4973				
6-SNP								
PSO	2, 6, 8, 16, 18, 25	3-1-1-2-3-2	3	1	1.000	3.001	0.31-28.86	0.341
		Other	4997	4999				
CPSO	2, 11, 16, 18, 22, 23	1-2-1-2-3-2	14	9	1.089	1.557	0.67-3.60	0.301
		Other	4986	4991				
DBMPSO	1, 2, 7, 10, 11, 21	1-1-3-1-2-1	27	12	<b>6.417</b>	<b>2.257</b>	<b>1.14-4.46</b>	<b>0.019</b>
		Other	4973	4988				
7-SNP								
PSO	1, 3, 6, 12, 21, 24, 26	1-3-2-1-3-1-1	2	0	1.000	3.001	0.31-28.86	0.341
		Other	4998	5000				
CPSO	1, 2, 3, 9, 19, 21, 24	1-1-3-1-1-2-3	4	1	1.801	4.002	0.45-35.82	0.215
		Other	4996	4999				
DBMPSO	1, 3, 5, 9, 17, 23, 24	1-3-2-1-2-2-1	10	1	<b>7.372</b>	<b>10.018</b>	<b>1.28-78.29</b>	<b>0.028</b>
		Other	4990	4999				

studies are the main method to determine the association between SNPs. Many breast cancer studies have analysed the associations between important related genes [28-34], hypothesizing that disease risk may be associated with the cooccurrence of SNPs displaying a jointed effect, including genes related to DNA repair [35, 36], chemokine ligand-receptor interactions [37], and estrogen-response genes [4].

Evolutionary algorithms are applied to identify good models of gene-gene interaction [7, 9]. Previous studies have used the difference between case and control data sets to design the fitness function, allowing for the identification of models with high difference values for all SNP combinations. However, the highest difference between the case and control data sets is not necessarily statistically significant

TABLE 3: Estimation of the best protection model of gene-gene interaction on the occurrence of breast cancer as determined by PSO, CPSO, and DBMPSO.

	Combined SNP	SNP genotypes	Cases number	Controls number	$\chi^2$ value	Odds ratio	95% CI	P value
2-SNP								
PSO	1, 8	1-3	816	941	10.789	0.841	0.76–0.93	0.001
		Other	4184	4059				
CPSO	1, 8	1-3	816	941	10.789	0.841	0.76–0.93	0.001
		Other	4184	4059				
DBMPSO	1, 8	1-3	816	941	10.789	0.841	0.76–0.93	0.001
		Other	4184	4059				
3-SNP								
PSO	8, 9, 22	3-1-2	225	269	4.123	0.829	0.69–0.99	0.043
		Other	4775	4731				
CPSO	3, 8, 9	1-3-1	319	371	4.209	0.850	0.73–0.99	0.040
		Other	4681	4629				
DBMPSO	1, 8, 15	1-3-1	527	624	<b>9.238</b>	<b>0.826</b>	<b>0.73–0.94</b>	<b>0.002</b>
		Other	4473	4376				
4-SNP								
PSO	4, 8, 14, 22	2-3-1-2	76	99	3.077	0.764	0.57–1.03	0.080
		Other	4924	4901				
CPSO	10, 17, 21, 23	2-1-1-1	223	268	4.337	0.824	0.69–0.99	0.038
		Other	4777	4732				
DBMPSO	1, 10, 17, 21	1-2-1-1	692	795	<b>8.381</b>	<b>0.850</b>	<b>0.76–0.95</b>	<b>0.004</b>
		Other	4308	4205				
5-SNP								
PSO	5, 6, 8, 9, 26	1-1-3-2-1	75	91	1.568	0.821	0.60–1.12	0.211
		Other	4925	4909				
CPSO	2, 4, 8, 11, 18	1-2-3-1-2	32	44	1.909	0.726	0.46–1.15	0.169
		Other	4968	4956				
DBMPSO	1, 2, 6, 8, 15	1-1-1-3-1	167	218	<b>7.026</b>	<b>0.758</b>	<b>0.62–0.93</b>	<b>0.008</b>
		Other	4833	4782				
6-SNP								
PSO	4, 8, 15, 19, 22, 24	1-3-2-2-1-3	0	2	1.000	0.333	0.04–3.20	0.341
		Other	5000	4998				
CPSO	3, 4, 12, 16, 20, 24	1-1-1-2-2-3	21	28	1.005	0.749	0.43–1.32	0.318
		Other	4979	4972				
DBMPSO	1, 10, 15, 17, 21, 26	1-2-1-1-1-1	327	394	<b>6.710</b>	<b>0.818</b>	<b>0.70–0.95</b>	<b>0.010</b>
		Other	4673	4606				
7-SNP								
PSO	5, 8, 11, 13, 14, 24, 25	1-1-3-1-1-2-1	3	6	1.001	0.500	0.13–2.00	0.327
		Other	4997	4994				
CPSO	10, 12, 16, 17, 19, 22, 26	2-2-2-1-2-2-1	20	27	1.047	0.740	0.41–1.32	0.308
		Other	4980	4973				
DBMPSO	1, 10, 13, 15, 17, 21, 26	1-2-2-1-1-1-1	141	185	<b>6.139</b>	<b>0.755</b>	<b>0.60–0.94</b>	<b>0.014</b>
		Other	4859	4815				

( $P < 0.05$ ). The chi-square test is a statistical tool to evaluate the difference between the observed and expected data sets under specific hypothetical conditions. A property of the chi-square test is that the chi-square value is inversely proportional to  $P$  value. Therefore, the chi-square test is used to design the fitness function in this study. PSO and CPSO [7] were used to search for good models based on the new fitness function, but the results (Tables 2 and 3) fail to identify high-order associations. However, DBM-PSO effectively identified

good risk and protection association models of gene-gene interactions for breast cancer. Statistical methods, such as  $P$  value, OR, and its 95% CI, provide strong validation of the search ability of DBM-PSO.

PSO and DBM-PSO use the fitness functional computation to calculate complexity. DBM-PSO can be observed in (15) and (18). Equation (18) is only used to amend the original PSO updating equation (15). Therefore, DBM-PSO does not increase the complexity of the PSO search process.

The computational complexity of DBM-PSO is big- $O(nm)$ , where  $n$  is the number of iterations and  $m$  is the number of particles.

The results of DBM-PSO are influenced by its parameters, including double-bottom chaotic maps (18), population size, iteration size, and  $c_1$  and  $c_2$  in the updating function (15). Yang et al. [13] tested the 22 most commonly used representative benchmark functions, selecting the optimal parameters ( $4\pi$ ) in the proposed double-bottom chaotic maps. Therefore, the parameter is suggested as  $4\pi$  in (18). The population and iteration sizes could be adjusted according to the size of the data set. Population size suggested a setting from 50 to 200 and the suggested number of iterations ranges from 100 to 1000.  $c_1$  and  $c_2$  are both suggested to be 2 [38].

#### 4. Conclusion

We proposed a new fitness function to identify good models of gene-gene interaction for the investigation of polygenic diseases and cancers. The fitness function based on chi-square test addresses the disadvantage of previously proposed fitness functions, in that the highest difference between the case and control data sets is not necessarily statistically significant ( $P < 0.05$ ). Our proposed DBM-PSO showed to be able to successfully determine the 26 SNP cross interactions for risk and protection models of gene-gene interactions in breast cancer. The results indicate that DBM-PSO can successfully use the chi-square test to identify good models by evaluating the difference between the observed and expected data sets under specific hypothetical conditions.

#### Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

#### Acknowledgments

This study was partly supported by the National Science Council of Taiwan for Grants NSC102-2221-E-151-024-MY3, NSC102-2622-E-151-003-CC3, NSC101-2221-E-214-075, NSC101-2622-E-151-027-CC3, NSC100-2221-E-151-049-MY3, and NSC100-2221-E-151-051-MY2, the National Sun Yat-sen University-KMU Joint Research Project (no. NSYSU-KMU 103-p014), and the Ministry of Health and Welfare, Taiwan (MOHW103-TD-B-111-05).

#### References

- [1] X. Li, H. Chen, J. Li, and Z. Zhang, "Gene function prediction with gene interaction networks: a context graph kernel approach," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 1, pp. 119–128, 2010.
- [2] P. Kraft and C. A. Haiman, "GWAS identifies a common breast cancer risk allele among BRCA1 carriers," *Nature Genetics*, vol. 42, no. 10, pp. 819–820, 2010.
- [3] D. Fanale, V. Amodeo, L. R. Corsini, S. Rizzo, V. Bazan, and A. Russo, "Breast cancer genome-wide association studies: there is strength in numbers," *Oncogene*, vol. 31, no. 17, pp. 2121–2128, 2012.
- [4] J.-C. Yu, C.-N. Hsiung, H.-M. Hsu et al., "Genetic variation in the genome-wide predicted estrogen response element-related sequences is associated with breast cancer development," *Breast Cancer Research*, vol. 13, no. 1, p. R13, 2011.
- [5] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics*, vol. 26, no. 4, pp. 445–455, 2010.
- [6] P. Yang, J. W. K. Ho, Y. H. Yang, and B. B. Zhou, "Gene-gene interaction filtering with ensemble of filters," *BMC Bioinformatics*, vol. 12, supplement 1, p. S10, 2011.
- [7] L.-Y. Chuang, H.-W. Chang, M.-C. Lin, and C.-H. Yang, "Chaotic particle swarm optimization for detecting SNP-SNP interactions for CXCL12-related genes in breast cancer prevention," *European Journal of Cancer Prevention*, vol. 21, pp. 336–342, 2012.
- [8] L. Y. Chuang, Y. D. Lin, H. W. Chang, and C. H. Yang, "An improved PSO algorithm for generating protective SNP barcodes in breast cancer," *PLoS ONE*, vol. 7, Article ID e37018, 2012.
- [9] C. H. Yang, L. Y. Chuang, Y. H. Cheng et al., "Single nucleotide polymorphism barcoding to evaluate oral cancer risk using odds ratio-based genetic algorithms," *Kaohsiung Journal of Medical Sciences*, vol. 28, pp. 362–368, 2012.
- [10] J. B. Chen, L. Y. Chuang, Y. D. Lin et al., "Preventive SNP-SNP interactions in the mitochondrial displacement loop (D-loop) from chronic dialysis patients," *Mitochondrion*, vol. 13, pp. 698–704, 2013.
- [11] C. H. Yang, Y. D. Lin, L. Y. Chuang, and H. W. Chang, "Evaluation of breast cancer susceptibility using improved genetic algorithms to generate genotype SNP barcodes," *IEEE-ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, pp. 361–371, 2013.
- [12] C. H. Yang, Y. D. Lin, L. Y. Chuang, J. B. Chen, and H. W. Chang, "MDR-ER: balancing functions for adjusting the ratio in risk classes and classification errors for imbalanced cases and controls using multifactor-dimensionality reduction," *PLoS ONE*, vol. 8, Article ID e79387, 2013.
- [13] C. H. Yang, S. W. Tsai, L. Y. Chuang, and C. H. Yang, "An improved particle swarm optimization with double-bottom chaotic maps for numerical optimization," *Applied Mathematics and Computation*, vol. 219, pp. 260–279, 2012.
- [14] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, pp. 1942–1948, Perth, Australia, December 1995.
- [15] E. Garcia-Gonzalo and J. Fernandez-Martinez, "A brief historical review of particle swarm optimization (PSO)," *Journal of Bioinformatics and Intelligent Control*, vol. 1, pp. 3–16, 2012.
- [16] D. Chen and C. Zhao, "Particle swarm optimization with adaptive population size and its application," *Applied Soft Computing Journal*, vol. 9, no. 1, pp. 39–48, 2009.
- [17] L. Ali, S. L. Sabat, and S. K. Udgata, "Particle swarm optimization with stochastic ranking for constrained numerical and engineering benchmark problems," *International Journal of Bio-Inspired Computation*, vol. 4, pp. 155–166, 2012.
- [18] H. M. Abdelsalam and A. M. Mohamed, "Optimal sequencing of design projects' activities using discrete particle swarm optimisation," *International Journal of Bio-Inspired Computation*, vol. 4, pp. 100–110, 2012.

- [19] Z. Cui and X. Cai, "Integral particle swarm optimization with dispersed accelerator information," *Fundamenta Informaticae*, vol. 95, no. 4, pp. 427–447, 2009.
- [20] Z. Cui, X. Cai, J. Zeng, and Y. Yin, "PID-controlled particle swarm optimization," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 16, no. 6, pp. 585–609, 2010.
- [21] C. Priya and P. Lakshmi, "Particle swarm optimisation applied to real time control of spherical tank system," *International Journal of Bio-Inspired Computation*, vol. 4, pp. 206–216, 2012.
- [22] M. Salehi Maleh, S. Soleymani, R. Rasouli Nezhad, and N. Ghadimi, "Using particle swarm optimization algorithm based on multi-objective function in reconfigured system for optimal placement of distributed generation," *Journal of Bioinformatics and Intelligent Control*, vol. 2, pp. 119–124, 2013.
- [23] T. Xiang, X. Liao, and K.-W. Wong, "An improved particle swarm optimization algorithm combined with piecewise linear chaotic map," *Applied Mathematics and Computation*, vol. 190, no. 2, pp. 1637–1645, 2007.
- [24] Y. Shi and R. C. Eberhart, "Empirical study of particle swarm optimization," in *Proceedings of the Congress on Evolutionary Computation*, pp. 1945–1949, Washington, DC, USA, 1999.
- [25] Y. Shi and R. Eberhart, "Empirical study of particle swarm optimization," in *Proceedings of the IEEE Congress on Evolutionary Computation*, pp. 1948–1950, 1999.
- [26] A. Ratnaweera, S. K. Halgamuge, and H. C. Watson, "Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 3, pp. 240–255, 2004.
- [27] L. E. Mechanic, B. T. Luke, J. E. Goodman, S. J. Chanock, and C. C. Harris, "Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions," *BMC Bioinformatics*, vol. 9, article 146, 2008.
- [28] S. L. Zheng, J. Sun, F. Wiklund et al., "Cumulative association of five genetic variants with prostate cancer," *The New England Journal of Medicine*, vol. 358, no. 9, pp. 910–919, 2008.
- [29] C.-Y. Yen, S.-Y. Liu, C.-H. Chen et al., "Combinational polymorphisms of four DNA repair genes XRCC1, XRCC2, XRCC3, and XRCC4 and their association with oral cancer in Taiwan," *Journal of Oral Pathology and Medicine*, vol. 37, no. 5, pp. 271–277, 2008.
- [30] G.-T. Lin, H.-F. Tseng, C.-K. Chang et al., "SNP combinations in chromosome-wide genes are associated with bone mineral density in Taiwanese women," *Chinese Journal of Physiology*, vol. 51, no. 1, pp. 32–41, 2008.
- [31] S. Cauchi, D. Meyre, E. Durand et al., "Post genome-wide association studies of novel genes associated with type 2 diabetes show gene-gene interaction and high predictive value," *PLoS ONE*, vol. 3, no. 5, Article ID e2031, 2008.
- [32] F. Ricceri, S. Guarrera, C. Sacerdote et al., "XRCC1 haplotypes modify bladder cancer risk: a case-control study," *DNA Repair*, vol. 9, no. 2, pp. 191–200, 2010.
- [33] J. Yin, K. Lu, J. Lin et al., "Genetic variants in TGF- $\beta$  pathway are associated with ovarian cancer risk," *PLoS ONE*, vol. 6, no. 9, Article ID e25559, 2011.
- [34] L. Chen, W. Li, L. Zhang et al., "Disease gene interaction pathways: a potential framework for how disease genes associate by disease-risk modules," *PLoS ONE*, vol. 6, no. 9, Article ID e24495, 2011.
- [35] W. Han, K.-Y. Kim, S.-J. Yang, D.-Y. Noh, D. Kang, and K. Kwack, "SNP-SNP interactions between DNA repair genes were associated with breast cancer risk in a Korean population," *Cancer*, vol. 118, no. 3, pp. 594–602, 2012.
- [36] J. Conde, S. N. Silva, A. P. Azevedo et al., "Association of common variants in mismatch repair genes and breast cancer susceptibility: a multigene study," *BMC Cancer*, vol. 9, article 344, 2009.
- [37] G.-T. Lin, H.-F. Tseng, C.-H. Yang et al., "Combinational polymorphisms of seven CXCL12-related genes are protective against breast cancer in Taiwan," *OMICS*, vol. 13, no. 2, pp. 165–172, 2009.
- [38] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, pp. 1942–1948, Perth, Australia, December 1995.
- [39] P. D. P. Pharoah, J. Tyrer, A. M. Dunning, D. F. Easton, and B. A. J. Ponder, "Association between common variation in 120 candidate genes and breast cancer risk," *PLoS Genetics*, vol. 3, no. 3, p. e42, 2007.

## Research Article

# Novel Design Strategy for Checkpoint Kinase 2 Inhibitors Using Pharmacophore Modeling, Combinatorial Fusion, and Virtual Screening

Chun-Yuan Lin<sup>1,2</sup> and Yen-Ling Wang<sup>1</sup>

<sup>1</sup> Department of Computer Science and Information Engineering, Chang Gung University, Taoyuan 33302, Taiwan

<sup>2</sup> Research Center for Emerging Viral Infections, Chang Gung University, Taoyuan 33302, Taiwan

Correspondence should be addressed to Chun-Yuan Lin; [cyulin@mail.cgu.edu.tw](mailto:cyulin@mail.cgu.edu.tw)

Received 29 November 2013; Accepted 19 February 2014; Published 23 April 2014

Academic Editor: Che-Lun Hung

Copyright © 2014 C.-Y. Lin and Y.-L. Wang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Checkpoint kinase 2 (Chk2) has a great effect on DNA-damage and plays an important role in response to DNA double-strand breaks and related lesions. In this study, we will concentrate on Chk2 and the purpose is to find the potential inhibitors by the pharmacophore hypotheses (PhModels), combinatorial fusion, and virtual screening techniques. Applying combinatorial fusion into PhModels and virtual screening techniques is a novel design strategy for drug design. We used combinatorial fusion to analyze the prediction results and then obtained the best correlation coefficient of the testing set ( $r_{\text{test}}$ ) with the value 0.816 by combining the  $\text{Best}_{\text{train}}\text{Best}_{\text{test}}$  and  $\text{Fast}_{\text{train}}\text{Fast}_{\text{test}}$  prediction results. The potential inhibitors were selected from NCI database by screening according to  $\text{Best}_{\text{train}}\text{Best}_{\text{test}} + \text{Fast}_{\text{train}}\text{Fast}_{\text{test}}$  prediction results and molecular docking with CDOCKER docking program. Finally, the selected compounds have high interaction energy between a ligand and a receptor. Through these approaches, 23 potential inhibitors for Chk2 are retrieved for further study.

## 1. Introduction

DNA-damage is induced by ionizing radiation, genotoxic chemicals, or collapsed replication forks, and when DNA was damaged or the responses of cells were failure, the mutation associated with the breast or ovarian cancer of genes may occur. To prevent and repair the DNA-damage, mammalian cells will control and stabilize the genome by cell cycle checkpoint. The checkpoint pathway consists of several kinases, such as ataxia telangiectasia mutated protein (ATM [1, 2]), ataxia telangiectasia and Rad3-related protein (ATR [1, 2]), checkpoint kinase 1 (Chk1 [3, 4]), and checkpoint kinase 2 (Chk2 [5–8]). ATM and ATR are upstream kinases passing messages to downstream kinases and phosphorylating several proteins that initiate the activation of the DNA-damage checkpoint. Moreover, ATM is a primarily pathway to activate p53 (protein 53 [9]) by Chk2, and ATR may influence the phosphorylation of Chk1. Both Chk1 and Chk2 are key components in DNA-damage; however, their cellular

activities are different. Chk1 is involved in S and G2 phases of the cell cycle with ATR pathway. By contrast, Chk2 is activated in all phases through ATM-dependent pathway and plays an important role in response to DNA double-strand breaks and related lesions. Furthermore, Chk1 is an unstable protein and lacks the forkhead-associated domain (FHA) which was involved in several processes that protect against cancer and can be found in Chk2. Therefore, we concentrate on Chk2 in this study.

Chk2 is a protein containing 543 amino acid residues and the structure of Chk2 consists of some functional elements, including the N-terminal SQ/TQ cluster domain (SCD), FHA, and the N-terminal serine/threonine kinase domain (KD) [5–8]. The SCD is known to be the preferred site with the residue Thr68 for phosphorylation to respond to DNA-damage by ATM/ATP kinases. The FHA domain is a phosphopeptide recognition domain found in many regulatory proteins and thought to bind to the phosphoThr68 segment of SCD [5–8, 10–14]. Hence it is a good candidate

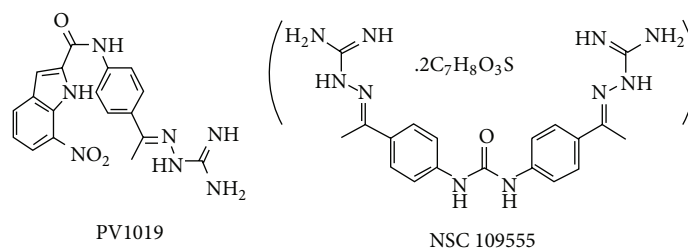


FIGURE 1: Two-dimensional chemical structures of known Chk2 inhibitors. The experimental  $IC_{50}$  of PV1019 and NSC 109555 were 138 nM and 240 nM, respectively.

for interactions of Chk2 with its upstream regulators or downstream targets in the cell-cycle-checkpoint signaling. The KD occupies almost the entire carboxy-terminal half of Chk2 and has been identified based on their homology with serine/threonine kinases. Some studies reported that when DNA was damaged, Chk2 is activated by ATM/ATR through the phosphorylation of residue Thr68. Moreover, Chk2 induces transautophosphorylation of residues Thr383 and Thr387 and then cis-phosphorylation of residue Ser516 [5–8, 10–14]. After that, Chk2 will phosphorylate several downstream substrates, such as BRCA1 (breast cancer 1, early onset [15, 16]), Cdc25A (cell division cycle 25 homolog A), Cdc25C, and p53 [7, 8, 10]. Several researches indicated that Chk2 phosphorylates Cdc25A which is considered an oncogene on the residue Ser123 in S phase of cell cycle, and it also phosphorylates Cdc25C on the residue Ser216 in G2 phase helping prevent mitotic entry in cells with damaged DNA [5]. Furthermore, BRCA1 and p53 are involved in DNA repair process in the breast or ovarian cancer. BRCA1 is a human caretaker gene and helps repair damaged DNA or destroys cells which cannot be repaired. The p53 is a tumor suppressor protein involved in preventing cancer in human and plays an important role in the G1 checkpoint in response to DNA damaging agents. We consider that the sites of the phosphorylations are important in the drug design for cell survival when DNA is damaged.

Recently, several studies identified the inhibitors of Chk2 [6–8, 10–14], and they also showed the crystal structures of Chk2 complex, such as PDB: 1GXG, 2W7X, and, and so forth. They are selective, reversible, and ATP-competitive Chk2 inhibitors demonstrating that they effectively restrain the radiation-induced phosphorylation of Chk2. In addition, several selective Chk2 inhibitors have been also identified (two examples were shown in Figure 1) and the researches indicated that they are potential and selective inhibitors of Chk2 with chemotherapeutic and radiosensitization potential. On structure-based drug design, several developments of Chk2 were published [17, 18]. Quantitative structure-activity relationship model (QSAR model) is a regression or classification model and is an important technique in the rational drug design. It is used to correlate the structure properties of compounds with their biological activities. The method to predict the quality by QSAR was improved by considering the three-dimensional structure QSAR (3D-QSAR) [19–24] of targeted inhibitor. Therefore, the compound structure can be directly optimized in the 3D space.

The comparative molecular field analyses (CoMFA) [18, 25–30] and the comparative molecular similarity indices analyses (CoMSIA) [18, 27–32] for Chk2 inhibitors were performed by ligand-based and receptor-guided alignment. They used the cocrystal structure from protein data bank (PDB code: 2CN8) [7], and then they identified new plausible binding modes used as template for 3D-QSAR [18]. There is another research of Chk2 studied in QSAR/QSPR [17] providing structures that will improve reducing the side effects of Chk2 inhibitors.

Pharmacophore [20–24, 33–35] is a set of structural features responsible for the biological activity of a molecule. It allowed compounds with diverse structures to find the common chemical features by ligand pharmacophore mapping, and that is different from CoMFA and CoMSIA with the common structure constraint. Thus, pharmacophore can explain how diverse ligands bind to a receptor site by these features and visualize the feature of potential chemical interactions between ligands and receptors. Moreover, pharmacophore can easily and quickly identify candidate inhibitors for a target protein based on 3D query. Therefore, in this work, we first used 3D-QSAR study to build pharmacophore hypotheses (denoted as PhModels) for Chk2 inhibitors by HypoGen Best, Fast, and Caesar algorithms, respectively. Then we used the combinatorial fusion to select and combine prediction results for improving the predictive accuracy in biological activities of inhibitors. Virtual screening is a computational technique used in drug discovery research. There are two categories of screening techniques: ligand-based and structure-based. In this work, for ligand-based virtual screening, we used the selected PhModels as 3D structure query by pharmacophore hypothesis screening that each compound in National Cancer Institute (NCI) database will be mapped onto the pharmacophoric features of selected PhModels. When the chemical features of a compound fit the generated PhModels, it will be selected. All of feasible compounds in NCI database were selected in this work. Finally, the potential inhibitors were retrieved from selected compounds by using molecular docking program to predict the conformation and interaction energy between Chk2 and ligand. Applying combinatorial fusion into PhModels and virtual screening techniques is a novel design strategy for drug design and can help medicinal chemists to identify or design new Chk2 inhibitors. Besides, the potential inhibitors of Chk2 retrieved in this work can be estimated by biologists for further study.

## 2. Materials and Methods

**2.1. Biological Data Collection.** In order to construct the PhModels, at first, we collected the Chk2 inhibitors with two-dimensional structures and the biological activity values from the ChEMBL database [36]. Then, according to the structure variations and chemical differences in the kinase inhibitor activity, 158 known Chk2 inhibitors were selected and retrieved. The biological activity of 158 known Chk2 inhibitors was represented as  $IC_{50}$  (nanomolar, nM). There are 260,071 compounds from the NCI database (release version 3, <http://cactus.nci.nih.gov/download/nci/>) which were used in the database screening and molecular docking approach in this work.

**2.2. Training and Testing Sets Selection.** Before generating PhModels, we should divide the 158 Chk2 inhibitors into the training set and testing set, respectively. The rules used to select training set inhibitors are according to the following requirements as suggested by the Accelrys Discovery Studio. (1) All selected inhibitors should have clear and concise information including structure features and activity range. (2) At a minimum, 16 diverse inhibitors for training set were selected to ensure the statistical significance. (3) The training set should contain the most and the least active inhibitors. (4) The biological activities of the inhibitors spanned at least 4 orders of magnitude. Based on the above four rules, the 158 Chk2 inhibitors were divided, and the scatter diagram of training set and testing set inhibitors was shown in Figure 2. Figure 2 demonstrates the distribution of the inhibitors in the training set and testing set, and the representative points of the testing set are close to those of the training set. The training set with 25 inhibitors is used to construct PhModels, and the  $IC_{50}$  values of these 25 inhibitors are ranged from 2.3 to 100,000 nM (Table 1). The testing set with remaining 133 inhibitors is used to test the predictive ability of generated PhModels, and the  $IC_{50}$  values of the 133 testing set inhibitors are ranged from 3.4 to 74,000 nM (Table 2). After selecting the training set and testing set inhibitors, we established PhModels at first, and then we used the correlation analysis to estimate the prediction abilities of PhModels.

**2.3. Pharmacophore Generation.** The workflow of PhModel generation for Chk2 inhibitors was shown in Figure 3. In this study, we used the HypoGen program [37] in Accelrys Discovery Studio 2.1 to generate PhModels. At the initial step, 3D conformations of the training set inhibitors were generated by using “3D-QSAR Pharmacophore Generation protocol” with the Best, Fast, and Caesar generating algorithms, respectively, based on the CHARMM-like force field. The conformational-space energy was constrained  $\leq 20$  kcal/mol which represented the maximum allowed energy above the global minimum energy. For each training set inhibitor, the number of the diverse 3D conformations was set to  $\leq 255$ . All other parameters were set as default values. Following the above rules, the 3D conformations were generated, and then we can construct the PhModel by using “Ligand Pharmacophore Mapping protocol.” Each of the ten PhModels using

HypoGen Best, Fast, and Caesar algorithms were generated in this study.

**2.4. Combinatorial Fusion.** In this study, we use a combinatorial fusion technique to facilitate prediction results selection and combination for improving predictive accuracy in biological activities of inhibitors. The combinatorial fusion we take is analogous to that used in information retrieval [38, 39], pattern recognition [40], molecular similarity searching and structure-based screening [41], and microarray gene expression analysis [42]. These works have demonstrated the following remark [43].

*Remark 1.* For a set of multiple scoring systems, each with a score function and a rank function, we have that (a) the combination of multiple scoring systems would improve the prediction accuracy only if (1) each of the systems has a relatively high performance, and (2) the individual systems are distinctive (or diversified), and (b) rank combination performs better than score combination under certain conditions.

Given an inhibitor and for each prediction result  $A$ , let  $s_A$  be a function as the predicted biological activity and it is represented as a real number. We view the function  $s_A$  as the score function. Since  $s_A$  only assigns a number not a set of numbers, in this work, no rank function would be used for an inhibitor. Therefore, the rank combination and the rule (b) in Remark 1 are not considered in the study. Suppose we have  $m$  prediction results ( $m$  scoring functions). There are combinatorially  $2^m - 1$  combinations for all  $m$  individual prediction results ( $\sum_{k=1}^m \binom{m}{k} = 2^m - 1$ ) with score functions. The total number of combinations to be considered for predicting biological activity of an inhibitor is  $2^m - 1$ . This number of combinations can become huge when the number of prediction results  $m$  is large. Moreover, we have to evaluate the predictive power of each combination across all inhibitors. This study would start with combining only two prediction results which still retain fairly good prediction power.

Suppose  $m$  prediction results  $A_i$ ,  $i = 1, 2, \dots, m$ , are given with score function  $s_{A_i}$ ; there are several different ways of combination. Among others, there are score combination, voting, linear average combination, and weighted combination [38–42]. Voting is computationally simple and better than simple linear combinations when applied to the situation with large number of prediction results. However, a better alternative is to reduce the number of prediction results to a smaller number and then these prediction results are combined. In this paper, we reduce the set of prediction results to those which perform relatively well and then use the rank/score function to decide whether to combine by score. In this paper, we use the rules (a) (1) and (a) (2) stated in Remark 1 as our guiding principle to select prediction results and to decide on the method of combination. After generating each of the ten PhModels by using HypoGen Best, Fast, and Caesar algorithms for training set inhibitors, each of the best PhModel (denoted as  $Best_{train}$ ,  $Fast_{train}$ , and  $Caesar_{train}$ ) was evaluated by its correlation coefficient of the

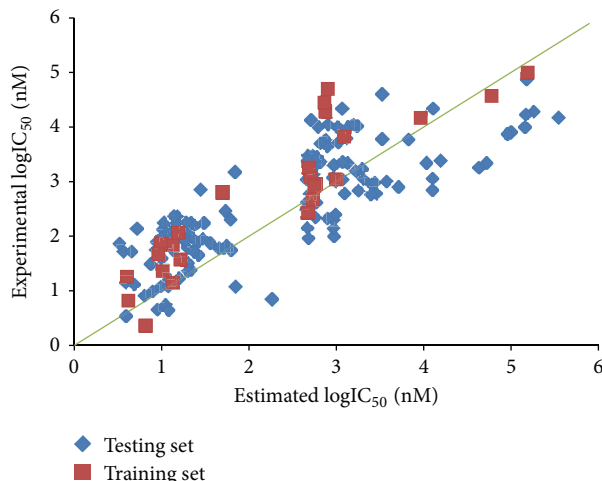


FIGURE 2: The scatter diagram of training set and testing set inhibitors.

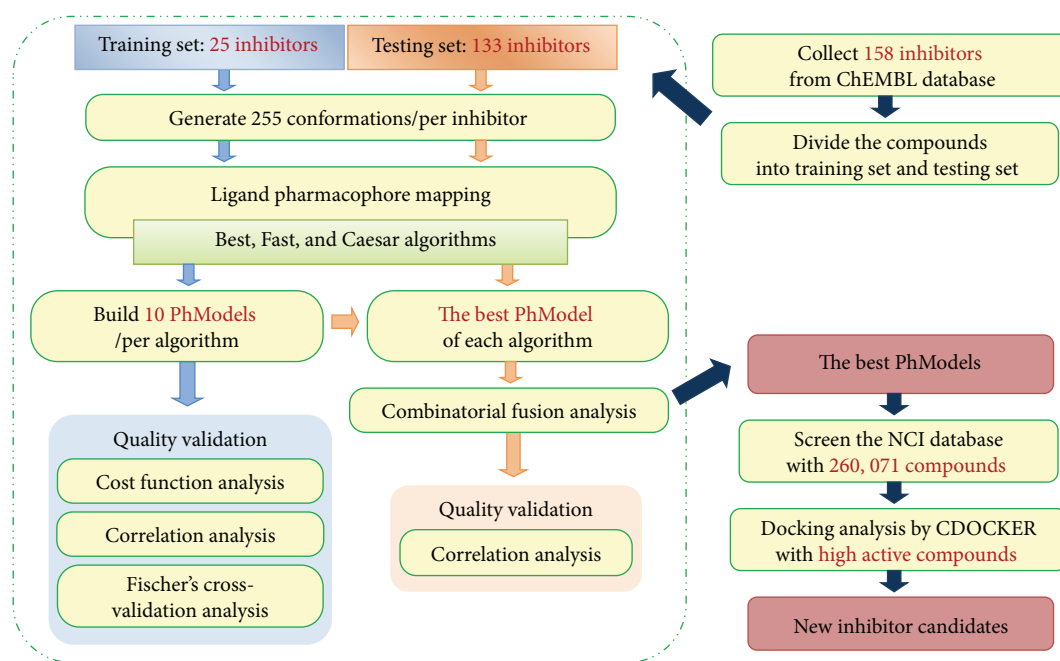


FIGURE 3: The workflow of PhModel generation for Chk2 inhibitors.

training set ( $r_{\text{train}}$ ). Then these best PhModels were used to predict the biological activities of testing set inhibitors by using HypoGen Best, Fast, and Caesar algorithms. Therefore, there are nine prediction results (denoted as  $Z_{\text{train}} \times Z_{\text{test}}$ ,  $Z = \{\text{Best, Fast, Caesar}\}$ , that is,  $\text{Best}_{\text{train}}\text{Best}_{\text{test}}$ ) generated for testing set inhibitors. Using data fusion, results from various prediction results are combined to obtain predictions with larger accuracy rate. The diversity rank/score function is used to select the most suitable prediction results for combination. If these three best PhModels were selected, there are nine prediction results and then there are  $2^9 - 1 = 511$  combinations. According to the rule (a) (1) in Remark 1, the  $r_{\text{train}}$  of  $\text{Casear}_{\text{train}}$  is far less than those of  $\text{Best}_{\text{train}}$  and  $\text{Fast}_{\text{train}}$  (Table 1); then, the  $\text{Casear}_{\text{train}}$  was not

considered in the combinations. Therefore, there are six prediction results ( $Z_{1\text{train}} \times Z_{2\text{test}}$ ,  $Z_1 = \{\text{Best, Fast}\}$  and  $Z_2 = \{\text{Best, Fast, Caesar}\}$ ) and  $2^6 - 1 = 63$  combinations. A special diversity rank/score graph was used to choose the best discriminating prediction results for further combination.

For an inhibitor  $p_i$  in the testing set  $P = \{p_1, p_2, \dots, p_t\}$  and the pair of prediction results  $A$  and  $B$ , the diversity score function  $d_i(A, B)$  is defined as  $d_i(A, B) = \sum |s_A - s_B|$ . When there are  $q$  prediction results selected (in this study,  $q = 6$ ), there are  $\binom{q}{2} = q(q-1)/2$  (in this study, the number is 15) diversity score functions. If we let  $i$  vary and fix the prediction result pair  $(A, B)$ , then  $d_i(A, B)$  is the diversity score function  $s_{(A,B)}$  from  $P = \{p_1, p_2, \dots, p_t\}$ . Sorting  $s_{(A,B)}$  into descending order would lead to the diversity



TABLE I: Experimental and estimated IC<sub>50</sub> values of training set inhibitors.

ChEMBL ID	Experimental IC <sub>50</sub> (nM)	Estimated IC <sub>50</sub> (nM)		
		Best <sub>train</sub>	Fast <sub>train</sub>	Caesar <sub>train</sub>
CHEMBL195041	2.3	15	9.9	1129
CHEMBL193990	6.6	6.8	6.2	942
CHEMBL248935	14	20	20	833
CHEMBL195320	18	8.5	6.2	942
CHEMBL176164	23	19	23	1151
CHEMBL250765	37	30	22	950
CHEMBL362677	47	23	23	1153
CHEMBL249959	70	110	20	1000
CHEMBL250992	72	47	6.9	72
CHEMBL251155	110	220	23	756
CHEMBL588536	270	670	790	78578
CHEMBL400772	470	2200	268	231
CHEMBL367390	640	2000	2237	1028
CHEMBL608262	830	1200	1456	94262
CHEMBL401105	900	1000	235	20
CHEMBL176115	1100	970	1044	1449
CHEMBL253542	1200	1100	189	3.8
CHEMBL592490	1800	860	1275	93360
CHEMBL589090	6700	1700	1419	3561
CHEMBL199299	15000	22000	233	1745
CHEMBL251629	19000	3600	615	411
CHEMBL259084	28000	6800	31827	5300
CHEMBL251628	37000	63000	1360	24786
CHEMBL438485	50000	16000	320	243
CHEMBL589501	100000	160000	48276	96926
Correlation coefficient ( $r_{\text{train}}$ )		0.955	0.840	0.238

rank function  $r_{(A,B)}$ . Consequently, the diversity rank/score function  $f_{(A,B)}$  is defined as  $f_{(A,B)} = (s_{(A,B)} \circ r_{(A,B)}^{-1})(j) = s_{(A,B)}(r_{(A,B)}^{-1}(j))$ , where  $j$  is in  $T = \{1, 2, 3, \dots, t\}$ . We note that the set  $T$  is different from the set  $P$  which is the testing set considered. The set  $T$  is used as the index set for the diversity rank function value and  $|T| = t$  is indeed the cardinality of  $P$ . The diversity rank/score function  $f_{(A,B)}$  so defined exhibits the diversity trend of the prediction result pair  $(A, B)$  across the whole spectrum of input set of  $t$  inhibitors and is independent of the specific inhibitor under study. For two prediction results  $A$  and  $B$ , the graph of the diversity rank/score function  $f_{(A,B)}(j)$  is called the diversity rank/score graph. This study aims to examine all the  $q(q-1)/2$  diversity rank/score graphs to see which pair of prediction results would give the larger diversity measurement according to the rule (a) (2) in Remark 1.

**2.5. Database Screen.** After examining 15 diversity rank/score graphs, the PhModels  $A$  and  $B$  determined from the best prediction result pair were used to screen the NCI database for new Chk2 inhibitor candidates. Under the PhModel, pharmacophore hypothesis screening can be used to screen small molecule database to retrieve the compounds as potential inhibitors that fit the pharmacophoric features.

In this study, the "Search 3D Database protocol" with the Best/Fast/Caesar Search option in Accelrys Discovery Studio 2.1 was employed to search the NCI database with 260,071 compounds. We could filter out and select the compounds in the NCI database based on the estimated activity and chemical features of PhModel.

**2.6. Molecular Docking.** After the database screening approach, the selected compounds can be further estimated according to the interaction energy between a receptor and a ligand through the molecular docking approach. In this study, selected compounds in the NCI database were docked into Chk2 active sites by CDOCKER docking program, and then their CDOCKER interaction energies were estimated. Finally, new potential candidates were retrieved from the NCI database with high interaction energy. The workflow of database screening and molecular docking approach was shown in Figure 4.

### 3. Results

**3.1. PhModel Generation Results.** Each of the ten PhModels using 25 training set inhibitors and HypoGen Best, Fast, and Caesar algorithms was generated by selecting hydrogen bond

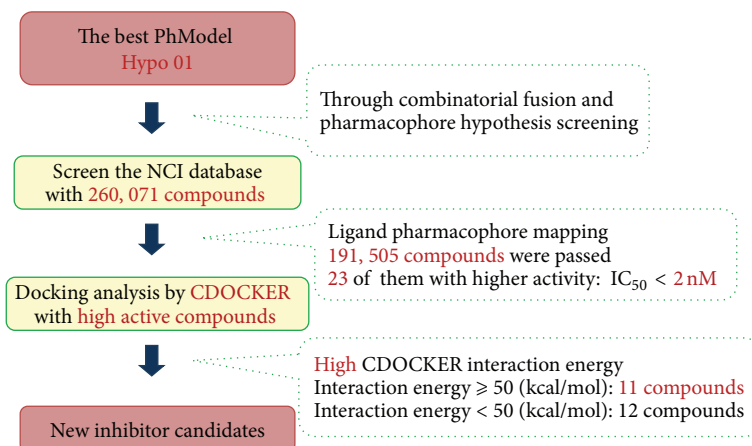


FIGURE 4: The workflow of database screening and molecular docking approach for new Chk2 inhibitor candidates.

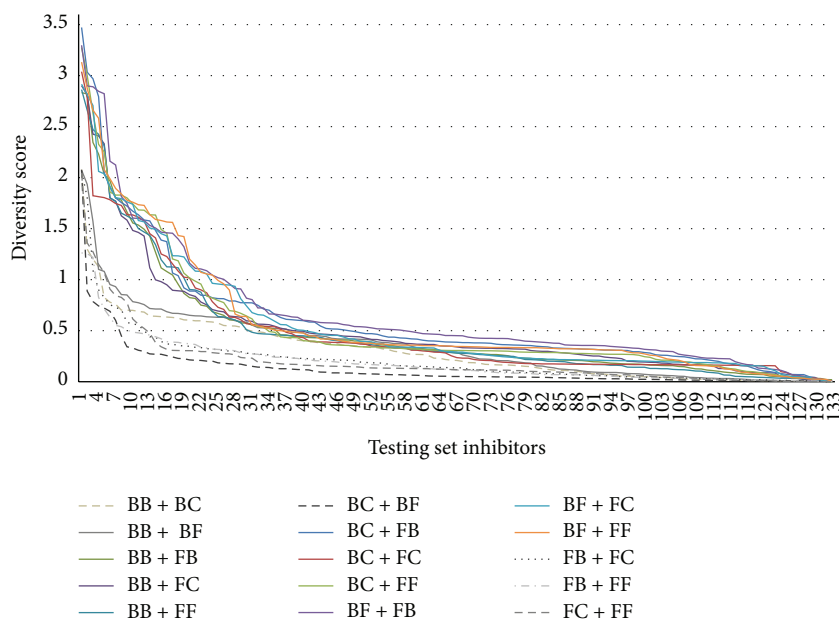


FIGURE 5: The diversity rank/score graphs for 15 combinations of prediction results.

acceptor (A), hydrogen bond donor (D), and hydrophobic (H) and hydrophobic aromatic (HYAR) features. Each of the best PhModels,  $Best_{train}$ ,  $Fast_{train}$ , and  $Casear_{train}$ , was evaluated with the best  $r_{train}$ , and the predicted biological activities of training set inhibitors and  $r_{train}$  were listed in Table 1, respectively. From Table 1, the  $Best_{train}$  obtained better  $r_{train}$  of value 0.955 than those by  $Fast_{train}$  and  $Casear_{train}$ . Moreover, the  $r_{train}$  of  $Casear_{train}$  is far less than those of  $Best_{train}$  and  $Fast_{train}$ . Hence, HypoGen Best algorithm was used individually to generate the PhModels for most of target genes in the past. According to rule (a) (1) in Remark 1, the  $Casear_{train}$  was not considered to be used for the prediction of testing set inhibitors.

**3.2. Correlation Analysis of Testing Set Inhibitors.** The testing set inhibitors were predicted by  $Best_{train}$  and  $Fast_{train}$  with HypoGen Best, Fast, and Caesar algorithms. Therefore, there

are six prediction results,  $Best_{train}Best_{test}$  (denoted as BB),  $Best_{train}Fast_{test}$  (denoted as BF),  $Best_{train}Casear_{test}$  (denoted as BC),  $Fast_{train}Best_{test}$  (denoted as FB),  $Fast_{train}Fast_{test}$  (denoted as FF), and  $Fast_{train}Casear_{test}$  (denoted as FC), for testing set inhibitors. The predicted biological activities of testing set inhibitors and  $r_{test}$  by these six prediction results were listed in Table 2, respectively. From Table 2, for the  $Best_{train}$ , the best  $r_{test}$  of value 0.81 was achieved by the  $Best_{train}Best_{test}$ ; for the  $Fast_{train}$ , the best  $r_{test}$  of value 0.728 was achieved by the  $Fast_{train}Fast_{test}$ . However, the  $Best_{train}Best_{test}$  obtained the best  $r_{test}$  in overall; moreover, the prediction results in the  $Best_{train}$  all outperform those in the  $Fast_{train}$ .

**3.3. Combinatorial Fusion Analysis.** Under the six prediction results, the diversity score function  $d_i(A, B)$  was calculated for each testing set inhibitor by a pair of prediction results (A, B). There are 15 diversity score functions  $s_{(A,B)}$  that were

TABLE 2: Experimental and estimated IC<sub>50</sub> values of testing set inhibitors.

ChEMBL ID	Experimental IC <sub>50</sub> (nM)	Estimated IC <sub>50</sub> (nM)						
		Best <sub>train</sub>		Caesar <sub>test</sub>	Fast <sub>train</sub>		Caesar <sub>test</sub>	Best <sub>train</sub> Best <sub>test</sub> + Fast <sub>train</sub> Fast <sub>test</sub>
		Best <sub>test</sub>	Fast <sub>test</sub>		Best <sub>test</sub>	Fast <sub>test</sub>		
CHEMBL195177	3.4	3.9	5.2	5.1	14.8	6.2	12.3	6.2
CHEMBL359881	4.4	12.0	46.1	42.3	17.3	22.5	29.4	22.5
CHEMBL179717	4.5	9.0	42.1	43.2	14.3	21.9	29.3	21.9
CHEMBL175553	5.5	11.2	57.8	43.1	17.6	20.5	29.1	20.5
CHEMBL192161	7	183.4	74.6	36.7	258.7	253.9	291.4	253.9
CHEMBL191969	8.2	6.4	15.3	14.3	10.5	9.4	13.5	9.4
CHEMBL175472	9.8	7.9	48.4	9.9	12.4	22.5	29.1	22.5
CHEMBL361378	12	9.9	48.1	43.7	14.7	21.7	29.3	21.7
CHEMBL362255	12	11.9	55.4	42.3	15.5	22.9	29.4	22.9
CHEMBL369254	12	70.2	51.3	45.6	38.0	23.7	31.5	23.7
CHEMBL364978	13	4.8	4.9	5.1	13.5	6.5	12.3	6.5
CHEMBL195846	14	3.9	5.0	5.1	12.8	6.2	12.3	6.2
CHEMBL179583	16	12.2	48.5	43.1	13.4	23.2	28.9	23.2
CHEMBL178972	17	15.9	52.5	43.1	18.0	23.3	29.1	23.3
CHEMBL250360	23	19.8	46.8	43.7	17.6	21.7	29.4	21.7
CHEMBL175879	24	21.6	57.0	42.3	24.0	23.2	29.2	23.2
CHEMBL179267	31	7.4	63.5	42.8	20.5	21.2	29.4	21.2
CHEMBL192022	32	20.3	47.5	42.6	23.9	22.4	29.4	22.4
CHEMBL250158	39	10.1	20.4	43.7	19.1	21.5	29.3	21.5
CHEMBL363339	41	10.1	59.8	42.2	18.1	24.0	29.5	24.0
CHEMBL250555	45	26.3	48.2	42.9	24.5	22.7	29.4	22.7
CHEMBL250359	52	3.7	5.2	3.5	20.6	7.0	28.8	7.0
CHEMBL251585	52	4.5	3.0	3.1	11.9	4.9	10.7	4.9
CHEMBL398529	53	23.0	48.4	43.5	21.4	22.4	29.4	22.4
CHEMBL178971	55	62.4	44.9	43.1	41.7	22.1	29.3	22.1
CHEMBL427879	55	13.7	45.4	42.4	16.9	19.9	29.5	19.9
CHEMBL250963	57	8.6	44.1	42.9	17.8	21.5	27.2	21.5
CHEMBL251170	60	51.3	45.3	43.1	26.3	21.5	29.4	21.5
CHEMBL250759	61	45.6	47.4	43.3	36.6	23.2	29.5	23.2
CHEMBL367263	61	19.7	50.3	9.6	17.3	23.9	29.1	23.9
CHEMBL250159	67	55.7	45.3	43.1	25.7	17.2	29.4	17.2
CHEMBL398467	70	11.4	46.5	42.7	21.0	20.4	29.4	20.4
CHEMBL250796	73	3.3	3.4	4.2	14.7	6.0	14.4	6.0
CHEMBL250957	74	36.1	44.9	43.7	30.5	20.4	29.5	20.4
CHEMBL206609	77	25.9	51.0	44.0	9.6	16.3	17.8	16.3
CHEMBL400755	78	8.8	26.6	43.0	12.9	22.5	29.4	22.5
CHEMBL249569	80	11.1	48.3	43.3	17.0	22.9	28.0	22.9
CHEMBL193397	81	9.9	43.8	42.8	14.5	21.2	28.4	21.2
CHEMBL438868	82	18.9	42.0	43.1	13.1	19.0	29.3	19.0
CHEMBL249566	86	17.6	48.8	43.1	23.5	22.8	29.6	22.8
CHEMBL249345	90	23.2	47.6	43.0	20.9	20.0	25.0	20.0
CHEMBL399146	90	29.8	47.9	42.7	28.4	22.1	29.3	22.1
CHEMBL602931	92	483.6	560.1	506.8	645.9	594.0	588.2	594.0
CHEMBL249347	95	10.1	46.3	43.1	27.2	20.1	29.2	20.1
CHEMBL193476	100	951.1	914.5	925.8	2435.2	1027.4	1981.9	1027.4

TABLE 2: Continued.

ChEMBL ID	Experimental IC <sub>50</sub> (nM)	Estimated IC <sub>50</sub> (nM)						
		Best <sub>train</sub>			Fast <sub>train</sub>			Best <sub>train</sub> Best <sub>test</sub> + Fast <sub>train</sub> Fast <sub>test</sub>
		Best <sub>test</sub>	Fast <sub>test</sub>	Caesar <sub>test</sub>	Best <sub>test</sub>	Fast <sub>test</sub>	Caesar <sub>test</sub>	
CHEMBL250361	100	14.2	6.6	14.5	14.9	20.9	29.6	20.9
CHEMBL248934	109	20.0	52.5	43.3	14.0	22.4	29.4	22.4
CHEMBL249750	110	12.7	49.3	43.5	20.7	21.6	29.3	21.6
CHEMBL208463	133	10.4	46.0	41.0	968.6	2768.3	2670.7	2768.3
CHEMBL250566	140	5.2	17.3	26.3	18.4	20.9	29.6	20.9
CHEMBL251256	140	936.4	2453.5	2240.6	192.6	215.1	215.5	215.1
CHEMBL437331	142	471.8	521.6	450.6	64.2	222.1	60.7	222.1
CHEMBL249541	157	23.8	42.5	43.0	16.0	18.8	29.3	18.8
CHEMBL249776	158	13.9	47.2	43.7	16.7	20.9	29.4	20.9
CHEMBL249350	174	30.6	51.2	43.4	22.5	23.3	29.4	23.3
CHEMBL249546	176	10.6	47.3	42.8	17.9	21.0	29.4	21.0
CHEMBL251364	176	21.7	43.9	43.1	17.8	19.7	29.1	19.7
CHEMBL399933	180	14.8	45.0	43.5	20.8	18.9	29.4	18.9
CHEMBL400287	180	19.3	45.5	43.5	21.5	19.8	29.4	19.8
CHEMBL175780	200	61.7	47.5	42.3	36.0	21.9	29.5	21.9
CHEMBL176326	200	935.6	926.3	913.6	1896.0	986.2	1980.7	986.2
CHEMBL590335	210	791.5	721.8	807.4	616.3	639.0	644.7	639.0
CHEMBL398561	220	575.1	926.2	571.5	209.1	278.3	287.9	278.3
CHEMBL249777	231	15.0	45.7	42.5	25.5	21.7	29.4	21.7
CHEMBL442282	233	13.9	46.9	42.3	21.5	22.5	29.4	22.5
CHEMBL195599	250	981.5	908.0	925.8	2266.7	1057.6	1981.9	1057.6
CHEMBL176015	290	54.1	52.8	54.4	24.0	29.3	28.0	29.3
CHEMBL251284	310	484.8	429.7	533.7	189.6	196.8	217.4	196.8
CHEMBL600441	310	454.6	559.9	516.5	300.5	513.5	254.1	513.5
CHEMBL599581	410	594.7	496.6	509.1	506.1	252.9	299.2	252.9
CHEMBL592784	420	462.2	492.0	475.1	216.7	203.8	198.3	203.8
CHEMBL1197465	580	2492.8	8163.9	5925.9	995.1	896.8	488.9	896.8
CHEMBL590809	600	492.4	539.0	534.3	816.4	536.2	549.3	536.2
CHEMBL1197456	610	2871.8	7537.6	6733.2	615.7	4139.2	3514.6	4139.2
CHEMBL590637	610	1251.9	1786.7	1121.1	2379.0	1650.9	1262.3	1650.9
CHEMBL591518	680	1797.9	1804.0	1516.5	6075.0	4714.4	2726.5	4714.4
CHEMBL598973	700	12585.6	151896.0	84151.4	1047.3	396.5	1318.1	396.5
CHEMBL251368	710	27.7	45.9	43.8	13.4	21.2	29.1	21.2
CHEMBL1197303	800	5153.6	36481.1	6594.4	1277.8	681.9	4077.9	681.9
CHEMBL1197320	890	2537.8	7191.1	6002.7	1559.4	517.5	420.0	517.5
CHEMBL1197528	960	2752.7	7737.1	5925.9	765.7	654.9	559.0	654.9
CHEMBL215803	1000	3760.6	140257.0	74847.0	9672.7	50053.5	49263.9	50053.5
CHEMBL253324	1000	996.2	2416.5	603.1	264.7	553.1	272.4	553.1
CHEMBL589347	1100	458.3	560.2	482.3	209.7	299.1	196.7	299.1
CHEMBL604784	1100	1188.4	1365.6	1205.7	2305.3	1962.3	1307.9	1962.3
CHEMBL1197529	1120	2047.4	11678.5	8090.0	3368.8	3648.8	3659.9	3648.8
CHEMBL1197326	1130	12645.5	45428.8	7432.7	1138.2	465.0	416.0	465.0
CHEMBL176041	1200	925.0	906.8	913.6	1933.6	1040.0	1980.7	1040.0
CHEMBL590079	1350	548.9	548.5	550.6	1000.9	860.3	855.1	860.3
CHEMBL605083	1400	489.2	613.8	554.4	1224.9	1401.0	1345.6	1401.0
CHEMBL175481	1500	69.2	57.7	49.2	1853.9	1526.1	1852.1	1526.1

TABLE 2: Continued.

ChEMBL ID	Experimental IC <sub>50</sub> (nM)	Estimated IC <sub>50</sub> (nM)						
		Best <sub>train</sub>			Fast <sub>train</sub>			Best <sub>train</sub> Best <sub>test</sub> + Fast <sub>train</sub> Fast <sub>test</sub>
		Best <sub>test</sub>	Fast <sub>test</sub>	Caesar <sub>test</sub>	Best <sub>test</sub>	Fast <sub>test</sub>	Caesar <sub>test</sub>	
CHEMBL205906	1540	1696.1	1109.3	1153.0	980.6	17655.5	937.5	17655.5
CHEMBL590808	1600	1662.5	1799.1	1709.4	51882.2	48813.2	48302.3	48813.2
CHEMBL1170748	1700	1974.2	1940.4	1652.8	3251.4	2734.9	407.6	2734.9
CHEMBL253541	1800	43066.9	42831.1	79622.8	1255.9	1099.1	15566.3	1099.1
CHEMBL176554	1900	572.3	7226.3	9126.2	1096.1	3352.6	3926.3	3352.6
CHEMBL377597	2000	938.5	6719.7	5644.6	50.4	116.9	105.4	116.9
CHEMBL1170749	2200	10874.0	30873.8	3995.6	5835.1	25044.3	1090.5	25044.3
CHEMBL590336	2200	507.8	594.0	516.4	568.9	544.3	521.1	544.3
CHEMBL590807	2200	52944.6	70480.3	52639.4	44898.4	49196.7	36195.2	49196.7
CHEMBL600868	2200	1342.8	1603.8	1312.9	7373.6	5797.7	4864.3	5797.7
CHEMBL398759	2300	666.0	2847.2	1010.2	447.2	924.3	309.0	924.3
CHEMBL604459	2300	1176.9	2299.2	1369.3	614.2	877.8	682.5	877.8
CHEMBL179383	2400	15724.9	14621.7	14257.4	4508.5	4132.8	4023.9	4132.8
CHEMBL592489	2400	469.2	494.6	486.9	204.7	240.5	208.4	240.5
CHEMBL425904	2800	605.0	651.0	534.9	478.1	451.9	488.6	451.9
CHEMBL150894	3000	537.8	844.9	797.6	186.8	600.2	289.0	600.2
CHEMBL590793	3000	475.4	2480.4	1251.5	192.0	251.7	205.4	251.7
CHEMBL600865	4400	796.9	2072.3	1223.3	198.1	261.3	228.1	261.3
CHEMBL249253	5000	659.8	503.6	499.5	406.6	227.9	309.5	227.9
CHEMBL587506	5200	1050.8	1293.3	1083.2	2201.3	2162.7	1656.0	2162.7
CHEMBL204930	5800	755.5	1260.7	1105.5	47881.6	47912.3	47868.0	47912.3
CHEMBL554900	5900	6722.8	580526.0	807309.0	745.4	428.2	47897.1	428.2
CHEMBL176276	6000	3360.2	8955.9	8476.6	2075.6	2054.0	1970.7	2054.0
CHEMBL589091	6100	1319.5	1308.4	1172.0	579.2	603.1	600.9	603.1
CHEMBL559781	7400	91783.7	822917.0	1230010.0	414.5	6865.5	47911.9	6865.5
CHEMBL249252	8000	100671.0	38586.8	31129.8	590.5	491.6	467.9	491.6
CHEMBL589089	9800	1070.7	1321.5	1002.7	38589.3	48136.2	41143.2	48136.2
CHEMBL217090	10000	628.8	1010.9	1548.5	390.6	451.8	910.1	451.8
CHEMBL217092	10000	1030.2	1199.3	1908.8	379.9	625.6	707.2	625.6
CHEMBL382588	10000	1365.2	5880.3	5488.7	2560.7	3993.4	3649.4	3993.4
CHEMBL590581	10000	145206.0	149922.0	108067.0	50609.5	49348.5	48302.3	49348.5
CHEMBL242753	10300	1742.5	3393.1	1941.1	1202.5	2613.0	1297.3	2613.0
CHEMBL398758	11000	1582.8	187965.0	1573.6	237.0	4333.9	369.0	4333.9
CHEMBL399151	11000	812.1	1263.5	2069.9	334.5	1193.9	1271.6	1193.9
CHEMBL395080	13450	513.5	484.0	458.6	198.5	188.9	183.3	188.9
CHEMBL1171533	15000	349604.0	296515.0	159368.0	26784.5	26699.3	45506.4	26699.3
CHEMBL602729	17000	148756.0	149052.0	139465.0	224.1	324.7	229.4	324.7
CHEMBL249255	19000	182486.0	42615.8	41578.3	1698.2	616.1	683.9	616.1
CHEMBL202930	21730	12828.4	11945.6	12145.8	213.3	219.7	217.6	219.7
CHEMBL589986	22000	1167.1	1337.4	1143.5	53747.5	49155.6	49316.8	49155.6
CHEMBL251471	40000	3358.9	1946.2	2075.7	512.2	423.7	3427.4	423.7
CHEMBL560056	74000	152006.0	156723.0	208466.0	223.2	208.1	190.4	208.1
Correlation coefficient ( $r_{test}$ )		0.810	0.771	0.783	0.710	0.728	0.714	<b>0.816</b>

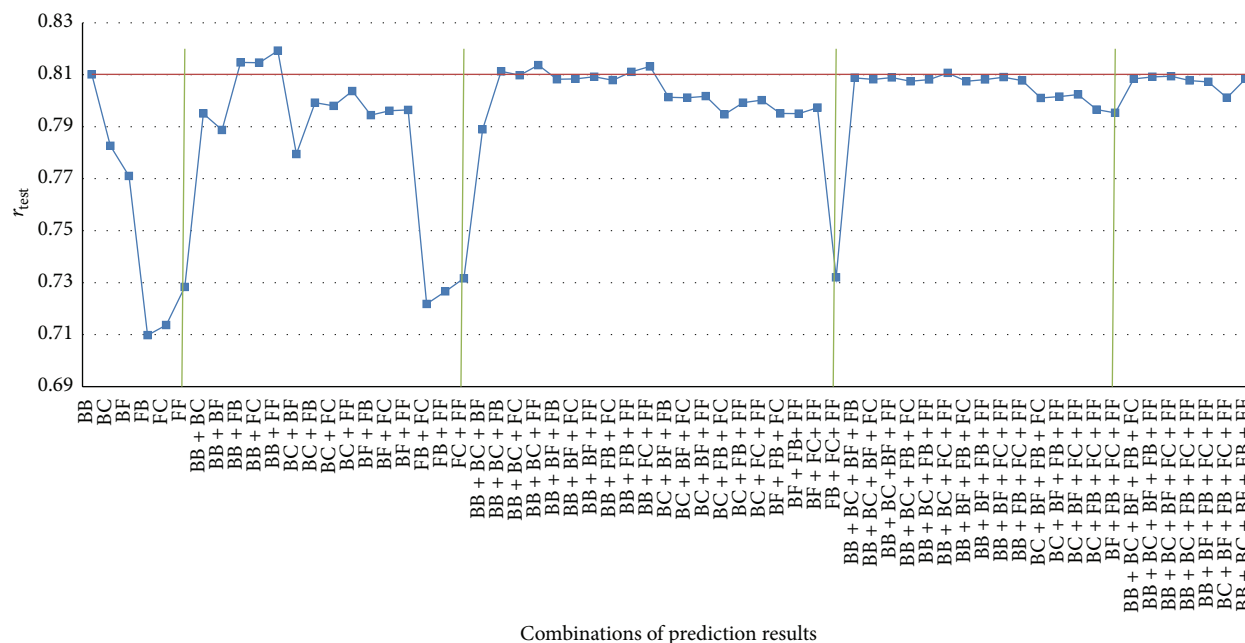


FIGURE 6: The  $r_{\text{test}}$  for all of 63 combinations from six prediction results.

performed at first and then these diversity score functions were sorted to become the diversity rank function  $r_{(A,B)}$ , respectively. Finally, 15 diversity rank/score functions  $f_{(A,B)}$  were represented as diversity rank/score graphs shown in Figure 5. Among 15 diversity rank/score graphs, there are several combinations (gray color) that have less diversity scores than those by others, such as BB + BC, BB + BF, and FB + FB, shown in Figure 5. It means that these combinations may have less  $r_{\text{test}}$  than those by others according to rule (a) (2) in Remark 1. In other words, several combinations, such as BB + FC (purple color), BB + FF (blue color), and BF + FF (orange color), may have larger  $r_{\text{test}}$  than those by others due to larger diversity scores. For the six prediction results, all of the 63 combinations were preformed and evaluated by its  $r_{\text{test}}$ , respectively, as shown in Figure 6. In Figure 6, for 15 pairs of two prediction results, the combinations BB + FB, BB + FC, and BB + FF have larger  $r_{\text{test}}$  than those by others. Moreover, the combination BB + FF has best  $r_{\text{test}}$  of value 0.816 among 15 combinations, even for 63 combinations. Besides, the average  $r_{\text{test}}$  by the combinations is larger than the individual prediction results. It means that the predictive accuracy for Chk2 inhibitors may be improved by considering the Best<sub>train</sub> and Fast<sub>train</sub> concurrently.

**3.4. Database Screen Results.** The best PhModels, Best<sub>train</sub> and Fast<sub>train</sub>, were used to screen the NCI database with 260,071 compounds for new Chk2 inhibitor candidates by using HypoGen Best and Fast algorithms, respectively. The Best<sub>train</sub>, Best<sub>test</sub> and Fast<sub>train</sub>, Fast<sub>test</sub> prediction results for NCI database were combined in order to filter out possible false positive candidates. Of the 260,071 compounds, 191,505 passed the screening and best fitted to the chemical features in 3D space. 23 drug-like compounds that had an estimated IC<sub>50</sub>

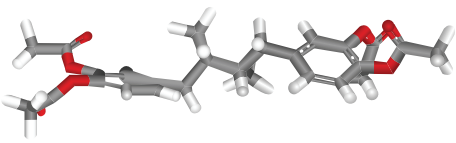
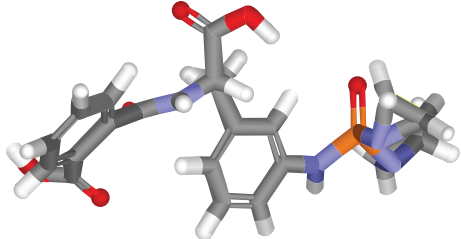
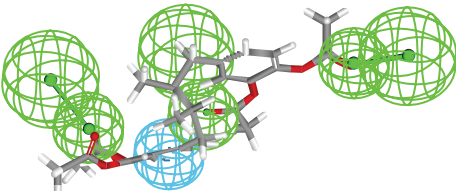
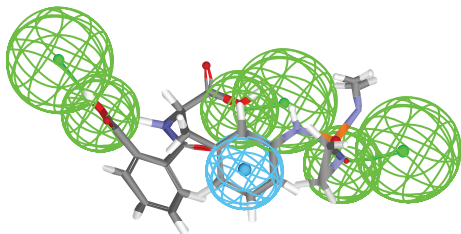
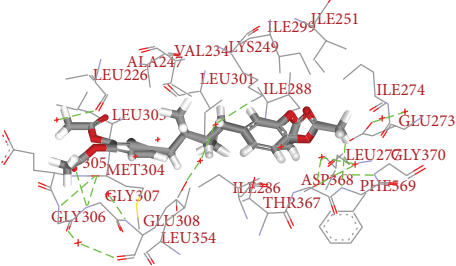
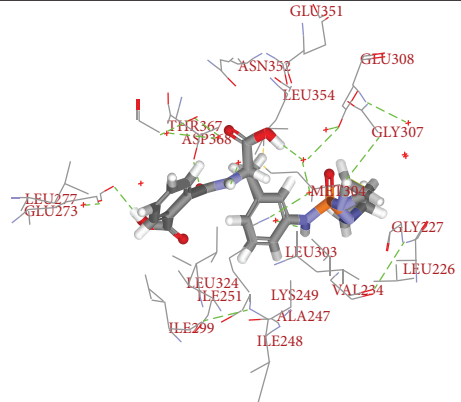
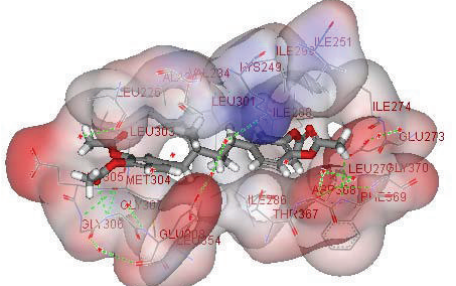
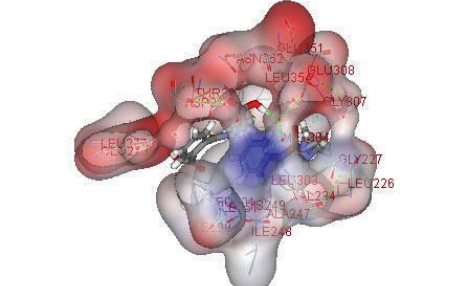
TABLE 3: The 21 drug-like compounds with their estimated IC<sub>50</sub> values and CDOCKER interaction energy greater than 37.786 (kal/mol).

Name	Estimated IC <sub>50</sub> (nM)	Interaction energy (kal/mol)
<b>NSC 136954</b>	<b>1.989</b>	<b>61.239</b>
<b>NSC 70804</b>	<b>1.682</b>	<b>58.967</b>
NSC 158029	1.885	57.944
NSC 603427	1.87	56.963
NSC 57782	1.6855	56.54
NSC 16739	1.5385	56.342
NSC 720227	1.914	55.839
NSC 618702	1.862	55.196
NSC 195178	1.7015	51.351
NSC 653142	1.557	51.19
NSC 653143	1.577	50.055
NSC 32200	1.901	49.439
NSC 342015	1.6515	47.327
NSC 343685	1.7615	46.436
NSC 205750	1.875	45.542
NSC 96538	1.705	44.344
NSC 210455	1.7935	42.258
NSC 314654	1.947	42.082
NSC 179894	1.6135	41.707
NSC 91710	1.701	40.533
NSC 370907	1.8785	40.502

value of less than 2 nM were studied in a molecular docking study (Figure 4).

**3.5. Molecular Docking Results.** 23 drug-like compounds along with the training set compounds were docked into the

TABLE 4: The structures and characteristics of the top 2 compounds.

	NSC 136954	NSC 70804
Structure		
Superposition		
Binding sites		
Docking results		

active sites that were defined based on the bound inhibitor, PV1019, in a crystal structure of Chk2 (PDB: 2W7X). We used CDOCKER program to confirm that inhibitor candidates bind to the receptor. CDOCKER uses molecular dynamics (MD) in conjunction with the CHARMM force field to individually dock the compounds into the binding sites. The coordinates of Chk2 from the Chk2/PV1019 crystal structure were used after removing PV1019 and solvent molecules and adding protein hydrogen atoms. After docking each screened compound, its interaction energy value was calculated. The PV1019 was redocked into the Chk2 binding site by the CDOCKER program. Its CDOCKER interaction energy was calculated by CDOCKER and determined to be

37.786 (kal/mol). The 23 drug-like compounds were docked into the Chk2 binding sites. Finally, there are 21 drug-like compounds with CDOCKER interaction energies greater than 37.786 (kal/mol). In addition, 11 drug-like compounds had high interaction value greater than 50 (kal/mol) (Figure 4) and the top 2 are NSC136954 with 61.239 (kal/mol) and NSC70804 with 58.967 (kal/mol), respectively, kept for future characterization as inhibitors. The 21 drug-like compounds with their estimated  $IC_{50}$  values and CDOCKER interaction energy greater than 37.786 (kal/mol) were shown in Table 3.

The structures and characteristics of the top 2 compounds are given in Table 4, and we can find that some active site residues were identified from the Chk2 complex. The

interaction sites of NSC136954 were Leu226, Val234, Ala247, Lys249, Ile251, Glu273, Ile274, Leu277, Ile286, Ile288, Ile299, Leu301, Leu303, Met304, Glu305, Gly306, Gly307, Glu308, Leu354, Thr367, Asp368, Phe369, and Gly370. On the other hand, the interaction sites of NSC70804 were Leu226, Leu227, Val234, Ala247, Ile248, Lys249, Ile251, Glu273, Leu277, Ile299, Leu301, Leu303, Met304, Gly307, Glu308, Glu351, Asn352, Leu354, Thr367, and Asp368. Several studies indicated that they are involved in hydrophobic interactions with Val234, Ile251, Leu354, Ile299, and the aliphatic portions of the side chains of Lys249, Thr367, and Asp368, in addition to several interactions of van der Waals or hydrophobic with Leu226, Val234, Leu303, Gly307, Leu354, and the aliphatic portions the side chains of Met304 and Glu308 [10, 11]. Furthermore, the quinazoline was sandwiched between the lipophilic side chains of Val234 and Leu354, with the side chains of Ala247, Leu301, and Leu303 also contributing to a hydrophobic surface surrounding the core and an interaction between the pyrazole and Lys249 is likely to account for the increase in Chk2 potency [12]. And residue Thr367 of Chk2 is a serine in Chk1. Portions of the glycine-rich P-loop in Chk2, which is located directly above the inhibitor, are disordered (residues 229–231), whereas this loop is well defined in the structure of Chk1, and Leu301 in Chk2 corresponds to the “gatekeeper” residue in many kinases, which has been found to form contacts with bound inhibitors and is poorly conserved [44].

#### 4. Conclusions

In this study, a novel design strategy for drug design was proposed to apply combinatorial fusion into PhModels and virtual screening techniques. 158 Chk2 inhibitors were divided into the training set and testing set, respectively. For 25 training set inhibitors, three best PhModels, Best<sub>train</sub>, Fast<sub>train</sub>, and Casear<sub>train</sub>, were generated at first, and then six prediction results for 133 testing set inhibitors were used for calculating 15 diversity rank/score functions. Finally, the combination Best<sub>train</sub>Best<sub>test</sub> and Fast<sub>train</sub>Fast<sub>test</sub> prediction results achieved the best  $r_{\text{test}}$  of value 0.816 among 63 combinations. Through these approaches, 23 potential Chk2 inhibitors with IC<sub>50</sub> value less than 2 nM and interaction energy value larger than 37.786 (kal/mol) are retrieved from NCI database. This study can help medicinal chemists to identify or design new Chk2 inhibitors. Besides, the potential inhibitors of Chk2 retrieved in this work can be estimated by biologists for further study.

#### 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 Science Council of Taiwan (under Grants NSC100-2221-E-182-057-MY3) and by Chang Gung Memorial Hospital (Grant CMRPD260033). The authors thank the National Center for

High-Performance Computing for computer time and use of its facilities.

#### References

- [1] K. M. Culligan, C. E. Robertson, J. Foreman, P. Doerner, and A. B. Britt, “ATR and ATM play both distinct and additive roles in response to ionizing radiation,” *Plant Journal*, vol. 48, no. 6, pp. 947–961, 2006.
- [2] J. Yang, Z.-P. Xu, Y. Huang, H. E. Hamrick, P. J. Duerksen-Hughes, and Y.-N. Yu, “ATM and ATR: sensing DNA damage,” *World Journal of Gastroenterology*, vol. 10, no. 2, pp. 155–160, 2004.
- [3] Q. Liu, S. Guntuku, X.-S. Cui et al., “Chk1 is an essential kinase that is regulated by Atr and required for the G2/M DNA damage checkpoint,” *Genes and Development*, vol. 14, no. 12, pp. 1448–1459, 2000.
- [4] C. Tapia-Alveal, T. M. Calonge, and M. J. O’Connell, “Regulation of Chk1,” *Cell Division*, vol. 4, article 8, 2009.
- [5] J. Bartek, J. Falck, and J. Lukas, “Chk2 kinase—a busy messenger,” *Nature Reviews Molecular Cell Biology*, vol. 2, no. 12, pp. 877–886, 2001.
- [6] J. Li, B. L. Williams, L. F. Haire et al., “Structural and functional versatility of the FHA domain in DNA-damage signaling by the tumor suppressor kinase Chk2,” *Molecular Cell*, vol. 9, no. 5, pp. 1045–1054, 2002.
- [7] A. W. Oliver, A. Paul, K. J. Boxall et al., “Trans-activation of the DNA-damage signalling protein kinase Chk2 by T-loop exchange,” *EMBO Journal*, vol. 25, no. 13, pp. 3179–3190, 2006.
- [8] Z. Cai, N. H. Chehab, and N. P. Pavletich, “Structure and activation mechanism of the CHK2 DNA damage checkpoint kinase,” *Molecular Cell*, vol. 35, no. 6, pp. 818–829, 2009.
- [9] Z. A. Stewart and J. A. Pietsenpol, “p53 signaling and cell cycle checkpoints,” *Chemical Research in Toxicology*, vol. 14, no. 3, pp. 243–263, 2001.
- [10] G. T. Lountos, A. G. Jobson, J. E. Tropea et al., “Structural characterization of inhibitor complexes with checkpoint kinase 2 (Chk2), a drug target for cancer therapy,” *Journal of Structural Biology*, vol. 176, no. 3, pp. 292–301, 2011.
- [11] A. G. Jobson, G. T. Lountos, P. L. Lorenzi et al., “Cellular inhibition of checkpoint kinase 2 (Chk2) and potentiation of camptothecins and radiation by the novel Chk2 inhibitor PV1019 [7-nitro-1H-indole-2-carboxylic acid 4-[1-(guanidinohydrazono)ethyl]-phenyl-amide],” *Journal of Pharmacology and Experimental Therapeutics*, vol. 331, no. 3, pp. 816–826, 2009.
- [12] J. J. Caldwell, E. J. Welsh, C. Matijssen et al., “Structure-based design of potent and selective 2-(quinazolin-2-yl)phenol inhibitors of checkpoint kinase 2,” *Journal of Medicinal Chemistry*, vol. 54, no. 2, pp. 580–590, 2011.
- [13] G. T. Lountos, A. G. Jobson, J. E. Tropea et al., “X-ray structures of checkpoint kinase 2 in complex with inhibitors that target its gatekeeper-dependent hydrophobic pocket,” *FEBS Letters*, vol. 585, no. 20, pp. 3245–3249, 2011.
- [14] S. Hilton, S. Naud, J. J. Caldwell et al., “Corrigendum to ‘Identification and characterisation of 2-aminopyridine inhibitors of checkpoint kinase 2’,” *Bioorganic and Medicinal Chemistry*, vol. 18, no. 12, p. 4591, 2010.
- [15] E. M. Rosen, S. Fan, R. G. Pestell, and I. D. Goldberg, “BRCA1 gene in breast cancer,” *Journal of Cellular Physiology*, vol. 196, no. 1, pp. 19–41, 2003.



- [16] E. S. Yang and F. Xia, "BRCA1 16 years later: DNA damage-induced BRCA1 shuttling," *FEBS Journal*, vol. 277, no. 15, pp. 3079–3085, 2010.
- [17] M. Gupta, S. Gupta, H. Dureja, and A. K. Madan, "Superaugmented eccentric distance sum connectivity indices: Novel highly discriminating topological descriptors for QSAR/QSPR," *Chemical Biology and Drug Design*, vol. 79, no. 1, pp. 38–52, 2012.
- [18] F. A. Pasha, M. Muddassar, and S. Joo Cho, "Molecular docking and 3D QSAR studies of Chk2 inhibitors," *Chemical Biology and Drug Design*, vol. 73, no. 3, pp. 292–300, 2009.
- [19] H. Kubinyi, G. Folkers, and Y. C. Martin, *3D QSAR in Drug Design*, Springer, 2002.
- [20] Y.-K. Jiang, "Molecular docking and 3D-QSAR studies on  $\beta$ -phenylalanine derivatives as dipeptidyl peptidase IV inhibitors," *Journal of Molecular Modeling*, vol. 16, no. 7, pp. 1239–1249, 2010.
- [21] R. R. S. Pissurlenkar, M. S. Shaikh, and E. C. Coutinho, "3D-QSAR studies of Dipeptidyl peptidase IV inhibitors using a docking based alignment," *Journal of Molecular Modeling*, vol. 13, no. 10, pp. 1047–1071, 2007.
- [22] W. Sippl, *3D-QSAR—Applications, Recent Advances, and Limitations. Recent Advances in QSAR Studies*, Springer, 2010.
- [23] A. Lauria, M. Ippolito, M. Fazzari et al., "IKK- $\beta$  inhibitors: an analysis of drug-receptor interaction by using Molecular Docking and Pharmacophore 3D-QSAR approaches," *Journal of Molecular Graphics and Modelling*, vol. 29, no. 1, pp. 72–81, 2010.
- [24] S. John, S. Thangapandian, M. Arooj, J. C. Hong, K. D. Kim, and K. W. Lee, "Development, evaluation and application of 3D QSAR Pharmacophore model in the discovery of potential human renin inhibitors," *BMC Bioinformatics*, vol. 12, pp. 1–14, 2011.
- [25] H. Kubinyi and Comparative Molecular Field Analysis (CoMFA), *Handbook of Chemoinformatics: From Data To Knowledge in 4 Volumes*, Wiley, 2008.
- [26] K. W. Lee and J. M. Briggs, "Comparative molecular field analysis (CoMFA) study of epothilones-tubulin depolymerization inhibitors: pharmacophore development using 3D QSAR methods," *Journal of Computer-Aided Molecular Design*, vol. 15, no. 1, pp. 41–55, 2001.
- [27] S. Durdagi, T. Mavromoustakos, and M. G. Papadopoulos, "3D QSAR CoMFA/CoMSIA, molecular docking and molecular dynamics studies of fullerene-based HIV-1 PR inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 18, no. 23, pp. 6283–6289, 2008.
- [28] M.-E. Suh, S.-Y. Park, and H.-J. Lee, "Comparison of QSAR methods (CoMFA, CoMSIA, HQSAR) of anticancer 1-N-substituted imidazoquinoline-4,9-dione derivatives," *Bulletin of the Korean Chemical Society*, vol. 23, no. 3, pp. 417–422, 2002.
- [29] S. J. Bang and S. J. Cho, "Comparative molecular field analysis (CoMFA) and comparative molecular similarity index analysis (CoMSIA) study of mutagen X," *Bulletin of the Korean Chemical Society*, vol. 25, no. 10, pp. 1525–1530, 2004.
- [30] L. Ghemtio, Y. Zhang, and H. Xhaard, *Virtual Screening*, InTech, 2012.
- [31] G. Klebe, *3D QSAR in Drug Design*, vol. 3, Springer, 2002.
- [32] G. Klebe and U. Abraham, "Comparative Molecular Similarity Index Analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries," *Journal of Computer-Aided Molecular Design*, vol. 13, no. 1, pp. 1–10, 1999.
- [33] I. Mitra, A. Saha, and K. Roy, "Pharmacophore mapping of arylamino-substituted benzo[b]thiophenes as free radical scavengers," *Journal of Molecular Modeling*, vol. 16, no. 10, pp. 1585–1596, 2010.
- [34] K. Boppana, P. K. Dubey, S. A. R. P. Jagarlapudi, S. Vadivelan, and G. Rambabu, "Knowledge based identification of MAO-B selective inhibitors using pharmacophore and structure based virtual screening models," *European Journal of Medicinal Chemistry*, vol. 44, no. 9, pp. 3584–3590, 2009.
- [35] M. Chopra, R. Gupta, S. Gupta, and D. Saluja, "Molecular modeling study on chemically diverse series of cyclooxygenase-2 selective inhibitors: generation of predictive pharmacophore model using Catalyst," *Journal of Molecular Modeling*, vol. 14, no. 11, pp. 1087–1099, 2008.
- [36] A. Gaulton, L. J. Bellis, A. P. Bento et al., "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Research*, vol. 40, pp. D1100–D1107, 2012.
- [37] K. C. Shih, C. W. Shiau, T. S. Chen et al., "Automated chemical hypothesis generation and database searching with Catalyst," *Perspectives in Drug Discovery and Design*, vol. 3, pp. 1–20, 1995.
- [38] D. F. Hsu and I. Taksa, "Comparing rank and score combination methods for data fusion in information retrieval," *Information Retrieval*, vol. 8, no. 3, pp. 449–480, 2005.
- [39] C. C. Vogt and G. W. Cotrell, "Fusion via a linear combination of scores," *Information Retrieval*, vol. 1, no. 3, pp. 151–172, 1999.
- [40] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 3, pp. 418–435, 1992.
- [41] J.-M. Yang, Y.-F. Chen, T.-W. Shen, B. S. Kristal, and D. F. Hsu, "Consensus scoring criteria for improving enrichment in virtual screening," *Journal of Chemical Information and Modeling*, vol. 45, no. 4, pp. 1134–1146, 2005.
- [42] M. A. Kuriakose, W. T. Chen, Z. M. He et al., "Selection and validation of differentially expressed genes in head and neck cancer," *Cellular and Molecular Life Sciences*, vol. 61, no. 11, pp. 1372–1383, 2004.
- [43] K.-L. Lin, C.-Y. Lin, C.-D. Huang et al., "Feature selection and combination criteria for improving accuracy in protein structure prediction," *IEEE Transactions on Nanobioscience*, vol. 6, no. 2, pp. 186–196, 2007.
- [44] G. T. Lountos, J. E. Tropea, D. Zhang et al., "Crystal structure of checkpoint kinase 2 in complex with NSC 109555, a potent and selective inhibitor," *Protein Science*, vol. 18, no. 1, pp. 92–100, 2009.

## Research Article

# New Strategies for Evaluation and Analysis of SELEX Experiments

Rico Beier,<sup>1</sup> Elke Boschke,<sup>2</sup> and Dirk Labudde<sup>1</sup>

<sup>1</sup> *Bioinformatics Group, Department of Mathematics, Natural and Computer Sciences, University of Applied Sciences Mittweida, 09648 Mittweida, Germany*

<sup>2</sup> *Institute of Food Technology and Bioprocess Engineering, Department of Mechanical Engineering, Dresden University of Technology, 01062 Dresden, Germany*

Correspondence should be addressed to Rico Beier; [rbeier1@hs-mittweida.de](mailto:rbeier1@hs-mittweida.de)

Received 4 October 2013; Accepted 28 January 2014; Published 19 March 2014

Academic Editor: Chun-Yuan Lin

Copyright © 2014 Rico Beier et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aptamers are an interesting alternative to antibodies in pharmaceuticals and biosensorics, because they are able to bind to a multitude of possible target molecules with high affinity. Therefore the process of finding such aptamers, which is commonly a SELEX screening process, becomes crucial. The standard SELEX procedure schedules the validation of certain found aptamers via binding experiments, which is not leading to any detailed specification of the aptamer enrichment during the screening. For the purpose of advanced analysis of the accrued enrichment within the SELEX library we used sequence information gathered by next generation sequencing techniques in addition to the standard SELEX procedure. As sequence motifs are one possibility of enrichment description, the need of finding those recurring sequence motifs corresponding to substructures within the aptamers, which are characteristically fitted to specific binding sites of the target, arises. In this paper a motif search algorithm is presented, which helps to describe the aptamers enrichment in more detail. The extensive characterization of target and binding aptamers may later reveal a functional connection between these molecules, which can be modeled and used to optimize future SELEX runs in case of the generation of target-specific starting libraries.

## 1. Introduction

The inhibition of protein interactions, such as receptor-ligand interactions or the interplay during pathogen infections, is one main functional principle of therapeutics to influence biologically relevant processes. In this context usually antibodies are used to bind to specific target proteins and thus wield biological influence. Although antibodies and corresponding technologies are widely distributed, they are accompanied with some major drawbacks. A first hindrance is the antibody's large size that limits the access to smaller biological compartments and thus also its bioavailability. It is also problematic that antibodies are often immunogenic and cannot be used after their denaturation. If we consider the production process of antibodies, it becomes apparent that this process is difficult to scale up and susceptible to bacterial or viral contamination [1, 2]. The need of finding other target-binding molecules as alternatives for antibodies

draws the attention now to another surrogate, the aptamer, which is also qualified for target binding [1].

These aptamers are short and stable, single-stranded nucleotide oligomers folding into complex three-dimensional structures. They are composed of helical parts and different variants of loops like hairpins, inner loops, bulges, and junctions, which allow branching of the structure. Unpaired nucleotides have a higher potential to take part in intermolecular, noncovalent chemical bonding via hydrogen bonds, hydrophobic, and electrostatic interactions on the nucleotides preferred binding sites [3]. Aptamers can target a diverse multitude of particles from small molecules like organic dyes [4] and amino acids [5] and larger molecules like antibiotics [6] and proteins [7] as well as whole cell surfaces [8]. The focus on therapeutically applied aptamers lies especially on proteins as target molecules. Notably, in respect of binding affinity they are comparable to antibodies. While a study has shown that an aptamer with an affinity of

$K_d = 50$  pM could be found for vascular endothelial growth factor as target, an antibody for the same target in comparison shows an affinity of  $K_d = 54$  pM [9, 10]. Furthermore there is growing evidence of a connection between regions of unpaired nucleotides and the concrete biological function of RNA molecules. This can analogously be assumed for DNA aptamers [11].

Since the production process of aptamers is purely chemical, it is readily scalable and less prone to bacterial or viral contamination, which poses an advantage over artificial synthesis of antibodies [1, 2]. The resulting aptamers are usually not immunogenic and smaller in size, which allows a less elaborate administration of aptamer based medication [12]. Although the aptamer denaturation is reversible, their half-life is limited by nuclease degradation. This vulnerability can only be opposed by chemical modification of the aptamers [1]. In summary, aptamers are an attractive alternative to antibodies and will lead to new issues in the fields of bioinformatics.

With the introduction of next generation sequencing (NGS) technologies it is possible to massively parallelize the sequencing process. That makes it easy to gather large amounts of sequence data in relatively short periods of time [13]. In this manner the NGS technology can be used for genome sequencing to speed up and enhance the shotgun sequencing. But that is not the only use of NGS. The sequencing technology is also applicable in fields of aptamer research, especially in the process of finding high affinity aptamers for a desired target molecule. Caused by the high complexity of the conformational space of aptamers it is a hard problem to find target-binding aptamers. Commonly a screening technology needs to be utilized to find these unique aptamers that are capable of binding to a specific target molecule. This technique is called SELEX (Systematic Evolution of Ligands by Exponential Enrichment) [14]. During the multiple steps of the experimental process there are several opportunities for performing NGS to gather sequence data useful for the purpose of later analysis.

The SELEX screening process starts with a chemically synthesized, random library of nucleotide oligomers of a fixed size. Although the size of this starting library is fairly large with a range of typically  $10^{13}$  to  $10^{16}$ , it can in practice only cover a small fraction of the possible sequence and structure space, because these spaces are growing exponentially with the desired aptamers lengths. Based on this library multiple subsequent selection rounds are performed, in which library and target molecules are incubated. As the multitude of aptamers contained in a rounds library is competing for the fewer binding sites available on the relatively small number of target molecules added, the arising selection pressure leads to the preferred binding of the highest affinity oligonucleotides of the library. Commonly some experimental parameters are adjusted during the execution to increase this selection pressure during the incubation. After each SELEX iteration nonbinding candidates are washed out and the bound aptamers are prepared for the next round. This includes the elution of aptamer candidates from target molecules and a following amplification to obtain a library sufficient in size

for the next round. Only oligonucleotides capable of binding to the target or background materials necessary for carrying out the experiment are enriched during that process [14]. This leads to the enrichment of specific and affine aptamers and thus a decrease of diversity in the resulting library can be observed.

NGS techniques now provide the possibility to better analyze such SELEX experiments. Benefits are provided by the magnitudes of higher sequencing coverage of the real library sequence diversity compared to classic sequencing technologies, such as Sanger sequencing [15], and the possibility to gain information from all SELEX rounds with reasonable effort. Hence, it is no longer only the final round that can be analyzed, but rather the development of the library during the whole experiment, which provides new chances in bioinformatics analysis. Nevertheless, the next generation sequencing technology is despite its advantages accompanied by some major drawbacks. NGS is a high throughput sequencing technique, which means that one has to consider sequencing errors. Although the probability of each single base being sequenced incorrectly is quite low, denoted by Phred values up to 41, the large number of single base reads within each data set will induce many sequencing errors [16]. Another problem is that the limits of conventional algorithms and their implementations can easily be reached when processing large NGS data sets.

If one is able to handle these difficulties, the additional information source provided by the NGS technology when performing SELEX experiments allows a deeper analysis and understanding of the SELEX process. So the analysis of only the first rounds of a SELEX experiment may show specific enrichment of the library and thus draws a deduction towards the enrichment of the final round. This could be a first hint for sequence characteristics that yield target-specific binding affinity. Those observations would allow interrupting a running SELEX experiment, skipping some intermediate selection rounds, and instead continuing with a computationally enriched pool at later position, saving time and material expenses. The enrichment of the aptamer library during the SELEX process can be observed when analyzing the sequence data gained from the different rounds. Using the NGS data, a diversity indicator can be calculated and compared, showing that the number of different sequences effectively decreases. It is very important to find a proper description for the observed aptamer enrichment in the later SELEX rounds. Though the simple description of the enrichment as a list of most frequently observed aptamers in the data set is sufficient for conventional validation of the experiments success through concrete binding experiments, a better way of description has to be found when aiming at the improvement of prospective SELEX runs.

The enrichment has to be characterized and more detailed, because occurring commonalities between the different found aptamer sequences indicate characteristics of the aptamers at different physical positions, which are relevant for binding to the target. Sequence motifs are one opportunity to describe those shared features on sequence level. These motifs are in turn corresponding to substructures

within the aptamers, which are characteristically fitted to specific binding sites located on the target molecules surface and therefore are present in all binding aptamers. Using a position specific scoring matrix as motif representation allows the definition of variable regions, which better reflects the natural divergence and thus preserves the informational content gained from the NGS sequence data. Once found, the sequence motifs can be utilized to generate an enriched and thus improved and target-specific starting library for SELEX experiments, which will positively affect the progress of future SELEX runs on the same target molecule. This would imply that for each improved SELEX run another experiment has to be performed to gain the information needed for generating the target-specific starting library for the main experiment. The real practical benefit of the motif description of the sequence libraries enrichment during the SELEX experiments becomes apparent, when later using the motifs as descriptors for the target molecule. The effect can be extended by using multiple bioinformatics technologies, ranging from sequence analysis by employing sequence alignment strategies and clustering techniques to secondary and tertiary structure prediction as well as the aforementioned motif search. Other technologies like electrostatic calculation and docking simulation are utilizing concrete three-dimensional structure information, which can be acquired from databases, through own structural clarification or structure prediction. Combining all these techniques it will be possible to extract a set of descriptors for both, target molecule and found aptamers, which characterize the aptamer-target-binding. These descriptors now need to be correlated appropriately to build an abstract model describing the aptamer-target-binding relation. The model can then be applied to an unknown target molecule in an effort to obtain information on the composition and architecture of binding aptamers only based on information about the desired target. The generation of target-specific SELEX starting libraries without the need of concrete performed previous experiments with the desired target would greatly improve the aptamer finding process.

This paper will present a search technique using suffix trees to find recurring motifs in large NGS nucleotide sequence data sets as one methodology besides the other mentioned techniques allowing deriving target-related descriptors for the later generation of target-specific SELEX starting libraries. This method is exemplarily attempted on an NGS data set supplied from a SELEX experiment targeting a *Norovirus* capsid protein.

## 2. Data Set and Investigated Target

In the past a SELEX experiment was performed to find a DNA aptamer capable of binding to the *Norovirus* genotype II.4 capsid protein VP1 as its target [17]. This aptamer may be used for efficient *Norovirus* detection or infection control. For validation of the successful enrichment of sequences during the experiment and further analysis profiting from the much higher coverage, next generation sequencing was performed to gather sequence data for all screening rounds.

**2.1. Target.** The *Norovirus* has been detected in 1972 in Norwalk, USA, for the first time. Since then this virus could be found in a variety of different genotypes spread all over the world. The *Norovirus* belongs to the family Caliciviridae and is genetically diverse. *Noroviruses* are the major cause of viral epidemic gastroenteritis worldwide, often resulting in large and persisting outbreaks. Two of the five major genogroups, GI and GII, especially the genotype GII.4, are responsible for the majority of human infections. Since only few viruses are already able to cause an infection, they are highly contagious. To the present there is no vaccine available, which could prevent a *Norovirus* disease outbreak [18]. The *Norovirus* contains a single-stranded, positive-sensed RNA genome with an approximate size of 7.7 kb, which is enclosed in a nonenveloped protein coat. This coat exhibits distinct cup-shaped depressions. Its icosahedral capsid structure is formed by 90 dimers of the capsid viral protein 1 (VP1), which is assembled of two domains. The inner S domains form a shell around the RNA, whereas the P domains are protruding on top of the shell [19]. Another minor capsid protein (VP2) is only present in a few copies. The overall construct leads to thermal stability of the virus, allowing it to survive temperatures up to 55°C and a pH in the range of 3–7 [20].

At present, a *Norovirus* infection is usually diagnosed by reverse transcription PCR (RT-PCR) or enzyme-linked immunosorbent assay (ELISA) using anti-*Norovirus* antibodies. Although the cost-intensive RT-PCR is the most sensitive method known so far, the genetic diversity of *Noroviruses* does not allow testing for all genotypes in one assay. Attributable to their low sensitivity ELISA assays can only be used for screening, where the results are confirmed by a following RT-PCR [21]. In a recent development an immunochromatographic detection assay based on antibodies was rated to have a high sensitivity and specificity [22]. As there is still a strong need for point-of-care methods for *Norovirus* detection, a solution using aptamers as receptor units may be another chance to develop real-time, label-free, and possibly low-cost biosensor systems for *Norovirus* detection. Targeting the attachment and internalization of the virus, one interesting approach would be to inhibit the binding of the P2 subdomain to its receptor molecules by competitive interacting molecules. Hence, *Norovirus* binding aptamers might also be used in vivo to control *Norovirus* infection.

**2.2. Origin of Sequence Data.** The target capsid protein VP1 of *Norovirus* genotype II.4 was expressed as a recombinant with polyhistidine-tag appended for later immobilization. The sequences of the initial library contained a 49 nt long random section enclosed by the necessary primers. So the initial library is described by the following template sequence: 5' GCC TCT TGT GAG CCT CCT AAC -N<sub>49</sub>- CAT GCT TAT TCT TGT CTC CC 3'. The SELEX experiment was performed in twelve rounds. After every third selection round an additional negative selection was performed to remove aptamer candidates binding to background materials

of the experiment or to fecal specimen, the later sample matrix.

For each round of the SELEX experiment the next generation sequencing supplied a sequencing file in FASTQ format containing the aptamer sequences remaining after this round. For each sequenced base the file further contains an additional coded quality value which approximates the error probability at this position. The sequences are flanked by parts of the Illumina primer sequences.

*2.3. Preparation of Sequence Data.* Prior to any concrete sequence analysis a preprocessing step of the raw data produced by the sequencer needs to be done. The aptamer sequences are flanked by primer sequences. At first these primer sequences, either fully preserved or just fragments, have to be recognized and removed. Raw sequences that did not contain the given primer sequences have been rejected. The remaining inserts are the object of the intended motif search.

Each sequenced base is annotated with a coded quality value which approximates the error probability at this position. Although these quality values are not regarded as an absolute quality indicator, conspicuously low values or continuous sections exhibiting low values may indicate sequencing errors. As the main goal of a SELEX experiment is the enrichment of the sequence pool with binding aptamers, a sequence occurring only with very small quantity can also be considered as deficient. Based on this information a filter can be applied, which discards sequences of possibly low quality. After preparation the data set contained approximately 233000 sequences, from which 5500 sequences were distinct.

### 3. Motif Search

As intended, this study is aimed at developing a search technique using suffix trees to find recurring sequence motifs, which are corresponding to concrete binding areas of the aptamers. The prepared sequence data of the SELEX experiment described above is the basis for the following search strategy, which will be presented in three main steps. After a short overview of different approaches of motif search utilizing suffix trees, the generation of a generalized suffix tree, which is used by a later exhaustive search, is described. Here, also the possibility of using only subsequences located on loop regions of the predicted structures is mentioned. Thereafter the benefit of the tree structure in doing a full search is outlined. The last part explains a couple of termination criteria for the search. Afterwards a possible way to handle the results of a motif search easier is specified.

*3.1. Suffix Tree Based Motif Search.* Over the last three decades suffix trees have been repeatedly utilized for sequence matching as they are known to provide very fast string operations [23]. The most simplistic problem is to find the exact motifs occurring in a subset of the given sequences. In particular, this can be done by traversing the tree to find nodes visited by the denoted minimum number of sequences. This basic problem increases in complexity when more

meaningful biological demands are considered. This includes the incorporation of character mismatches and sequence gaps during computation. With respect to DNA sequences and their corresponding structures, single motif elements can interact spatially and may be important for structure stabilization or even for defining the three-dimensional fold. However, such motif elements are not necessarily located in direct sequence neighborhood, which requires considering long gaps between elements. A number of approaches target the finding of such gap-containing motifs. Early algorithms permitted only fixed gap lengths—a restriction, which limits the number of possible motif arrangements. More sophisticated algorithms are also able to handle motifs interrupted by gaps of variable lengths [24–26]. In addition, integrating sequence-specific biological relevance to the problem of pattern and motif identification requires an appropriate ranking and processing scheme [27]. The other aspect of this problem lies in defining mismatch acceptance within motif hits, which would allow regarding mutations occurring in evolutionary processes, such as SELEX. These algorithms are usually intended to find motifs containing up to a fixed number of mismatches within each occurrence. In most cases, the mismatches are not restricted by special rules [28, 29]. The aforementioned algorithms directly aim at finding motifs within the suffix tree. An alternate approach affords ranking the search space to find a subspace (subtree) containing appropriate motif hits [30].

Brazma et al. introduced the Pattern Discovery Algorithm, which realizes an exhaustive search for three different classes of motifs. One of these classes called “patterns with character groups” describes mismatches by means of a well-defined regular expression syntax, which helps to specify the motif variability more precisely. The algorithm uses a suffix tree where nodes are annotated with symbols of the employed regular expression syntax, which means the character groups. This massively increases the tree size and thereby limits its practice [31]. Following the example of an exhaustive search over all possible patterns including variable regions within a huge number of nucleotide sequences, the single string search for one pattern in a single sequence needs to be optimized in order to minimize computational costs. In contrast to the Pattern Discovery Algorithm, our approach uses the generalized suffix tree annotated with the letters of the sequence alphabet. The consideration of variability is realized by merging nodes during the later search phase, which reduces memory usage and avoids the creation of unnecessary subtrees that would be created in the character group based tree. In this study, biological relevance is derived from predicted secondary structure information. In particular, free energy estimations of predicted structures are employed to ranking corresponding sequences prior to motif search, which, to our knowledge, poses a novelty in this field.

*3.2. Tree Construction.* To project sequences onto this tree structure, each edge of the generalized suffix tree is annotated with one of the possible characters of the underlying alphabet. Internally each character is mapped onto a number to allow

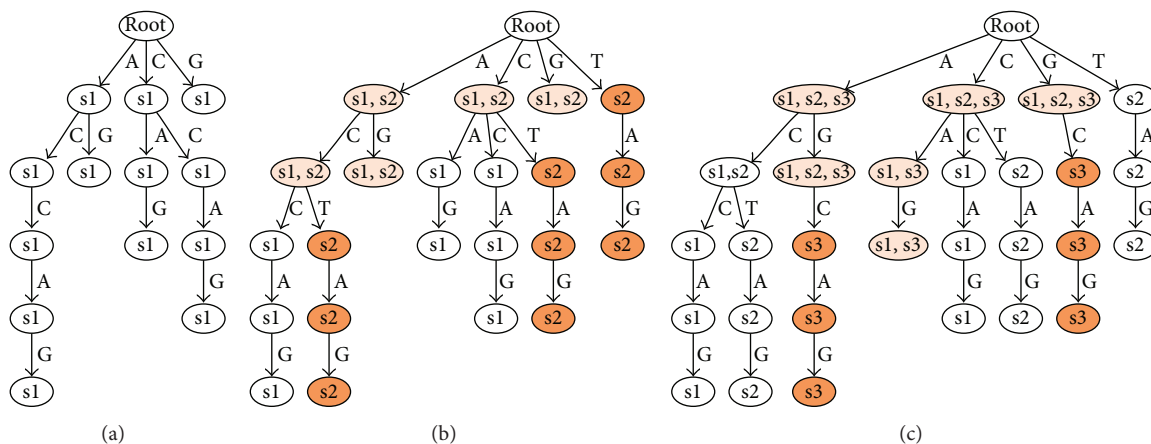


FIGURE 1: One example of the stepwise construction of the generalized suffix tree is shown, which later will be used within the search process. Parts (a), (b), and (c) of the graphic show the state of the tree after inserting the sequences ACCAG as s1, ACTAG as s2, and AGCAG as s3. Each edge is annotated by the corresponding letter of the underlying alphabet. The nodes themselves contain the list of sequence identifiers for all sequences containing the subsequence denoted by the path leading to the particular node. Starting from part b the coloring of the nodes indicates their status of update. Red colored nodes have been added in the latest construction step; orange nodes have been modified.

fast and direct access to the edges via arrays. This means that each path connecting a node with the tree root describes a designated subsequence, which is simply the concatenation of all annotated characters of the edges. This subsequence is implicitly assigned to the node, which itself comprises a list of all sequences containing its assigned subsequence. To find all relevant sequences containing a particular subsequence, it suffices to walk along the tree choosing the edges according to the successive characters of the searched subsequence. The last node now contains the list of all relevant sequences.

The tree is constructed by the repeated insertion of all sequences of the data set. As its model is not intended to map variable positions, all sequences containing variable characters are discarded as a first filtering step. The quality of today’s sequencing technologies and appropriate preprocessing keeps the impact of the filtering insignificant. A single sequence is inserted into the tree by traversing the tree, beginning from the root node. The depth of this insertion traversal is limited by the maximum allowed motif length. If the next edge and connected node, which are chosen by the next character in the inserted sequence, do not exist during traversal, they are created and the procedure is continued. Each node that is traversed during the insertion process will have placed the sequence ID of the inserted sequence into its internal list. Duplicate entries in the nodes internal lists are avoided. As we are creating a suffix tree, not only the sequence itself but also all possible suffixes of the inserted sequence need to be processed in the same manner to complete the insertion of a single sequence. According to this principle all sequences of the data set are inserted consecutively as shown in the example of tree creation in three steps in Figure 1. The time and space complexities of tree creation are within  $O(n \cdot l \cdot r)$ , where  $n$  is the number of sequences,  $l$  the sequence length, and  $r$  the maximal allowed motif length.

In particular loop regions of nucleotide aptamers are likely to interact with target molecules [11]. As the unpaired

nucleotides in loop regions do not take part in Watson-Crick or other kinds of nucleotide pairs, the related binding sites remain available for intermolecular chemical bonding. Loop regions should therefore be preferred when searching for common binding motifs. To adapt the presented strategy towards possible loop regions and potential binding motifs, the construction process of the tree was modified. To determine which parts of the sequences are placed on unpaired regions, the corresponding secondary structures need to be predicted. However, taking only into account the best predicted structure may lead to unintended findings, because predictions can only be trusted to the extension of their predictive performance. In the concrete binding situation many external impacts will influence the folding of the aptamer, so that the structure of the highest binding affinity does not necessarily correspond to the structure yielding minimal free energy. However, the latter is the objective in structure prediction algorithms. Thus, in the context of developing aptamer-target-binding models, RNA structure predictions have to be regarded with care and caution. Therefore a set of suboptimal structures is used as basis, which is predicted with the tool RNAsubopt of the Vienna RNA toolbox [32]. Hence the RNAsubopt application is primarily designed to be applied on RNA sequences; the prediction of DNA secondary structures requires a different energy parameterization [33, 34]. As the primer sequences are attached to the main aptamer sequence during the incubation phase, they are influencing its structural fold. Due to that the primer sequences need to be attached prior to predicting the aptamer secondary structures and neglected after prediction. For each of the predicted structures of each sequence, all loop subsequences are extracted and separately inserted into the tree. Loop regions that are contained in more than one suboptimal structure are now inserted multiple times. For a correct interpretation in the later pattern search, the inserted loop regions have to be weighted. The selection

GCTAGCTAGCTAGCTAGCTGACTGATCTCTTCATGATCGACTGATC	$E(x)$	$P(x)$
	-6.9	3.6%
	-6.9	3.6%
	-7.1	4.9%
	-7.1	4.9%
	-8.6	56.0%
	-6.6	2.2%
	-8.1	24.9%

FIGURE 2: The selection and weighting of secondary structure information prior to tree creation is illustrated. The top row shows the sequence of interest. In the following rows, for each secondary structure predicted by the Vienna RNA tool RNAsubopt [23], a representation in dot-bracket notation is placed. Mutual matching pairs of brackets denote base pairs, whereas dots are standing for unpaired bases. Each structure  $x$  is annotated with an energy value  $E(x)$  and the calculated probability  $P(x)$  used for weighting. The nonpaired subsequences, which will be inserted into the tree, are highlighted yellow. It is obvious that some subsequences are inserted multiple times and others are overlapping, which necessitates the mentioned weighting.

and weighting of unpaired structure elements is depicted in Figure 2. Besides a standard equal weighting of all structures, a special weighting according to the secondary structures annotated free energy values can be realized. Therefore a kind of probability for each sequence to be found in a natural mixture is calculated by the following formula using Boltzmann factors based on these energy values [35]:

$$P(x \in X) = \frac{1}{Z} \cdot e^{-1/(k_B T) \cdot \beta \cdot E(x)} \quad (1)$$

$$Z = \sum_{x \in X} e^{-1/(k_B T) \cdot \beta \cdot E(x)}$$

For each of the structures  $x$  of the ensemble  $X$ , an energy value  $E(x)$  needs to be available. The Boltzmann constant  $k_B$  and the absolute temperature  $T$  in kelvin are also required for calculation. The temperature value should be consistent with the settings used in secondary structure prediction. An additional parameter  $\beta$  allows customizing the characteristic of the weighting function. Larger values of  $\beta$  increase the up-weighting of better energy values; smaller values weaken the influence of the predicted free energy. A value of 0 for  $\beta$  comes to an equal weighting of all structures. The partition function  $Z$  can be seen as a normalization factor, so that the sum of all calculated probabilities will not exceed the limit 1.

Now the tree can be constructed either with or without using the information provided by the predicted secondary structures.

**3.3. Motif Search Using Node Merging.** The motif search algorithm shall be able to find motifs containing variable regions. As the underlying tree structure only models non-variable strings, the variability needs to be realized within the search process. Therefore a new composite alphabet is created and used as basis for the following search. This composite alphabet contains all standard characters taken from the normal sequence alphabet and all possible combinations as

special, variable characters. The composite alphabet can easily be restricted to only 2-letter or 3-letter combinations. With the help of this composite alphabet, the search now is able to cover variable motifs.

Performing the full search with the help of a suffix tree allows a very fast substring search strategy for both, normal and variable substrings. This strategy uses the principle of progressive node merging, which is a depth-first search. Each search process starts with the empty string, which is represented by the root node of the tree. As we are using node sets, the starting node set only consists of the trees root node. For each possible following character the search is continued. The following character can also be a combination of more than one character of the original sequence alphabet, because a composite alphabet is used. To continue the search, a new set of nodes covering the next searched substring with all variabilities needs to be found. This is done by aggregating all subnodes of nodes contained in the current node set, whose edges correspond to the composite character currently processed. The obtained node set is now merged to retrieve a single list of sequences containing the pattern represented by the nodes. This principle is demonstrated on two simple examples in Figure 3. Since a single sequence may occur in multiple nodes, the merged list of sequences has to be cleaned from redundant entries. Instead of holding a sorted list or linearly searching for each sequence prior to inserting it, an index list helps to ignore doublets with minimal time overhead, only requiring the sequences to carry a serial number. Now the merged node holds all sequences containing the searched pattern. The search effort for a pattern with one additional character is therefore minimal, because neither an actual string search nor a full tree traversal needs to be accomplished for each step of the motif search. The space complexity is within  $O(n)$ , where  $n$  is the number of sequences. However, due to the exhaustive search the maximum time complexity of the search is within  $O(|\Sigma^*|^r)$ , where  $\Sigma^*$  is the compound alphabet and  $r$  the maximal allowed motif length. The following termination criteria will reduce the required computational effort.

**3.4. Termination Criteria.** Two straightforward termination criteria are defined when starting the search procedure. The first criterion is the maximal motif length, which bounds the depth-first search at a specific depth. The second is the quantity of sequences containing the current motif. As a motif extended by one additional character must be equal or less frequent than the original, the search branch can be cut when the limit of quantity is reached.

For all further criteria, the motif actually contained in the found sequences needs to be constructed. Therefore all actual occurrences of the motif, which are located in the merged nodes internal list, are analyzed position by position. A result of this analysis is a position specific scoring matrix (PSSM), which now can be used for comparison and further calculation.

Another termination criterion is the formal integrity of a discovered motif. That means the exact match of the found motif with the motif actually contained in the sequences

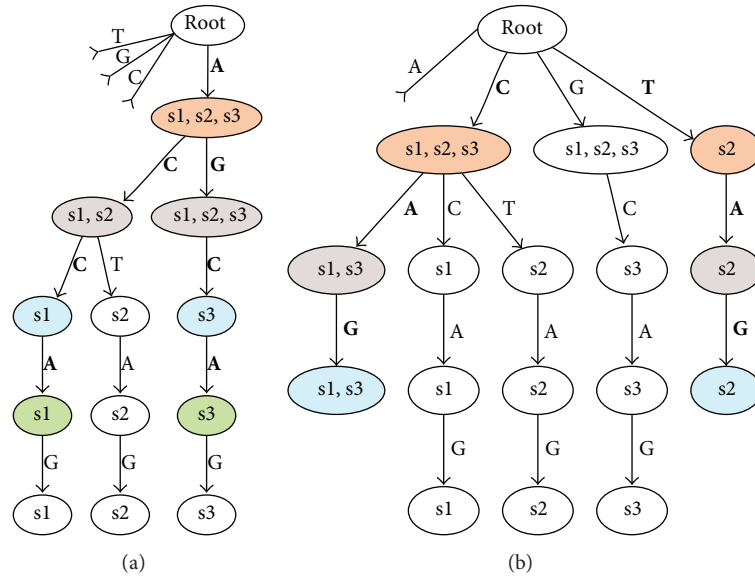


FIGURE 3: The search process within the suffix tree in two examples using variable motif positions is illustrated. In both examples parts of the tree have been omitted for visual perspicuity purposes. The basis is the tree, which was constructed in Figure 1. The changing colors from one row to another are for visual distinction of the consecutive node sets during search. Letters corresponding to the chosen edges are printed in bold font weight. In (a) the motif A[CG]CA is searched, which leads to a fork in the tree at step two. It is not necessary to traverse the tree down to the leaves. Merging the green nodes on the last marked row offers the list of search results (s1, s3). (b) searches for the motif [CT]A[GT]. In the last step there is no suitable edge found for the second allowed character T. Merging the cyan nodes on the last marked row offers the list of search results (s1, s2, s3).

which is denoted by the PSSM. If these motifs do not match, because at some position of the actual motif an original character is missing, the branch can also be rejected, because the presence of another (namely, the actual) motif covering that branch is mandatory.

Some other restrictions are only applicable for motif filtering, but not for termination of the search branches. Besides the minimal motif length, the entropy based total information of single positions of a motif and the average total information of all positions of the motif can be mentioned here. As the entropy  $H$  of an event, in this case of the event described by the probability distribution of one position in the PSSM, is a measure of the uncertainty, its complement can be used as a measure of expressiveness. We have chosen the Shannon entropy  $H = -\sum_{i=0}^N p_i \cdot \log_2 p_i$  which uses  $p_i$  as the values for probability or relative frequency of the characters in one column of the PSSM and  $N$  as the original alphabets length. It has a maximum value of  $H_{max} = \log_2 N$ . The total information  $E$  is then simply the difference  $E = H_{max} - H$ , which leads to values from 0 at uniform distribution to 2 for a nonvariable position [36].

However, a limitation of the total information values as described would result in the avoidance of possible gaps, which means positions of low total information. If they are desired to be found, defining another upper limit of total information to identify gaps, which are not validated by the standard total information criterion, will help. Motifs starting or ending with such a gap can be discarded without any consequence.

3.5. *Aggregation of Motif Results.* In consequence of the allowed variability and the used naive search strategy, a very large number of motifs will be eventually found, and thus the result of the algorithm will be difficult to manage. However, the resulting motif hits will naturally form a number of motif groups offering high mutual similarity, because the variability at each position leads to some kind of vacillation around a main motif. One possible solution to relieve the manageability is to group found patterns together by using an easy derivable consensus sequence of each pattern. A directed graph connecting a motif to other motifs, which are substrings of itself, is the preferred visual representation as seen in Figures 4 and 5.

## 4. Results

4.1. *Normal Motif Search.* For a motif search, the most frequent 1000 distinct sequences of the last round of the SELEX run have been chosen. The search was limited to motifs of length 7 to 11 and shall only show results with minimal total information of 1.8 bits, which occur in at least 95% of the approximately 233,000 concerned sequences. The variability was constrained by allowing only one or two original characters in each character of the composite alphabet.

The motif search results in approximately 150,000 motifs, which can be separated into 18 groups. The 18 groups are shown in Figure 4. The two longest consensus sequences of the groups are overlapping and thereby forming the motif



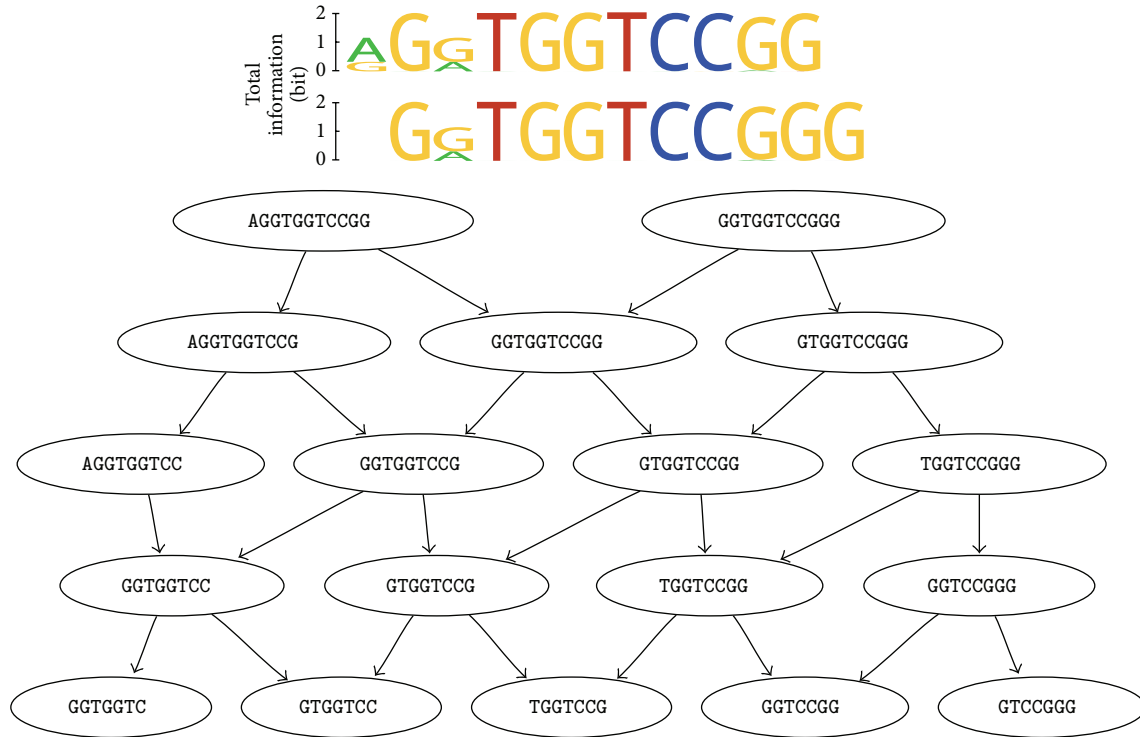


FIGURE 4: The result of the performed motif search is shown. In the lower part of the figure, the consensus sequences of the 18 discovered groups, which contain the actual motifs, are depicted. This is done in the form of a directed graph showing a substring relation. That means that a consensus sequence is connected to all other sequences, which are substrings of itself. The emerging hierarchy facilitates the understanding and selection of relevant finds. The upper part shows two concrete motifs in the form of weblogs, which have been picked one from each of the top consensus groups and then have been aligned to each other. The height of each column of the two motif weblog representations corresponds to the motif positions total information according to the scale on the left side. The letters are then sized by their relative frequency within that motif position.

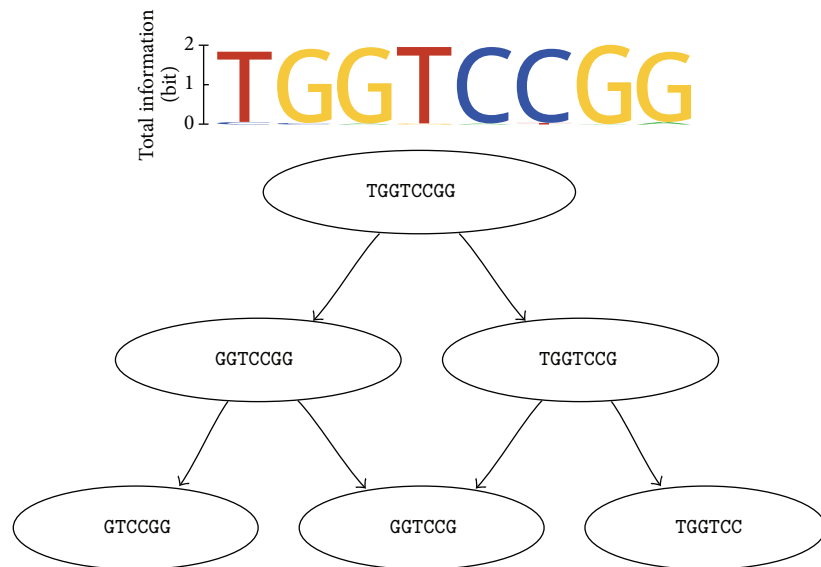


FIGURE 5: The result of the performed motif search using the secondary structure restrictions is shown. In the lower part of the figure, the consensus sequences of the 6 discovered groups, which contain the actual motifs, are depicted. This is done in the form of a directed graph as in Figure 4. The upper part shows one concrete motif in the form of a weblog, which has been picked from the top consensus group. See Figure 4 for further explanation of the weblogs representation.

(A)GGTGGTCCGG(G). The other 16 found groups show consensus sequences, which are subsequences of the two largest finds. The main focus shall therefore be laid on the two longest finds. Looking at the concrete formation of the motifs contained in these two groups shows that the only noticeable variability lies in positions 1 and 3 of the motifs. This yields the overall motif description of [AG]G[AG]TGGTCCGGG.

The sequence data set was also submitted to different motif search webservices. Only two of the tested services were able to handle the large data set. DREME returned 50 motifs offering 4600 to 40 matches within the given 5500 distinct input sequences [37]. The DRIMust online service resulted in a list of overrepresented k-mers and one motif hit [38]. The first motif hit reported by DREME as well as the top elements of the overrepresented k-mers provided by DRIMust corresponds to the motif found by this approach, whereas the DRIMust motif and later motif hits reported by DREME do not match to our result. The extended use of variability in combination with the exhaustive search strategy facilitates the finding of motifs that fit the natural variation more precisely. Due to this a very strict threshold could be applied to sequence coverage (95%) during the motif search.

**4.2. Using Secondary Structure Information.** In a second run, the secondary structure information was used to select only subsequences for motif search, which are likely to be located on loop regions of the structure. For that reason a suboptimal secondary structure prediction with an allowed energy delta of 1 kcal/mol was chosen. The absolute temperature  $T$  was set to 310 K and parameter  $\beta$  was set to 1. As this selection restricts the number and length of subsequences, which provide the basis for the motif search, using the same severe parameters as above will cause the search to reveal a reduced result focused on the loop regions.

With the altered base set the algorithm discovers approximately 125 motifs, which are aggregated into 6 consensus groups shown in Figure 5. The group with the longest consensus sequence is TGGTCCGG, which is a subsequence of the motif discovered without using secondary structure information. The other finds are subsequences of this motif. The circumstance that the motif discovered with structural restrictions is a subsequence of the one found without such restraints supposes that the found motif is relevant for binding to the target.

As we initially introduced a weighting based on the predicted free energy of the secondary structures, each found motif now contains a value describing a kind of propensity or probability for this motif to be found on loop regions of the aptamers structure. The longer the desired motif, the lower the expected propensity. So the longest motif TGGTCCGG is accompanied by a value of around 65%. The most common group TGGTCC in contrast ranges from values of 71% to 80% and is therefore probably assembled of unpaired nucleotides.

**4.3. Validation.** As a manual validation the 25 most frequently occurring sequences of the data set have been checked. After the aggregation of the sequences into six groups of mutual global similarity, the consensus sequences

of these groups were inspected. All except one sequence did contain the motif [AG]G[AG]TGGTCC[GA]GG, where only a small percentage is responsible for the last variable position. The one remaining sequence does only contain the motif TGGTC[]GGG with one missing C in the middle of the motif. One aptamer containing the found motif has also been experimentally confirmed to bind to the target.

For the top sequences of the groups determined above, secondary structures have been predicted separately to map the found motif onto the possible aptamer structures. The visualization of the structures was done with the online tool VARNA [39] and is presented in Figure 6. In some cases the optimal predicted structure contained the motif positioned on an unpaired region. Although the motif was positioned partially or even fully on paired regions in the other considered cases, a suboptimal structure with small energy difference existed, on which the motif was found nearly or fully on a structure element consisting of unpaired nucleotides. This finding can be attributed to the new methodology incorporating the predicted secondary structure information into the motif search process.

**4.4. Library Generation.** The SELEX experiment resulted in a final library with decreased diversity. Using the NGS data this decrease has been validated by calculating the diversity measures Simpson index and Shannon-Weaver index [40]. Corresponding to that diversity an enrichment of a number of aptamer sequences within the library can be observed. Besides a simple grouping of the sequences by global similarity, another approach, the motif search, was pursued. As a result of this performed motif search a short motif was revealed, which could be found in more than 95% of the investigated sequences. This motif is furthermore positioned on a loop region of suboptimally predicted secondary structures in the majority of the cases. This leads to the assumption that the motif TGGTCCGG is especially relevant for target binding, because loop regions offer unpaired nucleotides whose binding sites remain available for intermolecular chemical bonding.

As shown above, the motif corresponds to similar substructures within the different enriched aptamers, which may fit characteristically onto a specific binding site located on the target protein. This circumstance can be used to generate an enhanced SELEX starting library, which in turn will positively affect the progress of future SELEX runs on the same target molecule. As the discovered motif is described by a position specific scoring matrix, the natural divergence is captured and can be used when creating the new library. The motif itself represents a kind of indication for a preferred aptamer binding site; it is not a fully qualified predefinition of the optimal and exact binding aptamer. A SELEX library should therefore be enriched by the motif. One possibility is to create a small preliminary library highly enriched with that motif, which is modified and thereby inflated in the process of postrandomization. Another way would be a randomized sequence generation with the restriction, so that the resulting sequences have to contain a small number of possible variant instances of the desired motif. By this means,

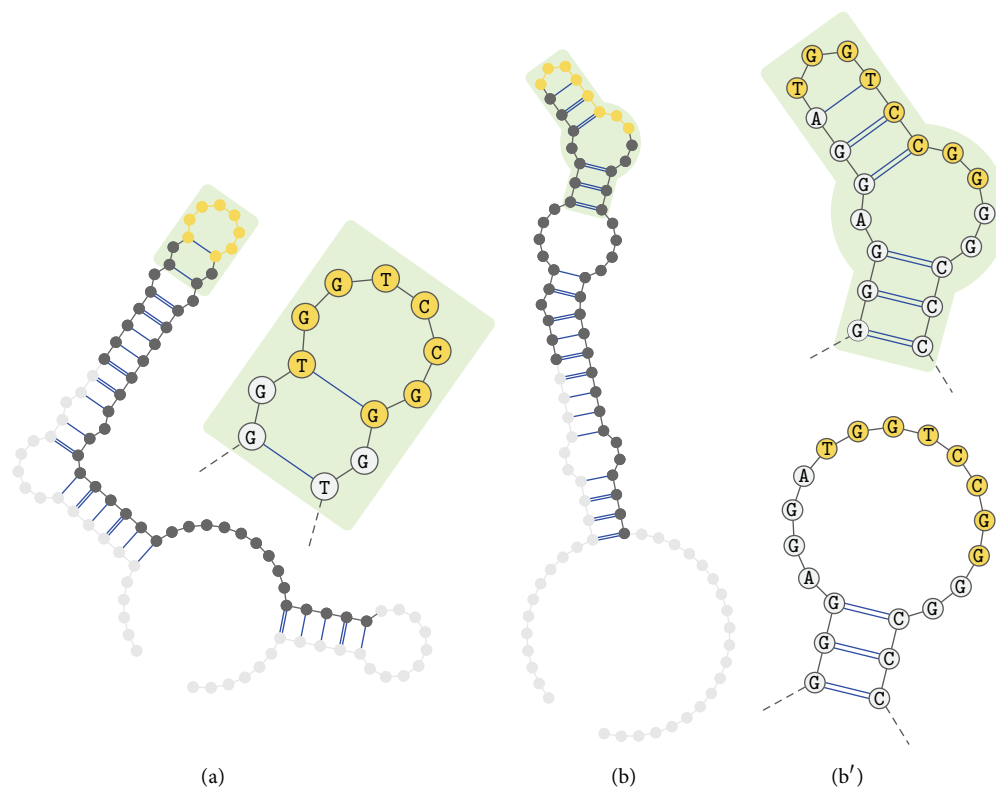


FIGURE 6: The result of mapping the found motif TGGTCCGG onto different predicted secondary structures of aptamers frequently occurring within the final SELEX round is shown. This is done using two examples. In both cases, based on the output of the VARNA [27] online tool, the optimally predicted secondary structure is schematically drawn with the following coloring. Light gray circles are nucleotides of the primer sequences, whereas dark gray and yellow circles are nucleotides of the actual aptamer. The latter are containing the searched motif. The area containing the motif is shaded in a light green tone and additionally presented in a separated detail view besides. In (a), the motif is exactly matching a hairpin loop. In (b) the motif is distributed over paired and unpaired nucleotides. A second detailed view (b') shows the same part based on a suboptimal structure instead providing a larger loop as an only difference, which holds the motif.

the highly complex conformation space of the aptamers is filled diversely with structures containing different configurations of the potential binding motif. This ensures that also conformational changes of the aptamers induced by the influence of the target molecule and other environmental impacts while binding are abstractly regarded in the libraries creation process. Following SELEX runs can eventually profit from the target-specific enhanced starting library, which was designed by using the additionally gathered NGS sequence data.

## 5. Discussion

In a narrow sense, the correct application of the described method would imply that for each SELEX run, which shall profit from the target specifically generated new libraries, another SELEX experiment has to be performed to gather the sequence data required for finding the relevant motifs. In the direct manner this can be used after a performed SELEX experiment offering only aptamers of relatively low affinity. If motifs can be determined, a following SELEX experiment with optimized library could be used to find

aptamers with higher affinity in fewer rounds. Another application is the optimization of the SELEX procedure. In normal cases the diversity decreases slowly in the later rounds of the experiment. The strategy discussed in this paper could reduce the number of necessary SELEX runs by introducing a sequence analysis step. After the analysis the experiment will be continued with a motif-based enriched library to have better chances to capture higher affinity aptamers.

The found motifs can be seen as one descriptor for the target, because aptamers containing that motif are likely to bind to that intended target molecule. This can be a consequence of physiochemical preferences of the amino acids and nucleotides as well as concrete structural preferences of the motif. The shown method can be extended and thereby practically enhanced by making use of other available, mostly complex descriptors for the target and also for the aptamers. This starts with descriptors based on the pure sequence, for example, sequence alignments, consensus sequences, clusterings, and base or amino acid distributions, but is not limited to these. It is also possible to use available secondary or tertiary structures of the binding partners or to predict these structures, which then can be analyzed in terms of physical surface formation, electrostatics, buriedness, and availability

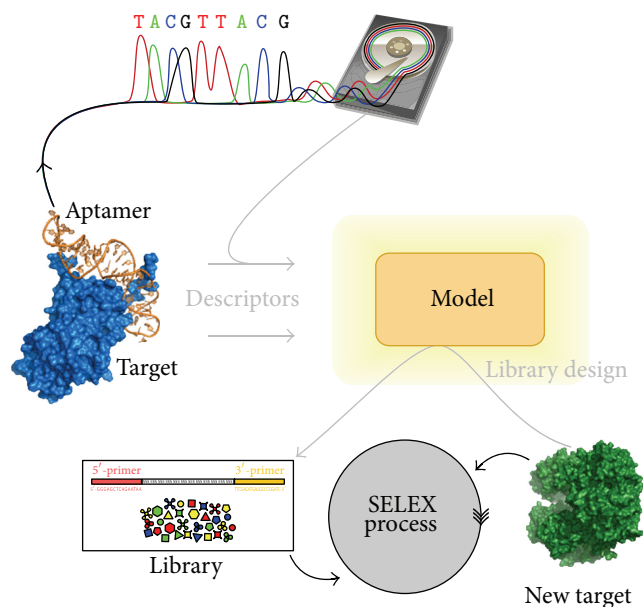


FIGURE 7: A schematic depiction of the longer term goal is shown. The left upper area illustrates the creation of an abstract model based on aptamer-target-binding information gained in the form of a multitude of descriptors for both, target and aptamer. The lower section illustrates the usage of the abstract model to generate a target-specific SELEX starting library only based on information about the desired new target molecule.

of the different amino acids and nucleotides. It is also surmisable to use a docking simulation to validate or even identify potential binding sites, which then can be described in more detail. After describing both, target and aptamer, in an appropriate model by quantifiable descriptors, these values can be correlated in a new model abstractly describing the aptamer-target-binding relationship. Now the real practical benefit of the basic strategy becomes obvious. At this point, the model can significantly contribute to dry and wet lab investigations, since it is applicable to other, even structurally unknown target proteins, and can aid in gaining knowledge on the composition and architecture of binding aptamers only based on information about the desired target. The generation of target-specific SELEX starting libraries without the need of concrete performed previous experiments with the desired target as illustrated in Figure 7 would greatly improve the aptamer finding process in fields of biosensor development and medical treatment.

## 6. Conclusion

Performing NGS on SELEX experiments can yield benefits. Although this sequencing is not part of the standard SELEX procedure, the technique and following sequence analysis can help to find a better description of the developed enrichment within the library. In this paper the enrichment of a specific sequence motif has been shown by performing a motif search on the sequenced last round of a SELEX experiment. The high enrichment of sequences containing this motif and its

likelihood to be located on unpaired regions of the aptamers indicate the motifs relevance for binding to the target protein. According to that the motif corresponds to a specific characteristic of the target. This kind of target description is only a first step towards an abstract model describing the aptamer-target-binding relationship, which then can be utilized to predict information on composition and architecture of binding aptamers. Based on this information SELEX starting libraries can be generated target-specific, which in turn will save time and financial expenses.

## Conflict of Interests

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

## Acknowledgments

This work has been supported and funded by the Free State of Saxony and the European Social Fund (ESF).

## References

- [1] A. D. Keefe, S. Pai, and A. Ellington, "Aptamers as therapeutics," *Nature Reviews Drug Discovery*, vol. 9, no. 7, pp. 537–550, 2010.
- [2] J. Zhou, M. L. Bobbin, J. C. Burnett, and J. J. Rossi, "Current progress of RNA aptamer-based therapeutics," *Frontiers in Genetics*, vol. 3, article 234, 2012.
- [3] N. B. Leontis and E. Westhof, "Analysis of RNA motifs," *Current Opinion in Structural Biology*, vol. 13, no. 3, pp. 300–308, 2003.
- [4] L. A. Holeman, S. L. Robinson, J. W. Szostak, and C. Wilson, "Isolation and characterization of fluorophore-binding RNA aptamers," *Folding and Design*, vol. 3, no. 6, pp. 423–431, 1998.
- [5] K. Harada and A. D. Frankel, "Identification of two novel arginine binding DNAs," *EMBO Journal*, vol. 14, no. 23, pp. 5798–5811, 1995.
- [6] H. Schürer, K. Stembera, D. Knoll et al., "Aptamers that bind to the antibiotic moenomycin A," *Bioorganic and Medicinal Chemistry*, vol. 9, no. 10, pp. 2557–2563, 2001.
- [7] S. E. Lupold, B. J. Hicke, Y. Lin, and D. S. Coffey, "Identification and characterization of nuclease-stabilized RNA molecules that bind human prostate cancer cells via the prostate-specific membrane antigen," *Cancer Research*, vol. 62, no. 14, pp. 4029–4033, 2002.
- [8] M. S. L. Raddatz, A. Dolf, E. Endl, P. Knolle, M. Famulok, and G. Mayer, "Enrichment of cell-targeting and population-specific aptamers by fluorescence-activated cell sorting," *Angewandte Chemie*, vol. 47, no. 28, pp. 5190–5193, 2008.
- [9] J. H. Lee, M. D. Canny, A. de Erkenez et al., "A therapeutic aptamer inhibits angiogenesis by specifically targeting the heparin binding domain of VEGF165," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 52, pp. 18902–18907, 2005.
- [10] Y. Wu, Z. Zhong, J. Huber et al., "Anti-vascular endothelial growth factor receptor-1 antagonist antibody as a therapeutic agent for cancer," *Clinical Cancer Research*, vol. 12, no. 21, pp. 6573–6584, 2006.
- [11] J. Hoinka, E. Zotenko, A. Friedman, Z. E. Sauna, and T. M. Przytycka, "Identification of sequence—structure RNA binding

- motifs for SELEX-derived aptamers,” *Bioinformatics*, vol. 28, no. 12, pp. i215–i223, 2012.
- [12] P. S. Pendergrast, H. N. Marsh, D. Grate, J. M. Healy, and M. Stanton, “Nucleic acid aptamers for target validation and therapeutic applications,” *Journal of Biomolecular Techniques*, vol. 16, no. 3, pp. 224–234, 2005.
- [13] M. L. Metzker, “Sequencing technologies the next generation,” *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2010.
- [14] R. Stoltenburg, C. Reinemann, and B. Strehlitz, “SELEX—a (r) evolutionary method to generate high-affinity nucleic acid ligands,” *Biomolecular Engineering*, vol. 24, no. 4, pp. 381–403, 2007.
- [15] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 12, pp. 5463–5467, 1977.
- [16] C. Luo, D. Tsementzi, N. Kyrpides, T. Read, and K. T. Konstantinidis, “Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample,” *PLoS ONE*, vol. 7, no. 2, Article ID e30087, 2012.
- [17] R. Beier, C. Pahlke, P. Quenzel et al., “Selection of a DNA aptamer against norovirus capsid protein VP1,” *FEMS Microbiology Letters*, vol. 351, no. 2, pp. 162–169, 2014.
- [18] D. P. Zheng, T. Ando, R. L. Fankhauser, R. S. Beard, R. I. Glass, and S. S. Monroe, “Norovirus classification and proposed strain nomenclature,” *Virology*, vol. 346, no. 2, pp. 312–323, 2006.
- [19] B. V. Prasad, M. E. Hardy, T. Dokland, J. Bella, M. G. Rossmann, and M. K. Estes, “X-ray crystallographic structure of the Norwalk virus capsid,” *Science*, vol. 286, no. 5438, pp. 287–290, 1999.
- [20] S. F. Ausar, T. R. Foubert, M. H. Hudson, T. S. Vedvick, and C. R. Middaugh, “conformational stability and disassembly of norwalk virus-like particles: effect of pH and temperature,” *Journal of Biological Chemistry*, vol. 281, no. 28, pp. 19478–19488, 2006.
- [21] J. J. Gray, E. Kohli, F. M. Ruggeri et al., “European multicenter evaluation of commercial enzyme immunoassays for detecting norovirus antigen in fecal samples,” *Clinical and Vaccine Immunology*, vol. 14, no. 10, pp. 1349–1355, 2007.
- [22] L. D. Bruggink, K. J. Witlox, R. Sameer, M. G. Catton, and J. A. Marshall, “Evaluation of the RIDA QUICK immunochromatographic norovirus detection assay using specimens from Australian gastroenteritis incidents,” *Journal of Virological Methods*, vol. 173, no. 1, pp. 121–126, 2011.
- [23] R. Giegerich and S. Kurtz, “From Ukkonen to McCreight and Weiner: a unifying view of linear-time suffix tree construction,” *Algorithmica*, vol. 19, no. 3, pp. 331–353, 1997.
- [24] C. S. Iliopoulos, J. Mchugh, P. Peterlongo, N. Pisanti, W. Rytter, and M.-F. Sagot, “A first approach to finding common motifs with gaps,” *International Journal of Foundations of Computer Science*, vol. 16, no. 6, pp. 1145–1154, 2005.
- [25] P. Antoniou, M. Crochemore, C. Iliopoulos, and P. Peterlongo, “Application of suffix trees for the acquisition of common motifs with gaps in a set of strings,” in *Proceedings of the International Conference on Language and Automata Theory and Applications*, 2007.
- [26] L. Marsan and M.-F. Sagot, “Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification,” *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 345–362, 2000.
- [27] L. Leibovich and Z. Yakhini, “Efficient motif search in ranked lists and applications to variable gap motifs,” *Nucleic Acids Research*, vol. 40, no. 13, pp. 5832–5847.
- [28] F. Zare-Mirakabada, P. Davoodib, H. Ahrabiana, A. Nowzari-Dalinia, M. Sadeghic, and B. Goliaeia, “Finding motifs based on suffix trie,” *Advanced Modeling and Optimization*, vol. 11, no. 2, 2009.
- [29] M. F. Sagot, “Spelling approximate repeated or common motifs using a suffix tree,” in *LATIN’98: Theoretical Informatics*, pp. 374–390, Springer, Berlin, Germany, 1998.
- [30] A. Mohapatra, P. M. Mishra, and S. Padhy, “Motif search in DNA sequences using generalized suffix tree,” in *Proceedings of the 10th International Conference on Information Technology (ICIT ’07)*, pp. 100–103, Orissa, India, December 2007.
- [31] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen, “Predicting gene regulatory elements in silico on a genomic scale,” *Genome Research*, vol. 8, no. 11, pp. 1202–1215, 1998.
- [32] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdisen et al., “ViennaRNA package 2.0,” *Algorithms for Molecular Biology*, vol. 6, no. 1, article 26, 2011.
- [33] D. H. Turner and D. H. Mathews, “NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure,” *Nucleic Acids Research*, vol. 38, supplement 1, Article ID gkp892, pp. D280–D282, 2009.
- [34] J. SantaLucia Jr. and D. Hicks, “The thermodynamics of DNA structural motifs,” *Annual Review of Biophysics and Biomolecular Structure*, vol. 33, pp. 415–440, 2004.
- [35] I. Miklós, I. M. Meyer, and B. Nagy, “Moments of the Boltzmann distribution for RNA secondary structures,” *Bulletin of Mathematical Biology*, vol. 67, no. 5, pp. 1031–1047, 2005.
- [36] T. D. Schneider and R. M. Stephens, “Sequence logos: a new way to display consensus sequences,” *Nucleic Acids Research*, vol. 18, no. 20, pp. 6097–6100, 1990.
- [37] T. L. Bailey, “DREME: motif discovery in transcription factor ChIP-seq data,” *Bioinformatics*, vol. 27, no. 12, pp. 1653–1659, 2011.
- [38] L. Leibovich, I. Paz, Z. Yakhini, and Y. Mandel-Gutfreund, “DRIMust: a web server for discovering rank imbalanced motifs using suffix trees,” in *Nucleic Acids Research*, vol. 41, pp. W174–W179, 2013.
- [39] K. Darty, A. Denise, and Y. Ponty, “VARNA: interactive drawing and editing of the RNA secondary structure,” *Bioinformatics*, vol. 25, no. 15, pp. 1974–1975, 2009.
- [40] C. J. Keylock, “Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy,” *Oikos*, vol. 109, no. 1, pp. 203–207, 2005.

## Research Article

# A Novel Approach for Discovering Condition-Specific Correlations of Gene Expressions within Biological Pathways by Using Cloud Computing Technology

**Tzu-Hao Chang,<sup>1</sup> Shih-Lin Wu,<sup>2</sup> Wei-Jen Wang,<sup>3</sup>  
Jorng-Tzong Horng,<sup>3,4</sup> and Cheng-Wei Chang<sup>5</sup>**

<sup>1</sup> Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei 110, Taiwan

<sup>2</sup> Department of Computer Science and Information Engineering, College of Engineering, Chang Gung University, Taoyuan 333, Taiwan

<sup>3</sup> Department of Computer Science and Information Engineering, National Central University, Taoyuan 320, Taiwan

<sup>4</sup> Department of Biomedical Informatics, Asia University, Taichung 413, Taiwan

<sup>5</sup> Department of Information Management, Hsing Wu University, New Taipei City 244, Taiwan

Correspondence should be addressed to Jorng-Tzong Horng; horng@db.csie.edu.tw  
and Cheng-Wei Chang; 095010@mail.hwu.edu.tw

Received 25 July 2013; Revised 18 November 2013; Accepted 15 December 2013; Published 22 January 2014

Academic Editor: Che-Lun Hung

Copyright © 2014 Tzu-Hao Chang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Microarrays are widely used to assess gene expressions. Most microarray studies focus primarily on identifying differential gene expressions between conditions (e.g., cancer versus normal cells), for discovering the major factors that cause diseases. Because previous studies have not identified the correlations of differential gene expression between conditions, crucial but abnormal regulations that cause diseases might have been disregarded. This paper proposes an approach for discovering the condition-specific correlations of gene expressions within biological pathways. Because analyzing gene expression correlations is time consuming, an Apache Hadoop cloud computing platform was implemented. Three microarray data sets of breast cancer were collected from the Gene Expression Omnibus, and pathway information from the Kyoto Encyclopedia of Genes and Genomes was applied for discovering meaningful biological correlations. The results showed that adopting the Hadoop platform considerably decreased the computation time. Several correlations of differential gene expressions were discovered between the relapse and nonrelapse breast cancer samples, and most of them were involved in cancer regulation and cancer-related pathways. The results showed that breast cancer recurrence might be highly associated with the abnormal regulations of these gene pairs, rather than with their individual expression levels. The proposed method was computationally efficient and reliable, and stable results were obtained when different data sets were used. The proposed method is effective in identifying meaningful biological regulation patterns between conditions.

## 1. Introduction

Using microarray technology combined with computational analysis is one of the most efficient and cost-effective methods for studying cancer. Using this method has enabled scientists to investigate and understand a vast array of cancer information [1–5], and it is used to analyze the functionality of specific genes during the development of a disease. In a high-throughput and parallel manner, expression profiling is performed by monitoring the expression levels for the thousands of genes that are simultaneously on an array.

One of the most common and extensively used methods for determining the biological significance of genes when comparing cancerous and normal conditions is the identification of differential gene expressions. Multiple technologies, such as cloud computing, parallel systems, and data analysis strategies, have been developed to identify differential gene expressions by using microarray gene expression data [6–10].

In microarray experiments, the gene expression levels of the samples can be detected using the intensity of probes [6]. Heretofore, most studies have focused only on finding the differential gene expressions of various conditions; however,

lack of methods focused on analyzing the correlations of differential gene expressions between conditions. Therefore, some abnormal regulations causing the diseases might have been disregarded. In addition, the major challenge accompanying this broad approach is computational complexity, because too many differential gene expressions (probes) must be calculated.

Cloud computing and parallel processing are considered valuable techniques because using them can greatly reduce the computation time of a program by efficiently combining multiple computers and processors in parallel. Therefore, we implemented cloud computing techniques to complete the program in considerably little time. Apache Hadoop is a distributed parallel data-processing framework that supports MapReduce-type computations, enabling users to perform distributed computations effectively in increasingly brittle environments [11]. We propose an approach for discovering the condition-specific correlations of gene expressions within biological pathways that implement an Apache Hadoop cloud computing platform. Three microarray data sets of breast cancer were collected from the Gene Expression Omnibus (GEO), and pathway information from the Kyoto Encyclopedia of Genes and Genomes (KEGG) was applied for discovering meaningful biological correlations.

## 2. Related Work

Previously, several methods have been developed to identify differential gene expressions in biological pathways. Most of these methods focus on diagramming gene expression levels and calculating the correlation of clustered genes. Kanehisa et al. proposed an approach, Pathway Miner, which extracts gene-association networks from molecular pathways for predicting the biological significance of gene expression microarray data [12]. When pathways are extracted from Pathway Miner, the levels of gene expression can be discerned, but four samples could be shown at a time. ArrayXPath functions by focusing on mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway information and by displaying the results using Scalable Vector Graphics [13]. The genes can be annotated using different colors in a pathway according to condition. In addition, ArrayXPath can be used for conducting clustering analysis of the time series data of gene expression. Another function, PathMesh, facilitates analyzing the association between gene and disease terms by using MeSH. Thus far, however, lack of tool exists for clearly discerning coexpressional changes under different conditions.

The KEGG consists of a suite of databases that record an extensive amount of information on genes, enzymes, and regulation pathways [12], facilitating the retrieval of gene names and information on their interactions within biological pathways. The KEGG application programming interface (API) is essential for accessing the KEGG system and enables searching and computing the biochemical pathways involved in cellular processes or the analysis of all genes from a completely sequenced genome.

Apache Hadoop is an open source implementation of the Google MapReduce technology that simplifies programming tasks by automatically performing duties such as job scheduling, distributed aggregation, and fault tolerance [11, 13]. The Apache Hadoop software library is a framework that facilitates the distributed processing of enormous data sets across clusters of computers. Apache Hadoop involves using simple programming models such as the Hadoop Distributed File System, which provides high-throughput access to application data and duplicates the data on multiple nodes so that failures of nodes containing a portion of the data do not affect the computations [14]. Apache Hadoop is designed to scale from single servers to thousands of machines, with each offering local computation and storage. Recently, Hadoop platform has been widely applied for cloud computing of biological, genomics, and drug design [15–19].

## 3. Materials

Several experimental studies have examined the genetic profiles of breast cancer samples, and most of them have focused on identifying the differential gene expressions between relapse and nonrelapse breast cancer samples [20–24]. Previous studies have not identified the correlations of differential gene expressions between different conditions; therefore, some abnormal regulation correlations may not have been detected. The proposed approach focuses on discovering the condition-specific correlations of gene expressions within biological pathways, thus providing a more macroscopic result than that from using a single-gene approach and potentially facilitating the discovery of greater biological meaning from a microarray data. We collected three breast-cancer-related data sets (GSE2034, GSE1456, and GSE4922) from multiple arrays from the GEO [25] to determine the correlations of differential gene expressions between relapse and nonrelapse breast cancer samples. In addition, GSE2109, of which numerous samples were available, was used for examining the performance of the cloud computing platform. Table 1 lists the information and materials extracted from the GEO.

## 4. Methods

The system flow of the proposed approach is depicted in Figure 1. Three microarray data sets of breast cancer were collected from the GEO. An Apache Hadoop cloud computing platform was implemented for decreasing computation time, and pathway information from the KEGG was applied for discovering meaningful biological correlations. The aim of this study was to develop a system that can identify abnormal regulations within the biological pathways of the relapse and nonrelapse breast cancer samples by using microarray data. These selected differential correlations can facilitate identifying the factors involved in breast cancer relapse. The system was divided into three major parts: receiving and preprocessing data, analyzing gene expression correlations, and mapping condition-specific correlations within biological pathways.

TABLE I: Statistical data of used data set.

GEO no.	Sample no. (nonrelapse/relapse)	Platform	Probe no.	Description	Reference no.
GSE2034	288 (179/109)	Affymetrix U133a (GPL96)	22,283	Breast cancer	[20]
GSE1456	159 (119/40)	Affymetrix U133a (GPL96)	22,283	Breast cancer	[21]
GSE4922	249 (160/89)	Affymetrix U133a (GPL96)	22,283	Breast cancer	[22]
GSE2109	2,158	Affymetrix U133a Plus 2.0 (GPL570)	54,675	Different types of cancer	[23]

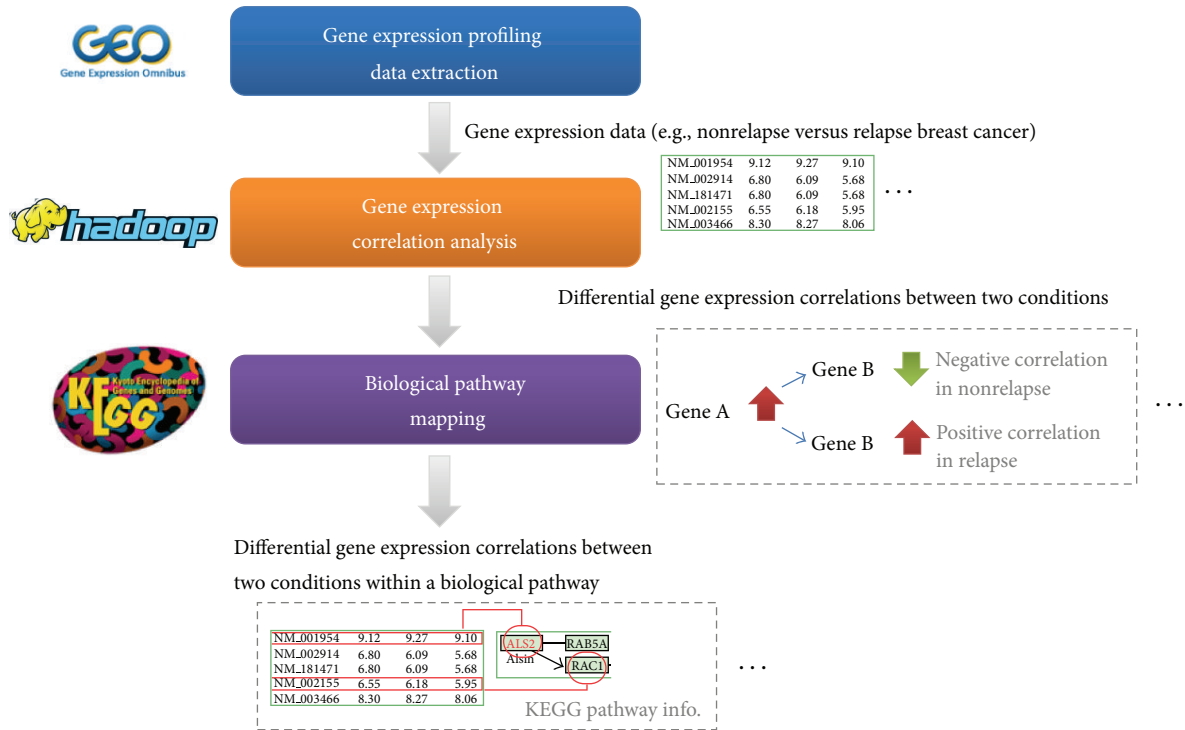


FIGURE 1: The flow chart of the system.

The microarray data were extracted from the GEO and segregated into a nonrelapse condition and relapse condition according to the sample descriptions. Gene profiling was performed using Affymetrix U133A arrays, and microarray quality control was performed using an R package *affyQCReport* [26]. The *gcrma* function of the R package *affy* was applied to normalize the CEL files by using the Robust Multiarray Averaging (RMA) method [27].

The master node divides the computation into multiple smaller subproblems and distributes them to worker nodes. The master node collects all of the answers submitted by the worker nodes and combines these answers into the output result. Two MapReduce programs were implemented to compute the linear regression validation and correlations for each pair of genes based on Hadoop Version 1.1.1. We used our programs to conduct a performance evaluation on a Hadoop cluster of 10 Xen virtual machines, where nine are the data nodes and one is the name node. Based on our settings, every two virtual machines were deployed on a physical machine. A physical machine was equipped with two CPUs with Intel Xeon E5504 4C 2.0 GHz and 16 GB of RAM; each virtual machine for the data nodes was equipped with a virtual CPU (VCPU) with two cores and 1GB of

RAM; the virtual machine for the name node was equipped with a VCPU with four cores and 4 GB of RAM. To evaluate the scalability of our MapReduce implementations, we also implemented two corresponding sequential Java programs as the basis for performance comparisons. In the performance evaluation experiments, we partitioned the data set into several pieces, uploaded them to the Hadoop file system, and submitted several MapReduce jobs to analyze the pieces of data.

The KEGG [25] Java APIs were used to obtain both pathway and interaction data. A gene map may possibly have more than one accession number, and we used the KEGG API to map the gene designations according to the KEGG Markup Language (KGML) file of each pathway. After mapping the gene expressions and interacting gene pairs, we identified substantial differences under the conditions of nonrelapsed and relapsed. The detailed steps are shown in Figure 2.

For example, the correlation of the gene pair A-B was calculated under different conditions. First, we used the expression intensity of genes A and B as the values of the *x*-axis and *y*-axis, respectively, for drawing a node. The number of nodes represented the number of samples. Subsequently,



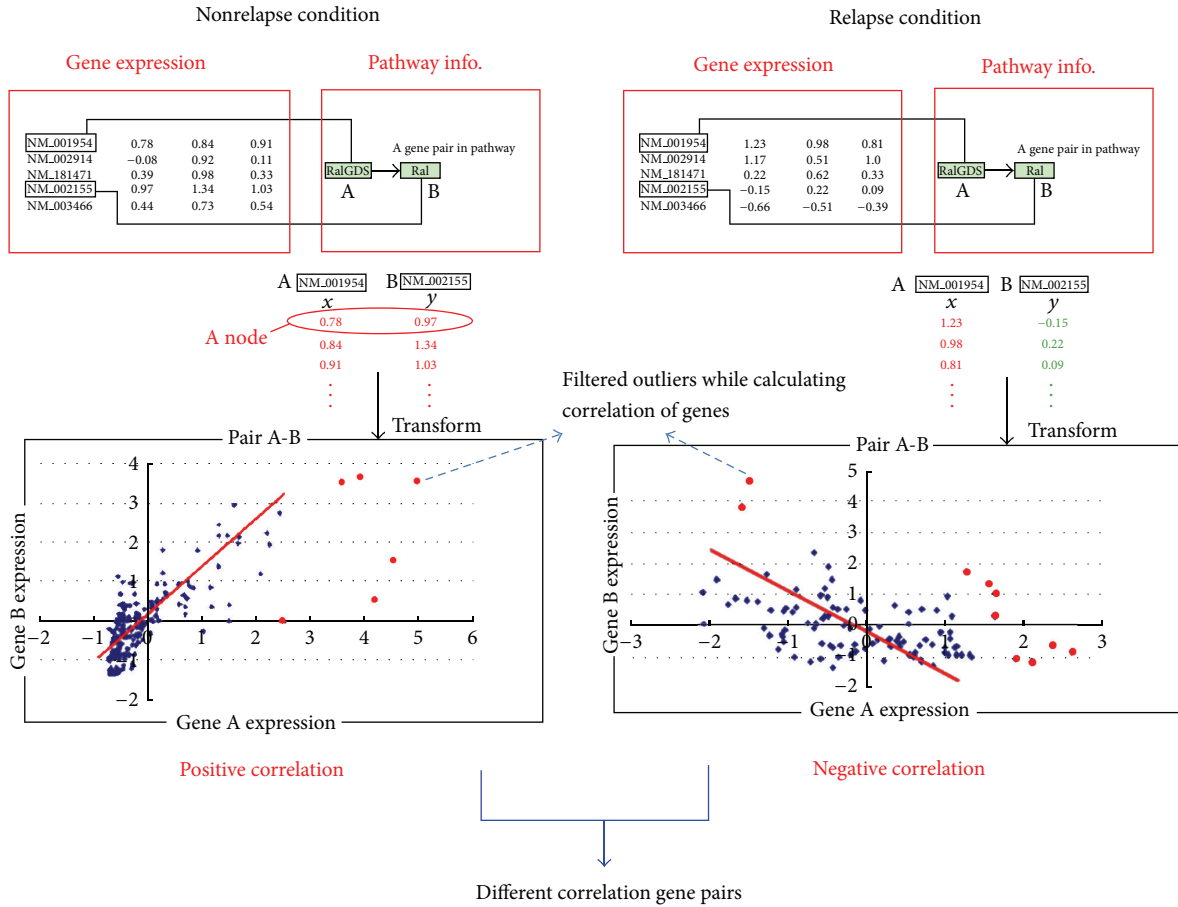


FIGURE 2: The process flow for identifying differential correlation of gene expression.

linear regression analysis was applied to discard outliers based on the least square value of each node. If the least square value of a node is greater than the average least square value-added threefold standard deviation, then the node would be discarded when calculating the correlation of gene expression. Eighty percent of the data was reserved, at least for the setting of the threshold, to make the data more representative.

After discarding the outliers, Pearson’s correlation coefficients were calculated for each gene pair. Pearson’s correlation,  $r$ , ranges between  $-1$  and  $1$ , and the greater the value of  $r$ , the stronger the coexpression between the members of the gene pair. In this study, coexpressed genes were defined as genes with a correlation greater than  $0.45$ , and reverse-expressed genes were defined as genes with a correlation less than  $-0.45$ . Unrelated genes were defined as genes with a correlation between  $-0.45$  and  $0.45$ . Genes with differential correlations between the nonrelapse and relapse conditions were collected when the difference of the correlation of the genes in the two conditions was greater than  $AVG + 3 * SD$  or less than  $AVG - 3 * SD$ , where  $AVG$  and  $SD$  are, respectively, the average and standard deviation of the correlations of differential gene expressions between the two conditions.

## 5. Results

**5.1. Performance Evaluation of Cloud Computing.** GSE2109 was used for evaluating the performance of the cloud computing platform. The size of the data set for analysis was  $54,675$  (probes)  $\times$   $2,158$  (samples) of floating numbers. Our MapReduce implementation for linear regression validation and correlation computation is a map-only program. It iteratively issues a MapReduce job by setting the number of Reducer to be zero, and the number of iterations depends on the input data size. That is, the outputs of the map tasks are written directly to the files system. At each iteration, our MapReduce implementation generates different number of map tasks for different input size. Based on our Hadoop MapReduce setting, it generates  $2, 3, 6, 12,$  and  $17$  map tasks for  $5,000, 10,000, 20,000, 40,000,$  and  $54,675$  probes, respectively.

Figure 3(a) shows the execution times of linear regression validation for each pair of genes by using the Hadoop MapReduce implementation and the sequential Java implementation, given different numbers of genes. Based on the measured performance values, the MapReduce program was generally faster than the sequential Java program that was executed on the name node, particularly when the input data became extremely large. The sequential Java program was

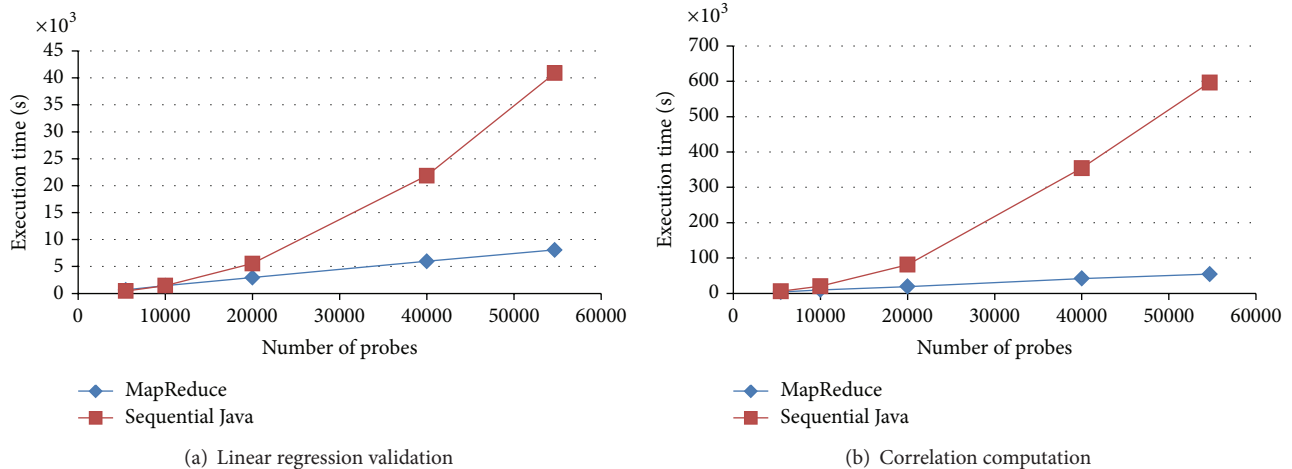


FIGURE 3: The execution times of linear regression validation and correlation computation by using the Hadoop MapReduce implementation and the sequential Java implementation, given different numbers of genes.

faster than the MapReduce program when the number of genes was smaller than 10,000. In addition, the execution time of the MapReduce program exhibited a linear growth for different numbers of genes, indicating that the Hadoop MapReduce framework was scalable in this case. By contrast, the execution time of the sequential Java program represented a quadric growth. Figure 3(b) shows a similar result to those of Figure 3(a), indicating that the Hadoop MapReduce framework is scalable for computing the correlations of gene pairs. The only difference is that the correlation computation jobs are more CPU-intensive than the linear regression validation jobs. The MapReduce program can accomplish tasks approximately 10 times as fast as the sequential Java program can.

**5.2. Analysis Results of Differential Gene Expression Correlations between Relapse and Nonrelapse Samples of Breast Cancer.** The number of differential gene expression correlations between the relapse and nonrelapse breast cancer samples from GSE2034, GSE1456, and GSE4922 is shown in Table S1 (see the Supplementary Material available online at <http://dx.doi.org/10.1155/2014/763237>), and Table 2 lists the differential gene expression correlations in the pathways, and there were 33, 32, and 50 gene expression correlations mapped in the KEGG pathways of GSE2034, GSE1456, and GSE4922, respectively. The pathways of the differential gene expression correlations are listed in Table 3. Several known cancer-related pathways were identified, including pathways in cancer, the PI3K-Akt signaling pathway, MAPK signaling pathway, and Wnt signaling pathway. The results showed that the number of correlations decreased as the number of samples increased. For example, there were 40 and 107 relapse samples in GSE1456 and GSE2034, respectively, and there were 3,630,906 and 1,595,963 positive correlations discovered within GSE1456 and GSE2034, respectively. These results indicated that using more samples may conduct more reliable correlations within pathways. As shown in Table 3,

two pathways, pathways in cancer and the PI3K-Akt signaling pathway, contained four correlated gene expressions between the relapse and nonrelapse breast cancer samples, and we used pathways in cancer for the demonstration and discussion of the discovery results presented in the following section.

**5.3. Differential Gene Expression Correlations between Relapse and Nonrelapse Samples of Breast Cancer in Pathways in Cancer of the KEGG.** As shown in Figure S1, four correlations of gene expression, NFKB2-PTGS2, JUN-MMP1, RUNX1-CEBPA, and JUN-FIGF, were identified in pathways in cancer. Table S2 shows the average log<sub>2</sub>-fold change in the gene expressions of NFKB2, PTGS2, JUN, MMP1, RUNX1, CEBPA, and FIGF between the relapse and nonrelapse samples, which were 0.02, -0.13, -0.16, 0.7, -0.08, -0.01, and -0.09, respectively. It shows that the differences of gene expression values between the two conditions were not substantial in these genes. However, the correlations between these genes were substantially different, and the differential correlations of NFKB2-PTGS2, JUN-MMP1, RUNX1-CEBPA, and JUN-FIGF were 0.37, 0.29, 0.29, and -0.27, respectively. The results imply that breast cancer recurrence may be induced by the abnormal regulations of these gene pairs, rather than their individual expression levels.

NF-kappa-B is a pleiotropic transcription factor and can be initiated by a vast array of stimuli related to many biological processes such as inflammation, immunity, differentiation, cell growth, tumorigenesis, and apoptosis through a series of signal transduction events. PTGS2 is regulated by specific stimulatory events and is responsible for the prostanoid biosynthesis involved in inflammation and mitogenesis. JUN is a putative transforming gene of avian sarcoma virus 17 and interacts directly with specific target DNA sequences to regulate gene expression. JUN is intronless and mapped to 1p32-p31, which is a chromosomal region involved in both translocations and deletions in human malignancies. The proteins of the matrix metalloproteinase (MMP)

TABLE 2: Correlations of gene expressions between nonrelapse and relapse samples in three data sets mapped in the KEGG pathway.

	Condition (no. of samples)	Number of correlated gene pairs mapped in the KEGG pathway		Number of differential correlations of gene pairs (AVG $\pm$ 3*SD)
		Positive (+) Cor. > 0.45	Negative (-) Cor. < -0.45	
GSE2034	Nonrelapse (179)	606	35	33
	relapse (107)	473	21	
GSE1456	Nonrelapse (119)	491	21	32
	relapse (40)	747	133	
GSE4922	Nonrelapse (160)	575	40	50
	relapse (89)	677	84	

TABLE 3: Pathways containing different correlated genes between relapse and nonrelapse breast cancer patients.

ID	Pathway names	Differential correlation genes (cor_Dif*)
hsa05200	Pathways in cancer	NFKB2 $\rightarrow$ PTGS2 (+0.37)
		JUN $\rightarrow$ MMP1 (+0.29)
		RUNX1 $\rightarrow$ CEBPA (+0.29)
		JUN $\rightarrow$ FIGF (-0.27)
hsa04151	PI3K-Akt signaling pathway	FGF4 $\rightarrow$ EGFR (-0.32)
		IRS1 $\rightarrow$ PIK3CB (-0.3)
		CSF1 $\rightarrow$ KIT (+0.3)
		EFNA1 $\rightarrow$ IGF1R (-0.28)
hsa04722	Neurotrophin signaling pathway	IRS1 $\rightarrow$ PIK3CG (+0.34)
		RPS6KA2 $\rightarrow$ NRAS (+0.3)
		IRS1 $\rightarrow$ PIK3CB (-0.3)
hsa04062	Chemokine signaling pathway	GNB5 $\rightarrow$ PIK3R5 (+0.31)
		CCL18 $\rightarrow$ XCR1 (-0.29)
hsa04150,	mTOR signaling pathway	IRS1 $\rightarrow$ PIK3CG (+0.34)
hsa04910,	Insulin signaling pathway	
hsa04930,	Type II diabetes mellitus	
hsa04960	Aldosterone-regulated sodium reabsorption	
hsa04010	MAPK signaling pathway	DUSP2 $\rightarrow$ MAPK8 (-0.34)
hsa04060	Cytokine-cytokine receptor interaction	IL17A $\rightarrow$ IL17RA (-0.32)
hsa04310	Wnt signaling pathway	SFRP5 $\rightarrow$ WNT11 (+0.3)
hsa04520	Adherens junction	WAS $\rightarrow$ ACTB (-0.39)
hsa04530	Tight junction	PRKCQ $\rightarrow$ ACTB (-0.34)
hsa04612	Antigen processing and presentation	HLA-E $\rightarrow$ KIR2DS1 (-0.29)
hsa04620	Toll-like receptor signaling pathway	RIPK1 $\rightarrow$ TRAF6 (+0.31)
hsa04630	Jak-STAT signaling pathway	STAT6 $\rightarrow$ SOCS1 (+0.31)
hsa04666	Fc gamma R-mediated phagocytosis	WASF3 $\rightarrow$ ARPC2 (+0.3)
hsa04725	Cholinergic synapse	GNB5 $\rightarrow$ PIK3R5 (+0.31)
hsa05012	Parkinson's disease	SLC25A4 $\rightarrow$ CYCS (+0.28)
hsa05020	Prion diseases	PRNP $\rightarrow$ BAX (+0.29)
hsa05152	Tuberculosis	MAPK3 $\rightarrow$ IL23A (+0.31)
hsa05211	Renal cell carcinoma	EGLN3 $\rightarrow$ EPAS1 (+0.28)

\*cor\_Dif: average correlation of genes in nonrelapse samples – average correlation of genes in relapse samples.

family are involved in the breakdown of the extracellular matrix in normal physiological processes, such as tissue remodeling, embryonic development, and disease processes, such as arthritis and metastasis. Core binding factor (CBF) is a heterodimeric transcription factor binding to the core element of many enhancers and promoters. Chromosomal translocations involving CBF are well documented and have been discovered to be associated with several types of leukemia. The protein encoded by FIGF is a member of the platelet-derived growth factor/vascular endothelial growth factor (PDGF/VEGF) family and is active in angiogenesis, lymphangiogenesis, and endothelial cell growth.

As mentioned, most of the discovered genes were related to cancer pathways and cancer regulations, including immunity, cell growth, tumorigenesis, and apoptosis. Although the gene expressions between relapse and nonrelapse samples were not substantially different, their correlations were substantially different. Thus, we believe these regulations of genes may be essential for regulating breast cancer recurrence.

## 6. Discussion and Conclusions

This study proposes an approach for discovering condition-specific correlations of gene expressions. An Apache Hadoop cloud computing platform was implemented to reduce the computation time. By using microarray data from the GEO, we discovered numerous differential gene expression correlations between the nonrelapse and relapse conditions of breast cancer. The results show that breast cancer recurrence is highly associated with the abnormal regulation of these gene pairs, rather than their individual expression levels. The pathways in cancer specifically show that NKFB2-PTGS2, JUN-FIGF, and RUNX1-CEBPA possess higher correlations of gene expression in nonrelapse samples and that JUN-MMP1 possesses higher correlations of gene expression in relapse samples. In addition, using the cloud computing technology successfully reduces the time required for conducting gene expression correlation analysis of microarray data, and it can be further applied for analyzing the correlation between different transcript isoforms using RNA sequencing data, which is helpful for deciphering the regulatory mechanisms of genes. The results show that our method is effective and can be extended to areas of biological analysis beyond that of breast cancer nonrelapse and relapse. We believe that the proposed method is effective for identifying meaningful biological regulation patterns between conditions and can be applied for developing coexpression networks and protein-protein interactions in the future.

## Conflict of Interests

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

## Authors' Contribution

Tzu-Hao Chang and Shih-Lin Wu contributed equally to this work.

## Acknowledgments

The authors thank the National Science Council of the Republic of China for financially supporting this research under Contracts no. NSC 101-2221-E-008-125-MY3, NSC 101-2923-E-182-001-MY3, NSC 102-2221-E-182-031, and NSC102-2221-E-266-005-. This work was supported in part by Taipei Medical University under the Grant TMUI01-AE1-B44.

## References

- [1] J. K. Cowell and L. Hawthorn, "The application of microarray technology to the analysis of the cancer genome," *Current Molecular Medicine*, vol. 7, no. 1, pp. 103–120, 2007.
- [2] X. Wang, Y. Gong, D. Wang et al., "Analysis of gene expression profiling in meningioma: deregulated signaling pathways associated with meningioma and EGFL6 overexpression in benign meningioma tissue and serum," *PLoS One*, vol. 7, no. 12, Article ID e52707, 2012.
- [3] P. E. Colombo, F. Milanezi, B. Weigelt, and J. S. Reis-Filho, "Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction," *Breast Cancer Research*, vol. 13, no. 3, p. 212, 2011.
- [4] J. Botling, K. Edlund, M. Lohr et al., "Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation," *Clinical Cancer Research*, vol. 19, no. 1, pp. 194–204, 2013.
- [5] P. Kwiatkowski, P. Wierzbicki, A. Kmiec, and J. Godlewski, "DNA microarray-based gene expression profiling in diagnosis, assessing prognosis and predicting response to therapy in colorectal cancer," *Postępy Higieny i Medycyny Doświadczalnej*, vol. 66, pp. 330–338, 2012.
- [6] A. E. Frolov, A. K. Godwin, and O. O. Favorova, "Differential gene expression analysis by DNA microarrays technology and its application in molecular oncology," *Molekulyarnaya Biologiya*, vol. 37, no. 4, pp. 573–584, 2003.
- [7] A. Gusnanto, S. Calza, and Y. Pawitan, "Identification of differentially expressed genes and false discovery rate in microarray studies," *Current Opinion in Lipidology*, vol. 18, no. 2, pp. 187–193, 2007.
- [8] H. Lee, Y. Yang, H. Chae et al., "BioVLAB-MMIA: a cloud environment for microRNA and mRNA integrated analysis (MMIA) on Amazon EC2," *IEEE Transactions on NanoBioscience*, vol. 11, no. 3, pp. 266–272, 2012.
- [9] C. Bernau and A. L. Boulesteix, "Application of microarray analysis on computer cluster and cloud platforms," *Methods of Information in Medicine*, vol. 52, no. 1, pp. 65–71, 2012.
- [10] L. Zhang, S. Gu, Y. Liu, B. Wang, and F. Azuaje, "Gene set analysis in the cloud," *Bioinformatics*, vol. 28, no. 2, pp. 294–295, 2012.
- [11] T. Gunarathne, T. Wu, J. Y. Choi, S. Bae, and J. Qiu, "Cloud computing paradigms for pleasingly parallel biomedical applications," *Concurrency Computation Practice and Experience*, vol. 23, no. 17, pp. 2338–2354, 2011.
- [12] M. Kanehisa, S. Goto, M. Hattori et al., "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Research*, vol. 34, pp. D354–D357, 2006.
- [13] M. C. Schatz, B. Langmead, and S. L. Salzberg, "Cloud computing and the DNA data race," *Nature Biotechnology*, vol. 28, no. 7, pp. 691–693, 2010.

- [14] J. Ekanayake, T. Gunarathne, and J. Qiu, "Cloud technologies for bioinformatics applications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 6, pp. 998–1011, 2011.
- [15] W. P. Chen, C. L. Hung, S. J. Tsai, and Y. L. Lin, "Novel and efficient tag SNPs selection algorithms," *Bio-Medical Materials and Engineering*, vol. 24, no. 1, pp. 1383–1389, 2014.
- [16] C. L. Hung and Y. L. Lin, "Implementation of a parallel protein structure alignment service on cloud," *International Journal of Genomics*, vol. 2013, Article ID 439681, 8 pages, 2013.
- [17] C. L. Hung and C. Y. Lin, "Open reading frame phylogenetic analysis on the cloud," *International Journal of Genomics*, vol. 2013, Article ID 614923, 9 pages, 2013.
- [18] C. L. Hung and G. J. Hua, "Cloud computing for protein-ligand binding site comparison," *BioMed Research International*, vol. 2013, Article ID 170356, 7 pages, 2013.
- [19] C. H. Hsu, C. Y. Lin, M. Ouyang, and Y. K. Guo, "Biocloud: cloud computing for biological, genomics, and drug design," *BioMed Research International*, vol. 2013, Article ID 909470, 3 pages, 2013.
- [20] Y. Wang, J. G. M. Klijn, Y. Zhang et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.
- [21] Y. Pawitan, J. Bjohle, L. Amler et al., "Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts," *Breast Cancer Research*, vol. 7, no. 6, pp. R953–R964, 2005.
- [22] A. V. Ivshina, J. George, O. Senko et al., "Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer," *Cancer Research*, vol. 66, no. 21, pp. 10292–10301, 2006.
- [23] IGC's Expression Project for Oncology (expO), <http://www.intgen.org/>.
- [24] S. A. Norman, S. L. Potashnik, M. L. Galantino, A. M. De Michele, L. House, and A. R. Localio, "Modifiable risk factors for breast cancer recurrence: what can we tell survivors?" *Journal of Women's Health*, vol. 16, no. 2, pp. 177–190, 2007.
- [25] R. Edgar and T. Barrett, "NCBI GEO standards and services for microarray data," *Nature Biotechnology*, vol. 24, no. 12, pp. 1471–1472, 2006.
- [26] C. Parman and C. Halling, AffyQCReport: QC Report Generation for affyBatch objects, 2005.
- [27] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "Affy—analysis of Affymetrix GeneChip data at the probe level," *Bioinformatics*, vol. 20, no. 3, pp. 307–315, 2004.

## Research Article

# Gene Prioritization of Resistant Rice Gene against *Xanthomonas oryzae* pv. *oryzae* by Using Text Mining Technologies

Jingbo Xia,<sup>1,2</sup> Xing Zhang,<sup>3</sup> Daojun Yuan,<sup>4</sup> Lingling Chen,<sup>5</sup>  
Jonathan Webster,<sup>2,3</sup> and Alex Chengyu Fang<sup>2,3</sup>

<sup>1</sup> College of Science, Huazhong Agricultural University, Wuhan 430070, Hubei, China

<sup>2</sup> Department of Chinese, Translation and Linguistics, City University of Hong Kong, Kowloon, Hong Kong

<sup>3</sup> The Halliday Centre for Intelligent Applications of Language Studies, City University of Hong Kong, Kowloon, Hong Kong

<sup>4</sup> College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, Hubei, China

<sup>5</sup> College of Life Science, Huazhong Agricultural University, Wuhan 430070, Hubei, China

Correspondence should be addressed to Alex Chengyu Fang; [acfang@cityu.edu.hk](mailto:acfang@cityu.edu.hk)

Received 4 May 2013; Revised 26 October 2013; Accepted 10 November 2013

Academic Editor: Huiru Zheng

Copyright © 2013 Jingbo Xia et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To effectively assess the possibility of the unknown rice protein resistant to *Xanthomonas oryzae* pv. *oryzae*, a hybrid strategy is proposed to enhance gene prioritization by combining text mining technologies with a sequence-based approach. The text mining technique of term frequency inverse document frequency is used to measure the importance of distinguished terms which reflect biomedical activity in rice before candidate genes are screened and vital terms are produced. Afterwards, a built-in classifier under the chaos games representation algorithm is used to sieve the best possible candidate gene. Our experiment results show that the combination of these two methods achieves enhanced gene prioritization.

## 1. Introduction

Due to the availability of abundant genomic resources, rice has become a model species for the genomic study. Taking into account that rice has been the main food for a large section of the world population, research issues related to yielding and antidisease have drawn much attention [1]. Bacterial blight, caused by *Xanthomonas oryzae* pv. *oryzae* (*Xoo*), is a worldwide devastating disease, which is second only to the *Pyricularia grisea*, and causes yield losses ranging from 20% to 30%, and in some areas of Asia the loss can be as high as 50% [2].

Traditionally, bacterial blight resistance genes have been cloned by a map-based cloning approach. To date, thirty bacterial blight resistance genes in rice have been identified. Among them, six genes, namely, Xa1, Xa5, Xa13, Xa21, Xa3/Xa26, and Xa27, have been reported to be isolated for bacterial blight resistance [3–6]. While on one hand the results of resistant gene discovery with map-based cloning

approach are accurate, these laboratory experiments take years of endeavor and a huge amount of input in terms of human and material resources. It is important to find a more effective way to locate vital resistant genes.

For a quicker discovery of R genes, the sequence-based approach in bioinformatics is an alternative strategy. In our previous work, Xia et al. [7] presented a novel disease-resistant gene predictor by using chaos games representation (CGR), and the predictor achieved a high accuracy of 98.13% by using a small database with 107 samples. Moreover, Xia et al. also applied this classifier onto the whole KOME database (Knowledge-based Oryza Molecular Biological Encyclopedia, [ftp://cdna01.dna.affrc.go.jp/pub/data//20081001/20081001/INE\\_FULL\\_SEQUENCE\\_DB\\_20081001.zip](ftp://cdna01.dna.affrc.go.jp/pub/data//20081001/20081001/INE_FULL_SEQUENCE_DB_20081001.zip)) and located the top 10 candidate genes, most of which own abundant annotation information in conserved domain information. Unfortunately, direct application of the classifier to the whole database shows a lack of confidence or reliability.

Additionally, the text mining strategy represents another effective way to improve the efficiency of gene discovery. This strategy usually adopts gene prioritization information among texts to find genes that are possibly related to R gene. For better use of the textual information about the gene, both structural and domain information for *Xoo*-resistant genes should be considered. According to the experimental results in literature [4], most of *Xoo*-resistant genes encode proteins containing conserved nucleotide binding site (NBS) domain and/or leucine-rich repeat (LRR) domain [8] or encode LRR receptor kinase-like proteins. These phenomena suggest a possible internal relation between the gene function and gene structure and offer clues for the text mining strategy [7].

Unfortunately, though both the sequence-based approach and the text mining strategy aim to improve the efficiency of discovery of the targeted R gene against *Xanthomonas oryzae pv. oryzae* (*Xoo*) in rice, the two methods have their own disadvantages. For example, the precision of the sequence-based methods is not high while the recall rate of text mining methods is low. It still has room for enhancement. Henceforth, the purpose of the research to be reported next is to integrate the above two methods into a combined gene discovery strategy so as to achieve a better precision of sequence-based methods and a higher recall of text mining methods as well.

In this paper, large-scale gene prioritization is enhanced with biomedical text mining technology. After extracting the 31 most distinguished terms in Medline files with term frequency-inverse document frequency, we retrieved 443 candidate proteins with 31 terms. With the classifier built in [7], 74 highly candidate proteins were screened. After searching in Conserved Domains and Protein Classification [9] (<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>), most of these proteins are proved to be related to *Xoo*-resistant gene in structure and super family information.

## 2. Related Work

Gene discovery based on bioinformatics and text mining are all related to gene prioritization. The definition of a standard definition of gene prioritization is given in [10], that is, given disease  $D$ , candidate gene set  $C$ , and training data set  $T$ ; input all these data to a predictor or classifier and the gene prioritization method will compute a score for each of the candidate genes. Genes with higher scores are those with higher probability of being disease  $D$ .

According to the type of input data, methods for gene prioritization can be classified into text and data mining methods, as well as network-based methods. Text and data mining methods use training data that includes gene expression [11–13], phenotypic data [14], PubMed abstracts [11], spatial gene expression profiles [12], gene ontology, and other resources [15, 16]. Subsequent computation then will produce scores of candidate genes by mining the genomes or processing currently available biomedical literature. Network-based methods use biological networks [17, 18] as the basis of the prioritization process. There are also network-based methods that combine data and text mining techniques to improve system performance [13, 19].

We can also divide current gene prioritization tools into two classes from the perspective of their working principles into functional annotation-based [11, 14, 20–22] and sequence feature based [15, 23]. There are also some studies, like [13], that try to combine these two methods together. Functional annotation tools are usually based on gene expression data. Its underlying principle is that; if a gene is found to be coexpressed with other genes that are involved in a given biological process, this gene can be predicted to be involved in the same process [24]. This principle proceeds from the observation that there is a strong correlation between co-expressions and functional relatedness [24]. The biggest problem for the functional annotation based method is annotation bias, as some genes lack sufficient annotation while others are annotated with abundant information [13]. On the other hand, sequence-based methods utilize information that can be readily computed from the gene sequence, such as gene length, homology and base composition [13]. This method avoids the limitation of annotation bias by making use of intrinsic characteristics of genes. However, it is based on the assumption that these genes have potential involvement in general diseases only rather than some specific disease in which the user is interested [13].

Gene seeker [14] is a useful tool to generate a starting list of candidate genes involved in human genetic disorders by gathering positional and expression/phenotypic data from 9 databases automatically. As a controlled vocabulary of anatomical terms, eVOC anatomical system ontology is designed in [11] to integrate clinical and molecular data through a combination of text and data mining methods. The candidate disease genes are selected according to their expression profiles by matching tissues associated with diseases to genes expressed in the tissues. Piro et al. [12] proves that spatially mapped gene expressions are suitable for candidate gene prioritization. The results demonstrate that spatial gene expression patterns have been successfully exploited to predict gene-phenotype associations for both mouse phenotypes and human central nervous system-related Mendelian disorders.

PROSPECTR [15] is a classifier based on sequence features to rank genes involved in Mendelian and oligogenic disorders. It uses a collection of features representing the structure, content, and phylogenetic extent of candidate genes without prior detailed phenotypic knowledge of the disease. In 2005, SUSPECT [13] combined annotation- and sequence-based approaches to prioritize genes on the principle that genes involved in that disease tend to share the same or similar annotation, so as to reflect common biological pathways. It tries to achieve higher precision of annotation-based methods and the better recall of sequence-based methods through four lines of evidence to score genes, that is, sequence features, extent of coexpressions, domain information, and semantic similarity.

## 3. Materials and Methods

**3.1. Data Set Construction.** To prepare the data set for literature text mining, texts are collected from NCBI PubMed data base (<http://www.ncbi.nlm.nih.gov/pubmed>) with MedLine

TABLE 1: Searching strategy for PubMed literature in rice.

Searching content	PubMed hit
Binding	1428
Catabolism	47
Expression	5170
Localization	816
Phosphorylation	226
Regulation	4067
Transcription	2624
All of the above events	6810
<i>Xanthomonas oryzae pv. oryzae</i> or <i>Xoo</i>	402
( <i>Oryza sativa</i> ) or rice	33349

format. In order to evaluate the effectiveness of terms for future extraction, ten sets of Medline texts were collected with different keywords, each of which represented fundamental biological function or event for rice gene/protein in literature. As can be seen from Table 1, the first document has a collection of literature related to binding events for rice, and 1428 hits were found, and the following documents collect corresponding biological event-related papers for rice, including catabolism, expression, localization, phosphorylation, regulation, transcription, all events, *Xoo*-related, and rice-related. Among these features, the first seven represent standard active biomedical events, the eighth one is the sum of the above events, and the last two features focus on *Xoo* gene and rice. In sum, the ten text databases reflect sufficient importance and relevance of the active *Xoo* resistant gene in rice.

### 3.2. Text Mining Based Approach: Choosing Controlled Phrase and Evaluation with Term Frequency-Inverse Document Frequency (TF \* IDF)

**3.2.1. Preparation of Phrase Dictionary for Candidate Gene Annotation.** In order to extract candidate genes from the whole data base, a phrase dictionary for candidate gene annotation is built on the annotation line in FASTA file for rice. A record in its standard format is shown as follows.

```
>gi|313507159|pdb|ICCR|A Chain A, Structure Of
Rice Ferricytochrome C At 2.0 Angstroms Resolution
```

There are 5 sections of information annotated in each record line.

- (1) ">gi" indicates the beginning of annotation line in NCBI.
- (2) "|313507159|" indicates the accession number in NCBI.
- (3) "pdb" indicates database Protein Data Bank (<http://www.rcsb.org/pdb/home/home.do>).
- (4) "|ICCR|" indicates the protein name in pdb database.
- (5) "A Chain A, Structure Of Rice Ferricytochrome C At 2.0 Angstroms Resolution" provides additional description.

In essence, the phrase dictionary collects information that can be automatically extracted from Section 5. The basic principle is to extract meaningful phrases. In the examples above for record 1, the 5th section is "A Chain A, Structure Of Rice Ferricytochrome C At 2.0 Angstroms Resolution"; there are two parts separated by a comma. In these cases, they will be considered as two separate phrases, that is "A Chain A" and "Structure Of Rice Ferricytochrome C At 2.0 Angstroms Resolution".

However, for those fragments extracted, some are meaningful themselves and some do not have any specific meaning. For example, in record 2, record 3, and record 4, there are "unknown protein", "hypothetical protein", and "unnamed protein" used in Section 5 for description. In these cases, they are not collected into the phrase dictionary because they lack specific reference. From the original annotation line of FASTA file for each rice protein, 12037 phrases were chosen on the basis of the above rules.

**3.2.2. Phrases Evaluation and Sequences Retrieving.** The term frequency-inverse document frequency (TF \* IDF) is a statistical measure for evaluating the importance or relevance of a specific word to a document among a series of documents or corpus.

For a given term  $t$  and a specific document  $d$  among a series of document  $D$ , we denote  $tf(t, d)$  as term frequency which means the occurrence of term  $t$  in document  $d$  and denote  $idf(t, D)$  as inverse document frequency; that is,

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}| + 1}. \quad (1)$$

Here,  $idf(t, D)$  is a measure of the general importance of the term  $t$  in documents  $D$ . Meanwhile, the TF \* IDF is defined as

$$TF * IDF(t, d, D) = tf(t, d) \times idf(t, D). \quad (2)$$

The smaller value of TF \* IDF shows more relevance between term  $t$  and document  $d$ . Therefore, related protein sequences can be retrieved according to vital phrases in conjunction with TF \* IDF value, after ranking top vital phrases among phrases in the built dictionary.

**3.3. Gene Priority with Hybrid Strategy.** We combine the text mining strategy and sequence-based approach to propose a hybrid algorithm for gene prioritization. See Algorithm 1.

Here, candidate proteins are chosen according to meaningful annotation screening. Afterwards, the candidate sequences are sent into a built-in classifier, and predictive values will be obtained. This classifier is a sequence-based predictor developed by Jingbo et al. [7] and is available for public use. In this classifier, proteins with a positive value will be regarded as possible *Xoo*-resistant rice gene.

Those proteins passing both tests in text-mining screening and the built-in classifier are chosen as the highly possible *Xoo*-resistant rice gene. Finally, standard bioinformatics methods are applied onto those positive samples for further evaluation.



*Step 1.* Collect NCBI literature in the rice research field, denote the text database as  $d_j$ , here  $d_{1,2,\dots,10}$  = “rice”, “Event”, “Binding”, “Catabolism”, “Expression”, “Localization”, “phosphorylation”, “regulation”, “transcript”, “Xoo”;

*Step 2.* Build phrase dictionary, denote the terms as  $t_i$ .

*Step 3.* Evaluate the relevance between  $t_i$  and  $d_j$  by computing  $TF * IDF(t_i, d_j, D)$ , here  $D$  is the total text data set.

*Step 4.* Rank important  $t_i$ .

*Step 5.* Retrieve protein in NCBI with annotation include  $t_i$ .

*Step 6.* Rank candidate protein by using the built-in classifier [17] which is sequence-based.

*Step 7.* Use Conserved Domain Data (CDD) and Gene Ontology (GO) to verify the result of prioritization.

ALGORITHM 1: Gene prioritization algorithm.

Thus, by combining both text mining candidate selection approach and sequence-based classifier, a novel hybrid strategy is proposed for gene priority with a specific function protein.

## 4. Results and Discussion

*4.1. Experiments Results.* As illustrated in Section 3, a phrase dictionary is built based on the annotation file for the whole rice protein sequence. The whole dictionary comprises 12037 terms, and  $t_i$  ( $i = 1, 2, \dots, 12037$ ) is the  $i$ th term,  $d_j$  ( $j = 1, 2, \dots, 10$ ) refers to “rice”, “rice event”, “blin”, “catabolism”, “expression”, “localization”, “phosphory”, “regulation”, “transcription”, and “*Xanthomonas oryzae versus oryzae*”, respectively, and  $D = d_1, d_2, \dots, d_{10}$ . So  $TF * IDF(t_i, d_j, D)$  is counted. The sample results are listed in Table 2.

In order to screen the key phrases with the most general importance, a voting strategy is used. For each  $t_i$  ( $i = 1, 2, \dots, 12037$ ) and  $d_j$  ( $j = 1, 2, \dots, 10$ ),  $TF * IDF(t_i, d_j, D)$  represents the relevance between  $t_i$  and  $d_j$ , the smaller the value, the higher the relevance, whereas zero means the nonexistence of  $t_i$  in  $d_j$ . For each fixed  $d_j$ , the value of  $TF * IDF(t_i, d_j, D)$  is sorted and the relevance of  $t_i$  and  $d_j$  is ranked, numbered as  $\text{Rank}(t_i, d_j, D)$ . The voting strategy is to choose  $t_i$  which satisfies

$$\# \{d_{j,(j=1,2,\dots,10)} \text{Rank}(t_i, d_j, D) < 100\} > 5, \quad (3)$$

where  $\#$  means the order/scale of the set. By using this voting strategy, only those  $t_i$  which are in the top 100 among at least 6 out of 10 documents can be chosen as the key phrases. Taking the construction rule of documents corpus into consideration, the majority agreement of relevance ensures the most general importance of chosen  $t_i$ .

After voting, thirty key phrases are chosen, which are “CR4”, “thioesterase”, “WRKY2”, “exonuclease-1”, “fibrillar”, “kinase-like”, “WRKY10”, “WRKY30”, “AML1”, “arginase”, “constans”, “decoy”, “glutaredoxin-like”, “glutathione-S-transferase”, “H2A”, “Metalloendopeptidase”, “PDR20”, “RISBZ5”, “SNF2P”, “YY2”, “CIA”, “CR9”, “EL3”, “MtN21”, “NPKL1”, “prohibitin”, “Ramy1”, “UreD”, “UreF”, and “UreG”, respectively. All of the key phrases with greatest importance are listed in Table 3.

By tracing these key phrases in FASTA annotation, 423 rice proteins are retrieved, each of which includes at least one key phrase in the annotation line. For simplicity and

clarity, the result of a small subset with 10 retrieved samples is listed in Table 4. Here, the entries in the first column refer to the NCBI numbers, the second column contains the key phrases, and the third column contains the corresponding gene annotations.

As an example, the GI code for the first sample sequence is 15721862 and its annotation line in FASTA file is “>gi15721862 dbj BAB68389.1 CR4 [*Oryza sativa*]”, which includes the phrase “CR4”.

Through the text mining approach, 423 rice protein sequences were chosen as the candidate genes which are regarded as relevant and functionally active. Finally, we test the *Xoo*-resistance for each candidate by using the built-in CGR classifier, and 74 sequences passed the testing procedure. Thus, they show possible positive effects on resistance with the screening ratio of 17.49%. With these 74 proteins, we obtain a candidate gene data set for resistant gene against *Xanthomonas oryzae pv. oryzae* (*Xoo*) in rice. In the following section, we aim to identify its positive resistance so as to obtain useful material for rice breeding.

*4.2. Validation Evaluation of the Candidate Gene Data Set by Conserved Domain Data and Gene Ontology Matching Results.* To evaluate the performance of gene prioritization method, the traditional method is map cloning which is time consuming, as mentioned in Section 1. Therefore, some popular bioinformatics validation methods are used. We use Conserved Domain Data (CDD) [9] and Gene Ontology (GO) [25] to observe information hidden in each gene sequence of candidate gene data set by checking both in conserved construction and function.

First, to observe the structure information of 74 screened proteins, CDD matching results are shown in Table 6. Hits in multidomain and super family in Table 6 clearly show a consistent tendency for the proteins we obtained. Most of the 74 proteins show a high consistency in CDD information. For simplicity and clarity, the domain information of the top 10 proteins is listed as below: the domain hits consist of 6 categories, that is, PLN00113, PKclike super family, LRRNT 2, PLN03150, PKc, and LRRRI, which are closely relevant with leucine-rich repeat or protein kinase. As mentioned in Section 1, most of *Xoo*-resistant genes encode proteins containing conserved nucleotide binding site (NBS) domain and/or leucine-rich repeat (LRR) domain or encode LRR receptor kinase-like proteins. Taking this evidence into account, five domains or super families (PLN03150 excluded)

TABLE 2: Sample list of evaluation of vital phrase by TF\*IDF ( $t_i, d_j, D$ ).

$t_i$	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$	$d_8$	$d_9$	$d_{10}$
WRKY14	0.79	0.79	0	0	0.79	0	0	0.79	0.79	0
RadA	3.02	2.41	2.41	0	2.41	2.41	0	2.41	0	0
UreD	0.6	0.6	0.6	0.6	0.6	0	0	0.6	0	0
CC-NBS-LRR	4.22	2.41	1.21	0	2.41	0	0	0.6	1.21	0
Urease	18.45	1.35	0.9	0.45	0.9	0	0	0.45	0.45	0
Hd6	7.85	3.02	0	0	3.02	0	0.6	3.02	0.6	0
Carboxypeptidase	15.85	8.56	0.32	0.95	6.02	0.95	0	5.07	0.63	0
EUI	2.2	1.8	0.2	0.2	1.6	0.2	0	1.6	0.6	0.2
H2A	1.9	1.59	0.32	0	0.95	0.32	0.32	0.32	0.95	0
Prolin	34.73	22.11	2.85	0.19	20.97	2.85	0.57	16.99	16.61	0.57
Polypeptide	36.82	18.6	5.69	0.19	14.14	2.85	1.14	8.92	7.78	0.66
Reductase	110.37	47.45	7.21	1.23	37.3	5.31	0.76	26.19	15.75	0.66

( $d_{1,2,\dots,10}$  = "rice", "event", "binding", "catabolism", "expression", "localization", "phosphorylation", "regulation", "transcript", and "Xoo".)

TABLE 3: Voting results of key phrases with greatest importance.

Term	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$	$d_8$	$d_9$	$d_{10}$	Vote
CR4	219	7	13	73	7	2	7	3	1	1	9
Thioesterase	106	6	1	63	6	1	6	14	20	8	9
WRKY2	88	62	4	65	74	9	130	91	96	21	9
Exonuclease-1	1	1	14	74	1	20	133	6	6	130	8
Fibrillarlin	2	2	15	75	2	21	134	7	7	131	8
Kinase-like	204	149	2	64	76	16	40	16	2	79	8
WRKY10	3	3	16	76	3	247	267	10	9	43	8
WRKY30	4	4	17	77	4	248	268	11	10	44	8
AML1	95	16	42	98	15	254	274	31	29	148	7
Arginase	91	60	19	32	66	292	310	12	11	133	7
Constans	96	17	43	99	16	255	275	32	30	149	7
Decoy	206	5	18	78	5	22	135	8	8	132	7
Glutaredoxin-like	6	9	35	94	11	38	149	20	362	376	7
Glutathione-S-transferase	227	181	32	91	196	17	12	92	3	7	7
H2A	103	145	5	66	75	10	20	4	95	211	7
Metalloendopeptidase	54	15	41	97	14	39	150	21	363	377	7
PDR20	7	10	36	95	12	252	272	29	27	146	7
RISBZ5	40	58	84	138	69	76	175	81	86	203	7
SNF2P	8	11	37	96	13	253	273	30	28	147	7
YY2	41	59	85	139	70	77	176	82	87	204	7
CIA	297	168	33	92	166	4	8	156	22	10	6
CR9	224	61	86	140	71	78	177	83	88	205	6
EL3	71	117	20	79	68	294	47	126	14	136	6
MtN21	55	85	462	463	24	43	153	24	25	144	6
NPKL1	5	8	445	446	22	41	65	17	360	374	6
Prohibitin	202	148	6	67	26	260	92	151	5	20	6
Ramy1	58	88	48	104	99	315	332	37	33	152	6
UreD	9	12	38	38	8	249	269	26	365	379	6
UreF	10	13	39	39	9	250	270	27	366	380	6
UreG	11	14	40	40	10	251	271	28	367	381	6

TABLE 4: The sample of retrieving protein sequences.

NCBI	Term	Annotation
15721862	CR4	>gi 15721862 dbj BAB68389.1 CR4 [Oryza sativa]
56201806	Thioesterase	>gi 56201806 dbj BAD73256.1 putative acyl-(acyl carrier protein) thioesterase [Oryza sativa Japonica Group]
50843956	WRKY2	>gi 50843956 gb AAT84156.1 transcription factor WRKY24 [Oryza sativa Indica Group]
54111120	Exonuclease-1	>gi 54111120 dbj BAD60834.1 exonuclease-1 [Oryza sativa Japonica Group]
18071363	Brillarin	>gi 18071363 gb AAL58222.1 AC09088225 putative brillarin [Oryza sativa Japonica Group]
1586408	Kinase-like	>gi 1586408 prf 2203451 A receptor kinase-like protein
50843970	WRKY10	>gi 50843970 gb AAT84163.1 transcription factor WRKY100 [Oryza sativa Indica Group]
58042751	WRKY30	>gi 58042751 gb AAW63719.1 WRKY30 [Oryza sativa Japonica Group]
52076187	AML1	>gi 52076187 dbj BAD46727.1 putative AML1 [Oryza sativa Japonica Group]
30134457	Arginase	>gi 30134457 gb ADK74000.1 arginase [Oryza sativa Indica Group]

TABLE 5: Multi Domain and Super family Data for Top 10 Sequence in CDD Hit.

Query	Hit type	Short name	Description	Evidence?
Q#1->gi 53793299	Multidom	PLN00113	LRR	Yes
	Superfamily	PKc.like superfamily	LRR and kinase	Yes
Q#2->gi 2586087	Superfamily	LRRNT_2 superfamily		
	Multidom	PLN00113		
	Multidom	PLN03150		
Q#3->gi 343466349	Specific	PKc	LRR and kinase	Yes
	Superfamily	PKc.like superfamily		
	Superfamily	LRRNT_2 superfamily		
	Superfamily	LRR_RI superfamily		
	Multidom	PLN00113		
Q#4->gi 343466347	Specific	PKc	LRR and kinase	Yes
	Superfamily	PKc.like superfamily		
	Superfamily	LRRNT_2 superfamily		
	Superfamily	LRR_RI superfamily		
	Multidom	PLN00113		
Q#5->gi 63098460	Superfamily	PKc.like superfamily	LRR and kinase	Yes
	Multidom	PLN00113		
Q#6->gi 63098462	Superfamily	PKc.like superfamily	LRR and kinase	Yes
	Multidom	PLN00113		
Q#7->gi 63098474	Superfamily	PKc.like superfamily	LRR and kinase	Yes
	Multidom	PLN00113		
Q#8->gi 63098472	Superfamily	PKc.like superfamily	LRR and kinase	Yes
	Multidom	PLN00113		
Q#9->gi 63098486	Superfamily	PKc.like superfamily	LRR and kinase	Yes
	Multidom	PLN00113		
Q#10->gi 63098454	Superfamily		LRR and kinase	Yes
	Multidom	PLN00113		

can be regarded as indirect structural evidence for resistance.

In terms of occurrence of LRR or kinase structure, all of the 10 proteins in Table 5 show consistent evidence, which shows that the genes in candidate gene data set demonstrate a good possibility of being resistant to *Xoo*.

Second, the functional information of the screened proteins is also considered by using the search engine of Gene Ontology (GO) [25], which is a popular bioinformatics

ontology aiming at standardizing the representation of gene and gene product attributes across species and databases (<http://www.geneontology.org/>). GO is also a powerful annotation tool providing a controlled vocabulary of functional terms and describing gene product characteristics. The annotation was performed with BLAST2GO [26, 27] which is based on sequence similarity. For the annotation, the configuration settings are as follows: BLASTP against NCBI nonredundant (nr) protein database, *E*-value filter  $\leq 10^{-3}$ ,

TABLE 6: Sequence distribution for biological process in GO database.

Go term	#Seq	Score	Parents	Evidence?
Cellular response to stimulus	50	30	Res, Cep	Yes
Regulation of biological process	50	18	Bir, Bip	
Response to stress	44	44	Res	Yes
Multicellular organismal development	41	72.4	Muo, Dep	
Response to biotic stimulus	40	40	Res	Yes
Primary metabolic process	38	21.4	Mep	
Response to external stimulus	37	37.8	Res	Yes
Anatomical structure development	37	31.2	Dep	
Cell death	34	34	Death, Cep	
Response to abiotic stimulus	33	33	Res	Yes
Establishment of localization	33	19.8	Loc, Bip	
Catabolic process	30	30	Mep	
Reproductive process	30	6.48	Bip, Rep	
Response to endogenous stimulus	10	10	Res	Yes
Macromolecule metabolic process	10	3.6	Mep	
Cellular metabolic process	10	3.42	Mep, Cep	
Cell cycle	5	5	Cep	
Regulation of biological quality	4	0.88	Bir	
Biosynthetic process	3	3	Mep	
Cell communication	3	3	Cep	
Nitrogen compound metabolic process	3	1.08	Mep	
Cellular homeostasis	1	1	Hop, Cep	

HSP length cutoff of 33, maximum 20 BLAST hits per sequence to sequence description tool, and annotation cutoff of 55. The sequence distribution results for biological process in GO are listed in Table 6.

As can be seen from Table 6, 50 out of 74 gene sequences are connected with GO terms related to cellular response to stimulus, and the hitting ratio is 67.57%. As cellular response to stimulus is a clear clue for resistant gene, the recall ratio is considerable. Observing entries in the first column of Table 6, which reflect gene function information, there are other five entries relevant to gene resistance, that is, “response to stress”, “response to biotic stimulus”, “response to external stimulus”, “response to abiotic stimulus”, and “response to endogenous stimulus”. Among them, 44 genes are hit for response to stress, 40 for response to biotic stimulus, 37 for response to external stimulus, 33 for response to abiotic stimulus, and 10 for response to endogenous stimulus. These results strongly support the hypothesis that proteins ranked in top list show evidence of resistance response. Since the final validation should be verified by the traditional laboratory experiment, the intensively selected candidate data set holds great potentials worthy of empirical testing and verification.

### 5. Conclusion

In this research, a hybrid strategy of gene prioritization is proposed, and reasonable results have been obtained. The flowchart of this strategy is shown in Figure 1. The protein

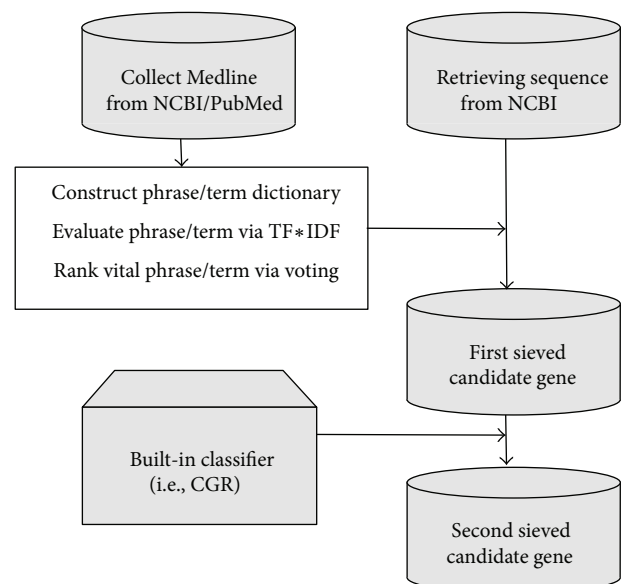


FIGURE 1: Flowchart of the Hybrid Strategy.

sequences and literature texts are both automatically collected from NCBI database, and our scheme consists of two sieves, the text-mining sieve and the classifier sieve. The first sieve is to screen candidate gene according to the important phrase evaluation through  $TF * IDF$  and voting scheme. After this

step, only those protein sequences with vital annotation are retained in the candidate set. Furthermore, the second sieve is a built-in classifier based on chaos games representation, and sequences predicted to be positive in this step show sufficient sequence similarity with 13 known *Xoo*-resistant proteins. The two sieves represent two popularly used but totally different methods for gene prioritization. After both sieving steps, the remaining sequence corresponds to those highly possible candidate genes. Thus a hybrid strategy for gene prioritization is proposed.

The effectiveness of this hybrid strategy stems from the successful combination of both a sequence-based classifier and text-mining based candidate screening. Generally, for a mere sequence-based predictor, the fraction of retrieved genes relevant to resistance is small, which leads to a low precision value and a high false positive rate. Meanwhile, for a mere text-mining based candidate screening method, the fraction of retrieved genes relevant to resistance is also low, which means a low recall rate. By balancing the high false positive rate and low recall rate, the hybrid strategy proposed in our work achieves a considerably accurate gene screening. The validation test of the candidate dataset shows that our proposed strategy is a significant attempt in large-scale gene prioritization.

The success of the hybrid strategy also benefits from the abundant information about the targeted gene. On the one hand, the disease resistant gene is quite a popular research model and there has been an increasingly large number of text and sequence resources about R gene. On the other hand, the disease gene resistance possesses many bio-specific properties which make it clear and convenient to locate resistance through texts by using key phrase matching during text mining.

More significantly, the strategy proposed in this paper is domain free, which means that it shows good potentials for use in other cases for different functional gene prediction. Currently, besides disease resistant gene, like *Xoo* resistant gene, more and more resistant genes are being investigated for better functional annotation or gene discovery, including cold resistant, drought resistant, and herbicide resistant genes. Therefore, the proposed hybrid methods are expected to be highly successful in achieving enhanced gene prioritization.

## Acknowledgments

Research described in this paper was supported in part by Grant received from the General Research Fund of the University Grant Council of Hong Kong (GRF Project no. 142711), City University of Hong Kong (Project nos. 7004091, 9610283, 7002793, 9610226, 9041694, and 9610188), the Fundamental Research Funds for the Central Universities of China (Project nos. 2013PY120, 52902-0900202346), and the National Natural Science Foundation of China (Grant no. 61202305). The authors would also like to acknowledge supports received from the Dialogue System Group, Department of Chinese, Translation and Linguistics, and the Halliday Center for Intelligent Applications of Language Studies, City University of Hong Kong.

## References

- [1] Q. Zhang, J. Li, Y. Xue, B. Han, and X. W. Deng, "Rice 2020: a call for an international coordinated effort in rice functional genomics," *Molecular Plant*, vol. 1, no. 5, pp. 715–719, 2008.
- [2] Z. Chu, M. Yuan, J. Yao et al., "Promoter mutations of an essential gene for pollen development result in disease resistance in rice," *Genes and Development*, vol. 20, no. 10, pp. 1250–1255, 2006.
- [3] K. Gu, B. Yang, D. Tian et al., "R gene expression induced by a type-III effector triggers disease resistance in rice," *Nature*, vol. 435, no. 7045, pp. 1122–1125, 2005.
- [4] A. S. Iyer and S. R. McCouch, "The rice bacterial blight resistance gene *xa5* encodes a novel form of disease resistance," *Molecular Plant-Microbe Interactions*, vol. 17, no. 12, pp. 1348–1354, 2004.
- [5] G. Ponciano, M. Yoshikawa, J. L. Lee, P. C. Ronald, and M. C. Whalen, "Pathogenesis-related gene expression in rice is correlated with developmentally controlled *Xa21*-mediated resistance against *Xanthomonas oryzae* pv. *oryzae*," *Physiological and Molecular Plant Pathology*, vol. 69, no. 4–6, pp. 131–139, 2006.
- [6] W.-Y. Song, G.-L. Wang, L.-L. Chen et al., "A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*," *Science*, vol. 270, no. 5243, pp. 1804–1806, 1995.
- [7] X. Jingbo, Z. Silan, S. Feng et al., "Using the concept of pseudo amino acid composition to predict resistance gene against *Xanthomonas oryzae* pv. *oryzae* in rice: an approach from chaos games representation," *Journal of Theoretical Biology*, vol. 284, no. 1, pp. 16–23, 2011.
- [8] S. A. Naveed, M. Babar, A. Arif et al., "Detection of bacterial blight resistant gene *xa5* using linked marker approaches," *African Journal of Biotechnology*, vol. 9, no. 24, pp. 3549–3554, 2010.
- [9] A. Marchler-Bauer, S. Lu, J. B. Anderson et al., "CDD: a Conserved Domain Database for the functional annotation of proteins," *Nucleic Acids Research*, vol. 39, no. 1, pp. D225–D229, 2011.
- [10] C. R. Arias, H. Y. Yeh, and V.W. Soo, *Disease Gene Prioritization, Bioinformatics*, INTECH, Rijeka, Croatia, 2011.
- [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] R. M. Piro, I. Molineris, U. Ala, P. Provero, and F. di Cunto, "Candidate gene prioritization based on spatially mapped gene expression: an application to XLMR," *Bioinformatics*, vol. 26, Proceedings of the 9th European Conference on Computational Biology, Ghent, Belgium, 2010, no. 18, pp. i618–i624, 2010.
- [13] 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.
- [14] 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.
- [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, article 55, 2005.
- [16] A. Schlicker, T. Lengauer, and M. Albrecht, "Improving disease gene prioritization using the semantic similarity of gene

- ontology terms,” *Bioinformatics*, vol. 26, Proceedings of the 9th European Conference on Computational Biology, Ghent, Belgium, 2010, no. 18, pp. i561–i567, 2010.
- [17] J. Chen, B. J. Aronow, and A. G. Jegga, “Disease candidate gene identification and prioritization using protein interaction networks,” *BMC Bioinformatics*, vol. 10, article 73, 2009.
- [18] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert, “GeneRank: using search engine technology for the analysis of microarray experiments,” *BMC Bioinformatics*, vol. 6, article 233, 2005.
- [19] J. E. Hutz, A. T. Kraja, H. L. McLeod, and M. A. Province, “CANDID: a flexible method for prioritizing candidate genes for complex human traits,” *Genetic Epidemiology*, vol. 32, no. 8, pp. 779–790, 2008.
- [20] 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.
- [21] J. Freudenberg and P. Propping, “A similarity-based method for genome-wide prediction of disease-relevant human genes,” *Bioinformatics*, vol. 18, no. 2, pp. S110–S115, 2002.
- [22] 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.
- [23] 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.
- [24] L. Miozzi, R. M. Piro, F. Rosa et al., “Functional annotation and identification of candidate disease genes by computational analysis of normal tissue gene expression data,” *PLoS ONE*, vol. 3, no. 6, Article ID e2439, 2008.
- [25] M. A. Harris, J. Clark, A. Ireland et al., “The Gene Oncology (GO) database and informatics resource,” *Nucleic Acids Research*, vol. 32, pp. D258–D261, 2004.
- [26] S. Götz, J. M. García-Gómez, J. Terol et al., “High-throughput functional annotation and data mining with the Blast2GO suite,” *Nucleic Acids Research*, vol. 36, no. 10, pp. 3420–3435, 2008.
- [27] A. Labarga, F. Valentin, M. Anderson, and R. Lopez, “Web services at the European bioinformatics institute,” *Nucleic Acids Research*, vol. 35, pp. W6–W11, 2007.

## Review Article

# Enabling Large-Scale Biomedical Analysis in the Cloud

Ying-Chih Lin,<sup>1,2</sup> Chin-Sheng Yu,<sup>1,3</sup> and Yen-Jen Lin<sup>4</sup>

<sup>1</sup> Master's Program in Biomedical Informatics and Biomedical Engineering, Feng Chia University, No. 100 Wenhwa Road, Seatwen, Taichung 40724, Taiwan

<sup>2</sup> Department of Applied Mathematics, Feng Chia University, No. 100 Wenhwa Road, Seatwen, Taichung 40724, Taiwan

<sup>3</sup> Department of Information Engineering and Computer Science, Feng Chia University, No. 100 Wenhwa Road, Seatwen, Taichung 40724, Taiwan

<sup>4</sup> Department of Computer Science, National Tsing Hua University, No. 101, Section 2, Kuang-Fu Road, Hsinchu 30013, Taiwan

Correspondence should be addressed to Ying-Chih Lin; [linian.tw@gmail.com](mailto:linian.tw@gmail.com)

Received 6 August 2013; Accepted 22 September 2013

Academic Editor: Chun-Yuan Lin

Copyright © 2013 Ying-Chih Lin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Recent progress in high-throughput instrumentations has led to an astonishing growth in both volume and complexity of biomedical data collected from various sources. The planet-size data brings serious challenges to the storage and computing technologies. Cloud computing is an alternative to crack the nut because it gives concurrent consideration to enable storage and high-performance computing on large-scale data. This work briefly introduces the data intensive computing system and summarizes existing cloud-based resources in bioinformatics. These developments and applications would facilitate biomedical research to make the vast amount of diversification data meaningful and usable.

## 1. Introduction

In more and more cases, the ability to gain experimental data has far surpassed the capability in doing further analyses. DNA sequencing presents a particularly good example of this trend. By current next-generation sequencing (NGS) technologies, an individual laboratory can generate terabase-scales of DNA and RNA sequencing data within a day at a reasonable cost [1–3]. However, the computing technologies required to maintain, process, and integrate the massive datasets are beyond the reach of small laboratories and introduce serious challenges even for larger institutes. Success at all fields will heavily rely on the ability to explain these large-scale and great diversification datasets, which drives us to adopt advances in computing methods.

The coming age of sharp data growth and increasing data diversification is a major challenge for biomedical research in the postgenome era. Cloud computing is an alternative to crack the nut because it gives concurrent consideration to enable storage and massive computing on large-scale data [4–6]. More than this cloud platform can considerably save costs in server hardware, administration, and maintenance by the virtualization technology, which allows systems to

act like real computers with flexible specification of the number of processors, memory, and disk size, operating system, and so on. With flexible cloud architectures that can harness petabyte scales of data, Internet-based companies, such as Google and Amazon, offer on-demand services to tens of thousands of users simultaneously. In addition, cloud storages allow large-scale and potentially shared datasets to be stored on the same infrastructure where further analyses can be run [7]. A good example is the data from the 1000 Genomes Project, which has grown to 200 terabytes of genomic data including DNA sequenced from more than 1,700 individuals, and it is now available on the Amazon cloud [8]. Developing translational biomedical applications with cloud technologies will enable significant breakthroughs in the diagnosis, prognosis, and high-quality healthcare. This study introduces the data-intensive computing system and summarizes existing cloud-based resources in bioinformatics. These developments and applications would facilitate biomedical research to make the massive datasets meaningful and usable.

This paper is organized as follows. Section 2 introduces the state of the art in the cloud developments of translational biomedical science. Subsequently, we review the

framework and platforms for massive computing in the cloud in Section 3. Finally, Section 4 draws our conclusion.

## 2. Translational Biomedical Science in the Cloud

Over the last decades, biomedical informatics has contributed a vast amount of data. In the genomic side, the data deluge comes from genotyping, gene expression, NGS data, and so on. The sequence read archive (SRA) provides the scientific community with an archival destination for raw sequence data, whose volume has reached 1.6 petabytes in 2013 [9]. A key goal of 1000 Genomes Project is to investigate the genetic contribution to human disease by characterizing the geographic and functional spectrum of genetic variation on a great deal of sequencing data [10]. More genome-wide association studies (GWAS) continue to identify common genetic factors that influence health or cause disease [11–13]. On the other hand, the diagnosis side constantly generates data from pharmacy prescription data, electronic medical and insurance records, healthcare information, and so forth. Electronic health record (EHR) is a digital data for the traditional document-based patient chart and has been essential to manage the wealth of existing clinical information. US health care data alone reached 150 exabytes ( $=10^9$  gigabytes) in 2011, while at this rate its volume would be zettabyte ( $=10^{12}$  gigabytes) scale soon [14]. In many respects, the two sides of biomedical data growth have yet to converge; however, the biomedical infrastructure for big data analysis lags behind the applications. The healthcare system has no capacity yet to distill the implicit meaning of the planet-size data for timely medical decision making. Despite the strong challenge of big data, there are considerable works in the bioinformatics community to develop feasible solutions. In what follows, existing cloud-based resources and GPU computing are summarized to the two types of biomedical data.

**2.1. Genomic-Driven Data.** Today new technologies in genomics/proteomics generate biomedical data with an explosive rate. With data volume getting larger more quickly than traditional storage and computation can afford, it is the time for biomedical studies to migrate these challenges to the cloud. Cloud computing offers new computational paradigms to not only deal with data and analyses at scale but also reduce the building and operation costs. By cloud technologies, numerous works have reported successful applications in bioinformatics (Table 1). These recent developments and applications would facilitate biomedical studies to harness the planet-size data.

Cloud-based tools in Table 1 combine distributed computing and large-scale storage to come with an effective solution in terms of data transfer, storage, computation, and analysis of big biomedical data. By deploying applications with these tools, small laboratories could maintain and process the large-scale datasets within affordable costs, which is increasingly thorny even for large institutes. For example, BioVLab infrastructure [28, 36] built on the cloud is developed for

genome analysis by utilizing the *virtual collaborative lab*, a suite of tools that allow scientists to orchestrate a sequence of data analysis tasks using remote computing resources and data storage facilities on demand from local devices. Furthermore, the Crossbow [21] genotyping program applies the MapReduce workflow on Hadoop to launch many copies of the short-read aligner Bowtie [20] in parallel. Once the aligned reads are generated, Hadoop automatically starts the MapReduce workflow of consensus calling to sort and aggregate the alignments. In the benchmark set on the Amazon EC2 cloud, Crossbow genotyped a human sample comprising 2.7 billion reads in less than 3 hours using a 320-CPU cluster for a total cost of \$85 [21].

**2.2. Diagnosis-Driven Data.** More and more requirements to the healthcare quality raise difficulties in processing both the heavy and heterogeneous biomedical data. For example, the high-resolution and dynamic data of medicinal images imply that the data transfer and image analysis are extremely time-consuming. Several works leverage the cloud approach to tackle the difficulties. MapReduce, the parallel computing framework in cloud, has been used to develop an ultrafast and scalable image reconstruction method for 4D cone-beam CT [37]. A solution to power the cloud infrastructure for digital imaging communication in medicine (DICOM) is introduced as a robust cloud-based service [38]. Whereas cloud-based medical image exchange is increasingly prevalent in medicine, its security and privacy issues to the data storage and communication need to be improved [39, 40].

An alternative to attack compute-intensive problems relies on the graphics processing unit (GPU), where there are two dominant APIs for GPU computing: CUDA and OpenCL [41]. GPU architectures feature several multiprocessors with each number of stream processors. The kernel is a function on GPU, while it splits works into blocks and threads. Blocks are assigned to run on multiprocessors, each of which is composed of a user-defined number of threads. The number of threads in a block can be different to the number of stream processors inside a multiprocessor because they run in groups of constant threads called warps. Stream processors are similar to CPU cores, but they share a single fetch-decode unit within the same multiprocessor, which forces threads to execute in lockstep. The mechanism likes the traditional single instruction multiple data (SIMD) instruction; however, any thread can diverge from the common execution path so as to increase the flexibility. Two review papers present the works on GPU accelerated medical image processing and cover algorithms that are specific to individual modalities [42, 43]. Intel quite recently unveiled its new Xeon Phi coprocessor as their many integrated core (MIC) product, while the China Tianhe-2 with the coprocessor inside was announced by TOP500 as the world's fastest supercomputer in 2013 [44]. The new coprocessor has a dramatic impact on the high-performance computing field and will drive more bioinformatics applications [45].

As to the clinical informatics, a major challenge is to integrate a wide range of heterogeneous data into a single and space-saving database for further queries and analyses. EHR could be an ideal solution because it is the patient-centered



TABLE 1: Cloud-based bioinformatics tools.

Program	Description	URL	Reference
Sequence alignment			
Cloud-Coffee	Multiple sequence alignment	<a href="http://www.tcoffee.org/">http://www.tcoffee.org/</a>	[15]
USM	MapReduce solution to sequence comparison	<a href="http://usm.github.io/">http://usm.github.io/</a>	[16]
Sequence mapping and assembly			
CloudBurst	Reference-based read mapping	<a href="http://cloudburst-bio.sourceforge.net/">http://cloudburst-bio.sourceforge.net/</a>	[17]
CloudAligner	Short read mapping	<a href="http://cloudaligner.sourceforge.net/">http://cloudaligner.sourceforge.net/</a>	[18]
SEAL	Short read mapping and duplicate removal	<a href="http://biidoop-seal.sourceforge.net/">http://biidoop-seal.sourceforge.net/</a>	[19]
Crossbow	Combine sequence aligner Bowtie and the SNP caller SOAPsnp [20]	<a href="http://bowtie-bio.sourceforge.net/crossbow/">http://bowtie-bio.sourceforge.net/crossbow/</a>	[21]
Contrail	<i>De novo</i> assembly	<a href="http://contrail-bio.sourceforge.net/">http://contrail-bio.sourceforge.net/</a>	[22]
Eoulsan	Sequencing data analysis	<a href="http://transcriptome.ens.fr/eoulsan/">http://transcriptome.ens.fr/eoulsan/</a>	[23]
Quake	Quality-aware detection and correction of sequencing errors	<a href="http://www.cbcb.umd.edu/software/quake/">http://www.cbcb.umd.edu/software/quake/</a>	[24]
Gene expression			
Myrna	Differential expression analysis for RNA-seq	<a href="http://bowtie-bio.sourceforge.net/myrna/">http://bowtie-bio.sourceforge.net/myrna/</a>	[25]
FX	RNA-seq analysis tool	<a href="http://fx.gmi.ac.kr/">http://fx.gmi.ac.kr/</a>	[26]
ArrayExpressHTS	RNA-seq process and quality assessment	<a href="http://www.ebi.ac.uk/services">http://www.ebi.ac.uk/services</a>	[27]
Comprehensive application			
BioVLab	A virtual collaborative lab for biomedical applications	<a href="https://sites.google.com/site/biovlab/">https://sites.google.com/site/biovlab/</a>	[28]
Hadoop-BAM	Directly manipulate NGS data	<a href="http://sourceforge.net/projects/hadoop-bam/">http://sourceforge.net/projects/hadoop-bam/</a>	[29]
SeqWare	A scalable NoSQL database for NGS data	<a href="http://seqware.sourceforge.net">http://seqware.sourceforge.net</a>	[30]
PeakRanger	Peak caller for ChIP-seq data	<a href="http://ranger.sourceforge.net/">http://ranger.sourceforge.net/</a>	[31]
YunBe	Gene set analysis for biomarker identification	<a href="http://tinyurl.com/yunbedownload/">http://tinyurl.com/yunbedownload/</a>	[32]
GATK	Genome analysis toolkit	<a href="http://www.broadinstitute.org/gatk/">http://www.broadinstitute.org/gatk/</a>	[33]
Cloud BioLinux	A virtual machine with over 135 bioinformatics packages	<a href="http://cloudbiolinux.org/">http://cloudbiolinux.org/</a>	[34]
CloVR	A virtual machine for automated sequence analysis	<a href="http://clovr.org/">http://clovr.org/</a>	[35]

record by integrating and managing personal medical information from various sources. EHRs are built to share information with other healthcare providers and organizations, while the cloud technologies can facilitate EHR integration and sharing. Developing EHR services on the cloud can not only reduce the building and operation costs but also support the interoperability and flexibility [46]. There are a great number of works that contributed different cloud-supported frameworks to improve EHR services. For instance, an e-health cloud system is defined to be capable of adapting itself to different diseases and growing numbers of patients, that is, improving the scalability [47]. Khansa et al. proposed an intelligent cloud-based EHR system and claimed that it has the potential to reduce medical errors and improve patients' quality of life [48]. A recent work introduces the state of cloud computing in healthcare [49]. Moreover, there are a number of security issues/concerns associated with cloud

computing, which is one of the major obstacles for the commercial considerations. As the emerging cloud technology to the healthcare system, more recent studies investigate the security and privacy issues [50–53].

### 3. Massive Computing in the Cloud

Cloud computing started with the promise of inexhaustible resources so that the data-intensive computing can be easily deployed. The three service models of cloud computing, that is, Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS), drive more complex and sophisticated markets. What makes cloud computing different from traditional IT technologies are mainly service delivery and consumer utilization models. Cloud platform is rapidly growing as a new paradigm for provisioning both storage and computing as a utility [54].

Based on the platforms, the IT capability is raised so that services can be easily deployed in a pay-as-you-go model. Subsequently, lots of resources could be acquired with a relatively low cost to test novel ideas or conduct extensive simulations. One could access more computing resources in lab to carry out his innovation based on a self-service and self-managed environment. Also, the feature of scalability for cloud platforms allows a lab-scale tool to be extended to a cloud application or a data-intensive scalable computing (DISC) system with fewer efforts [55, 56].

**3.1. MapReduce Framework.** One cannot mention DISC without mentioning MapReduce, while even many works regard MapReduce as the de facto standard for DISC [55, 57]. In 2004, Google announced a distributed computing framework, MapReduce, as the key technology for processing large datasets on a cluster made by upwards of one thousand commodity machines [58]. The MapReduce framework facilitates the management and development of massively parallel computing applications. A MapReduce program consists of two user-specified functions: map and reduce. The map function processes a <key, value> pair to generate a set of intermediate pairs, whereas the reduce function merges all intermediate results associated with the same key. In the beginning, the programming framework is used to assist Google in speedy searches, and nowadays more than 10,000 distant programs have been conducted at Google for the large-scale data analysis [57]. Once applications are modeled to the MapReduce manner, they all enjoy the scalability and fault-tolerance inherent in its execution platform supported by Google File System (GFS), whereas the successful implementation of the MapReduce model, the open-source platform Hadoop, along with the MapReduce framework, has been extensively used outside of Google by academia and industry [59]. Moreover, Ekanayake et al. compared the performances of Hadoop MapReduce, Microsoft Dryad-LINQ, and MPI implementation on two bioinformatics applications and suggested that the flexibility of MapReduce will become the preferred approach [60]. Recently, more and more MapReduce applications are proposed for bioinformatics studies [16–18, 33, 37, 61].

**3.2. Cloud Platform.** PaaS provides a substantial boost with the manageable cost, and there have been a number of solutions, such as Google App Engine (GAE), Amazon Elastic Compute Cloud (EC2), and Windows Azure. GAE offers a robust and extensible runtime environment for developing and hosting web-based applications in Google-managed infrastructure, rather than providing direct access to a customized virtual machine. Malawski et al. investigated how to use GAE service for free of charge execution of compute-intensive problems [62], while Prodan et al. compared GAE and Amazon EC2 in performance and resource consumption by four basic algorithms [63]. EC2 is a cloud service whereby one can rent virtual machines from Amazon data center and deploy scalable applications on them. Several works are conducted to evaluate EC2 performance [64]. Wall et al. concluded that the effort to transform existing comparative genomics algorithms from local infrastructures to cloud is

not trivial, but the cloud environment is an economical alternative in the speed and flexibility considerations [65]. Further, two works explore the biomedical cloud built on Amazon service with several case studies [66, 67].

Windows Azure platform provides a series of services for developing and deploying Windows-based applications on the cloud, and it makes use of Microsoft infrastructure to host services and scale them seamlessly [68–70]. Moreover, Aneka provides a flexible model for developing distributed applications, which can be integrated with external cloud platforms further. Aneka presents the possibility to avoid vendor lockin through a virtual infrastructure, a private datacentre, or a server, so that one could freely scale to cloud platforms when required. Its deadline-driven provisioning mechanism also supports QoS-aware execution of scientific applications in hybrid clouds [71]. It is handy to leverage famous PaaS platforms for compute-intensive applications; however, commercial cloud services charge for CPU time, storage space, bandwidth usage, and advanced functions. Apart from the service charge, the commercial cloud platform is still difficult for data-intensive applications. The critical factor is that current network infrastructure is too slow to enable terabytes of data to be routinely transferred. A feasible solution for transferring planet-size data is to copy the data into a big storage drive and then send the drive to the destination. In addition, the private cloud solution helps developers to construct cloud platforms for local use [72].

## 4. Conclusions

Recent technologies on next-generation sequencing and high-throughput experiments cause an exponential growth of biomedical data, and subsequently serious challenges arise in processing data volume and complexity. Numerous works have reported successful bioinformatics applications to harness the big data. Developing cloud-based biomedical applications can integrate the vast amount of diversification data in one place and analyze them on a continuous basis. This would make a significant breakthrough to launch a high-quality healthcare. This work briefly introduces the data-intensive computing systems and summarizes existing cloud-based resources in bioinformatics. These developments and applications would facilitate biomedical applications to make the planet-size data meaningful and usable.

## 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 Science Council under contract number NSC-102-2218-E-035-004.

## References

- [1] F. Luciani, R. A. Bull, and A. R. Lloyd, "Next generation deep sequencing and vaccine design: today and tomorrow," *Trends in Biotechnology*, vol. 30, no. 9, pp. 443–452, 2012.

- [2] L. Liu, Y. Li, S. Li et al., "Comparison of next-generation sequencing systems," *Journal of Biomedicine and Biotechnology*, vol. 2012, Article ID 251364, 11 pages, 2012.
- [3] L. D. Stein, "The case for cloud computing in genome informatics," *Genome Biology*, vol. 11, no. 5, article 207, 2010.
- [4] E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan, "Computational solutions to large-scale data management and analysis," *Nature Reviews Genetics*, vol. 11, no. 9, pp. 647–657, 2010.
- [5] A. Rosenthal, P. Mork, M. H. Li, J. Stanford, D. Koester, and P. Reynolds, "Cloud computing: a new business paradigm for biomedical information sharing," *Journal of Biomedical Informatics*, vol. 43, no. 2, pp. 342–353, 2010.
- [6] J. Chen, F. Qian, W. Yan, and B. Shen, "Translational biomedical informatics in the cloud: present and future," *BioMed Research International*, vol. 2013, Article ID 658925, 8 pages, 2013.
- [7] S. Sakr, A. Liu, D. M. Batista, and M. Alomari, "A survey of large scale data management approaches in cloud environments," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 3, pp. 311–336, 2011.
- [8] 1000 Genomes Project and AWS, <http://aws.amazon.com/1000genomes/>.
- [9] M. Shumway, G. Cochrane, and H. Sugawara, "Archiving next generation sequencing data," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D870–D871, 2009.
- [10] 1000 Genomes Project Consortium, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, pp. 56–65, 2012.
- [11] E. Evangelou and J. P. A. Ioannidis, "Meta-analysis methods for genome-wide association studies and beyond," *Nature Reviews Genetics*, vol. 14, pp. 379–389, 2013.
- [12] S. J. Chapman and A. V. S. Hill, "Human genetic susceptibility to infectious disease," *Nature Reviews Genetics*, vol. 13, no. 3, pp. 175–188, 2012.
- [13] G. Gibson, "Rare and common variants: twenty arguments," *Nature Reviews Genetics*, vol. 13, no. 2, pp. 135–145, 2012.
- [14] W. Hoover, *Transforming Health Care Through Big Data*, Institute for Health Technology Transformation, 2013.
- [15] P. di Tommaso, M. Orobitg, F. Guirado, F. Cores, T. Espinosa, and C. Notredame, "Cloud-Coffee: implementation of a parallel consistency-based multiple alignment algorithm in the T-coffee package and its benchmarking on the Amazon Elastic-Cloud," *Bioinformatics*, vol. 26, no. 15, pp. 1903–1904, 2010.
- [16] J. S. Almeida, A. Gruneberg, W. Maass, and S. Vinga, "Fractal MapReduce decomposition of sequence alignment," *Algorithms for Molecular Biology*, vol. 7, article 12, 2012.
- [17] M. C. Schatz, "CloudBurst: highly sensitive read mapping with MapReduce," *Bioinformatics*, vol. 25, no. 11, pp. 1363–1369, 2009.
- [18] T. Nguyen, W. Shi, and D. Ruden, "CloudAligner: a fast and full-featured MapReduce based tool for sequence mapping," *BMC Research Notes*, vol. 4, article 171, 2011.
- [19] L. Pireddu, S. Leo, and G. Zanetti, "Seal: a distributed short read mapping and duplicate removal tool," *Bioinformatics*, vol. 27, no. 15, pp. 2159–2160, 2011.
- [20] R. Li, Y. Li, X. Fang et al., "SNP detection for massively parallel whole-genome resequencing," *Genome Research*, vol. 19, no. 6, pp. 1124–1132, 2009.
- [21] B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg, "Searching for SNPs with cloud computing," *Genome Biology*, vol. 10, no. 11, article R134, 2009.
- [22] M. C. Schatz, A. L. Delcher, and S. L. Salzberg, "Assembly of large genomes using second-generation sequencing," *Genome Research*, vol. 20, no. 9, pp. 1165–1173, 2010.
- [23] L. Jourden, M. Bernard, M.-A. Dillies, and S. L. Crom, "Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses," *Bioinformatics*, vol. 28, no. 11, pp. 1542–1543, 2012.
- [24] D. R. Kelley, M. C. Schatz, and S. L. Salzberg, "Quake: quality-aware detection and correction of sequencing errors," *Genome Biology*, vol. 11, no. 11, article R116, 2010.
- [25] B. Langmead, K. D. Hansen, and J. T. Leek, "Cloud-scale RNA-sequencing differential expression analysis with Myrna," *Genome Biology*, vol. 11, article R83, 2010.
- [26] D. Hong, A. Rhie, S.-S. Park et al., "FX: an RNA-seq analysis tool on the cloud," *Bioinformatics*, vol. 28, no. 5, pp. 721–723, 2012.
- [27] A. Goncalves, A. Tikhonov, A. Brazma, and M. Kapushesky, "A pipeline for RNA-seq data processing and quality assessment," *Bioinformatics*, vol. 27, no. 6, pp. 867–869, 2011.
- [28] H. Lee, Y. Yang, H. Chae et al., "BioVLAB-MMIA: a cloud environment for microRNA and mRNA integrated analysis (MMIA) on Amazon EC2," *IEEE Transactions on Nanobioscience*, vol. 11, no. 3, pp. 266–272, 2012.
- [29] M. Niemenmaa, A. Kallio, A. Schumacher, P. Klemelä, E. Korpelainen, and K. Heljanko, "Hadoop-BAM: directly manipulating next generation sequencing data in the cloud," *Bioinformatics*, vol. 28, no. 6, pp. 876–877, 2012.
- [30] B. D. O'Connor, B. Merriman, and S. F. Nelson, "SeqWare Query Engine: storing and searching sequence data in the cloud," *BMC Bioinformatics*, vol. 11, no. 12, article S2, 2010.
- [31] X. Feng, R. Grossman, and L. Stein, "PeakRanger: a cloud-enabled peak caller for ChIP-seq data," *BMC Bioinformatics*, vol. 12, article 139, 2011.
- [32] L. Zhang, S. Gu, Y. Liu, B. Wang, and F. Azuaje, "Gene set analysis in the cloud," *Bioinformatics*, vol. 28, no. 2, pp. 294–295, 2012.
- [33] A. McKenna, M. Hanna, E. Banks et al., "The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Research*, vol. 20, no. 9, pp. 1297–1303, 2010.
- [34] K. Krampis, T. Booth, B. Chapman et al., "Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community," *BMC Bioinformatics*, vol. 13, article 42, 2012.
- [35] S. V. Angiuoli, M. Matalka, A. Gussman et al., "CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing," *BMC Bioinformatics*, vol. 12, article 356, 2011.
- [36] H. Chae, I. Jung, H. Lee et al., "Bio and health informatics meets cloud: BioVLab as an example," *Health Information Science and Systems*, vol. 1, no. 6, 9 pages, 2013.
- [37] B. Meng, G. Pratz, and L. Xing, "Ultrafast and scalable cone-beam CT reconstruction using MapReduce in a cloud computing environment," *Medical Physics*, vol. 38, no. 12, pp. 6603–6609, 2011.
- [38] G. Patel, "DICOM medical image management the challenges and solutions: cloud as a service (CaaS)," *Open Access Scientific Reports*, vol. 1, no. 4, 4 pages, 2012.
- [39] L. A. B. Silva, C. Costa, and J. L. Oliveira, "DICOM relay over the cloud," *International Journal of Computer Assisted Radiology and Surgery*, vol. 8, pp. 323–333, 2013.

- [40] S. G. Shinia, T. Thomas, and K. Chithranjana, "Cloud based medical image exchange-security challenges," *Procedia Engineering*, vol. 38, pp. 3454–3461, 2012.
- [41] J.-S. Varré, B. Schmidt, S. Janot, and M. Giraud, "Manycore high-performance computing in bioinformatics," in *Advances In Genomic Sequence Analysis and Pattern Discovery*, L. Elnitski, H. Piontkivska, and L. R. Welch, Eds., chapter 8, World Scientific, 2011.
- [42] A. Eklund, P. Dufort, D. Forsberg, and S. M. LaConte, "Medical image processing on the GPU—past, present and future," *Medical Image Analysis*, vol. 17, no. 8, pp. 1073–1094, 2013.
- [43] L. Shi, W. Liu, H. Zhang et al., "A survey of GPU-based medical image computing techniques," *Quantitative Imaging in Medicine and Surgery*, vol. 2, no. 3, pp. 188–206, 2012.
- [44] TOP500 Supercomputer Sites, <http://www.top500.org/>.
- [45] Intel, "Heterogeneous computing in the cloud: crunching big data and democratizing HPC access for the life sciences," *Intel White Paper*, 2013.
- [46] J. Haughton, "Look up: the right EHR may be in the cloud. Major advantages include interoperability and flexibility," *Health Management Technology*, vol. 32, no. 2, p. 52, 2011.
- [47] J. Vilaplana, F. Solsona, F. Abella et al., "The cloud paradigm applied to e-Health," *BMC Medical Informatics and Decision Making*, vol. 13, article 10, 2013.
- [48] L. Khansa, J. Forcade, G. Nambari et al., "Proposing an intelligent cloud-based electronic health record system," *International Journal of Business Data Communications and Networking*, vol. 8, no. 3, pp. 57–71, 2012.
- [49] S. P. Ahuja, S. Mani, and J. Zambrano, "A survey of the state of cloud computing in healthcare," *Network and Communication Technologies*, vol. 1, no. 2, pp. 12–19, 2012.
- [50] F. Magrabi, J. Aarts, C. Nohr et al., "A comparative review of patient safety initiatives for national health information technology," *International Journal of Medical Informatics*, vol. 82, pp. e139–e148, 2013.
- [51] H. Singh, J. S. Ash, and D. F. Sittig, "Safety assurance factors for electronic health record resilience (SAFER): study protocol," *BMC Medical Informatics and Decision Making*, vol. 13, article 8, 2013.
- [52] D. F. Sittig and H. Singh, "Electronic health records and national patient-safety goals," *The New England and Journal of Medicine*, vol. 367, no. 19, pp. 1854–1860, 2012.
- [53] T. S. Chen, C. H. Liu, T. L. Chen et al., "Secure dynamic access control scheme of PHR in cloud computing," *Journal of Medical Systems*, vol. 36, no. 6, pp. 4005–4020, 2012.
- [54] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599–616, 2009.
- [55] R. E. Bryant, "Data-intensive scalable computing for scientific applications," *Computing in Science and Engineering*, vol. 13, no. 6, pp. 25–33, 2011.
- [56] A. Iosup, S. Ostermann, N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, "Performance analysis of cloud computing services for many-tasks scientific computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 6, pp. 931–945, 2011.
- [57] J. Dean and S. Ghemawat, "Map Reduce: a flexible data processing tool," *Communications of the ACM*, vol. 53, no. 1, pp. 72–77, 2010.
- [58] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.
- [59] Apache Hadoop, <http://hadoop.apache.org/>.
- [60] J. Ekanayake, T. Gunarathne, and J. Qiu, "Cloud technologies for bioinformatics applications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 6, pp. 998–1011, 2011.
- [61] M. E. Colosimo, M. W. Peterson, S. Mardis, and L. Hirschman, "Nephele: genotyping via complete composition vectors and MapReduce," *Source Code for Biology and Medicine*, vol. 6, article 13, 2011.
- [62] M. Malawski, M. Kuzniar, P. Wojcik, and M. Bubak, "How to use Google App engine for free computing," *IEEE Internet Computing*, vol. 17, no. 1, pp. 50–59, 2013.
- [63] R. Prodan, M. Sperr, and S. Ostermann, "Evaluating high-performance computing on google app engine," *IEEE Software*, vol. 29, no. 2, pp. 52–58, 2012.
- [64] J. J. Rehr, F. D. Vila, J. P. Gardner, L. Svec, and M. Prange, "Scientific computing in the cloud," *Computing in Science and Engineering*, vol. 12, no. 3, pp. 34–43, 2010.
- [65] D. P. Wall, P. Kudtarkar, V. A. Fusaro, R. Pivovarov, P. Patil, and P. J. Tonellato, "Cloud computing for comparative genomics," *BMC Bioinformatics*, vol. 11, article 259, 2010.
- [66] V. A. Fusaro, P. Patil, E. Gafni, D. P. Wall, and P. J. Tonellato, "Biomedical cloud computing with amazon web services," *PLoS Computational Biology*, vol. 7, no. 8, Article ID e1002147, 2011.
- [67] R. L. Grossman and K. P. White, "A vision for a biomedical cloud," *Journal of Internal Medicine*, vol. 271, no. 2, pp. 122–130, 2012.
- [68] Q. Xing and E. Blaisten-Barojas, "A cloud computing system in windows azure platform for data analysis of crystalline materials," *Concurrency and Computation*, vol. 25, no. 15, pp. 2157–2169, 2013.
- [69] I. Kim, J.-Y. Jung, T. F. DeLuca et al., "Cloud computing for comparative genomics with windows azure platform," *Evolutionary Bioinformatics Online*, vol. 8, pp. 527–534, 2012.
- [70] S. J. Johnston, N. S. O'Brien, H. G. Lewis et al., "Clouds in space: scientific computing using windows azure," *Journal of Cloud Computing*, vol. 2, article 2, 2013.
- [71] C. Vecchiola, R. N. Calheiros, D. Karunamoorthy, and R. Buyya, "Deadline-driven provisioning of resources for scientific applications in hybrid clouds with Aneka," *Future Generation Computer Systems*, vol. 28, no. 1, pp. 58–65, 2012.
- [72] M. Taifi, A. Khreishah, and J. Y. Shi, "Building a private HPC cloud for compute and data-intensive applications," *International Journal on Cloud Computing*, vol. 3, no. 2, 20 pages, 2013.

## Methodology Report

# A Novel Framework for the Identification and Analysis of Duplicons between Human and Chimpanzee

Trees-Juen Chuang,<sup>1</sup> Shian-Zu Wu,<sup>2</sup> and Yao-Ting Huang<sup>2</sup>

<sup>1</sup> Genomics Research Center, Academia Sinica, Taipei, Taiwan

<sup>2</sup> Department of Computer Science and Information Engineering, National Chung Cheng University, No.168 University Road Chiayi, Taiwan

Correspondence should be addressed to Yao-Ting Huang; [ythuang@cs.ccu.edu.tw](mailto:ythuang@cs.ccu.edu.tw)

Received 24 April 2013; Revised 25 June 2013; Accepted 10 July 2013

Academic Editor: Che-Lun Hung

Copyright © 2013 Trees-Juen Chuang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Human and other primate genomes consist of many segmental duplications (SDs) due to fixation of copy number variations (CNVs). Structure of these duplications within the human genome has been shown to be a complex mosaic composed of juxtaposed subunits (called duplicons). These duplicons are difficult to be uncovered from the mosaic repeat structure. In addition, the distribution and evolution of duplicons among primates are still poorly investigated. In this paper, we develop a statistical framework for discovering duplicons via integration of a Hidden Markov Model (HMM) and a permutation test. Our comparative analysis indicates that the mosaic structure of duplicons is common in CNV/SD regions of both human and chimpanzee genomes, and a subset of core duplicons shared by the majority of CNVs/SDs. Phylogenetic analyses using duplicons suggested that most CNVs/SDs share common duplication ancestry. Many human/chimpanzee duplicons flank both ends of CNVs, which may be hotspots of nonallelic homologous recombination.

## 1. Introduction

Human genome and other primate genomes consist of many repetitive sequences. Many of these are hotspots for nonallelic homologous recombination (NAHR) [1] or genomic rearrangements. Current estimates suggest that approximately 4%–6% of our human genome is composed of segmental duplication (SD) [1–3]. SD is a DNA segment  $\geq 1$  kb in size that occurs greater than once within the genome and typically shares  $\geq 90\%$  sequence identity [1, 4]. Genomic regions of SDs have been shown to be hotspots of copy number variations (CNVs), which is a DNA segment 1 kb or larger in size and presents different number of copies in the population. A number of SDs and CNVs have been known to highly associate with several complex diseases such as HIV-1 infection, glomerulonephritis, Parkinson, and Alzheimer diseases [5–8].

The completion of several sequencing projects provided abundant resources for mapping SDs in mammalian genomes. SDs are usually identified by self-comparison of

the entire genome or by coverage analysis of overcollapsed shotgun sequences [2, 9]. For example, a genome-wide map of chimpanzee SDs was built by self-comparison of chimpanzee assembly and alignment of shotgun sequences to the human genome [10]. Through comparison of clone-ordered assemblies of human and mouse, She et al. [11] found that the amount of mouse SDs is comparable to that of human SDs. Recently, with the advent of array comparative genomic hybridization (aCGH), numerous CNVs have been discovered in several mammalian populations [12–14]. For example, Redon et al. [15] identified a total of 1,447 CNVs from 270 individuals across four populations, covering 360 megabases of the human genome. Perry et al. [16, 17] characterized a map of CNVs in chimpanzees and found that human and chimpanzee CNVs occur in orthologous regions far more than expected.

A number of statistical and combinatorial methods have been developed to identify SDs/CNVs on the basis of comparative genomics, microarray, or high-throughput sequencing platforms. For instance, comparative approaches aim to

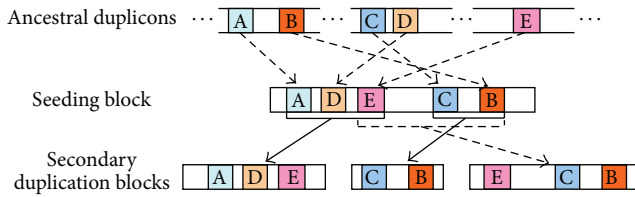


FIGURE 1: Ancestral duplicons are first aggregated into one seeding block that subsequently produces secondary duplication blocks.

uncover genomic sequences with high similarity from whole-genome sequence alignment [3, 10, 11]. Computational methods on top of microarray platforms often identify genomic regions with high density of unusual intensity signals [18, 19]. On the other hand, algorithms for high-throughput sequencing platforms search for genomic segments with ultrahigh/low read depth or aberrant mapping distances [20].

Even though many duplications have been discovered and studied in the last decade, the underlying mechanism leading to these large duplications is still not well understood. To date, NAHR and retrotransposition are two mechanisms known to support many duplication events. NAHR, also termed ectopic recombination or unequal crossover, is a recombination error during meiosis in which the exchanged chromosomes were misaligned, leading to gain or loss of DNA segments [1, 21, 22]. The misalignment of NAHR has been suspected due to repetitive elements widespread in the genome. On the other hand, the activation of retrotransposons, retrovirus, and endogenous retrovirus (ERV) may also mediate retrotransposition of a few genes via reversely transcribing RNAs into DNAs and inserting them back to the genome [23].

In recent years, a few studies started to investigate the sequence composition within large duplications and found that the structure is a complex mosaic composed of smaller subunits called *duplicons* (with a minimum length of 100 bp) [2, 24, 25]. A two-step model has been established to explain this mosaic structure [26, 27] (see Figure 1). In this model, ancestral duplicons are first transposed and aggregated into one seeding block, which subsequently produces secondary duplication blocks. Duplicons within this complex mosaic cannot be readily uncovered by conventional multiple sequence alignment approaches. Thus, Pevzner et al. [28] developed an *A*-Bruijn graph algorithm for identifying duplicons from this mosaic structure. The *A*-Bruijn graph algorithm was then revised to discover 4,692 ancestral duplicons using human SDs and outgroup mammalian genomes [24]. Subsequently, Jiang et al. [9] compiled a library of known duplication sequences and used this library to efficiently annotate SDs in a new genome.

The discovery of duplicons was based on comparing sequences of known SDs. In reality, due to the difficulty of assembling shotgun sequences in duplicated regions, large (>15 kb) and highly identical (>95%) SDs are often collapsed [11]. Furthermore, because these shotgun sequences are collected from only a few individuals in the population, SDs of unsampled individuals would be missed in the assembled genome [17]. Thus, a substantial amount of duplicons can be lost. In fact, CNVs have been viewed as a drifting and

polymorphic form of SDs, and both are probably mediated by similar mechanisms [29]. A few studies have reported that only ~24% of CNVs are overlapped with SDs [15, 22], implying that CNVs may serve as alternative repository of duplicons. Recently, analysis of a fosmid clone indicated that a large segment of CNV is deleted owing to NAHR mediated by flanking duplicons [9]. However, the distribution of duplicons within CNVs and their mosaic structures in human and other primates remains poorly investigated.

In this paper, we develop a Hidden Markov Model (HMM) for efficiently annotating duplicons within CNVs and assess the statistical significance of each duplicon. Our results indicate that the mosaic structure composed of duplicons is common in CNVs and SDs of both human and chimpanzee. Although our duplicons are annotated from a subset of CNVs, other CNV regions are found to have significantly higher density of these duplicons. Phylogenetic analyses suggest that many CNVs/SDs share common duplicons and ancestry, and these CNVs/SDs are usually centered around a few core duplicons shared by majority of duplications with common ancestry. In addition, a number of duplicons are found to flank both ends of human and chimpanzee CNVs, creating hotspots of nonallelic homologous recombination. Compared with previous functional analysis on CNVs, these duplicons are also enriched for regulation of immune process and response to stimulus but underrepresented in cell adhesion.

## 2. Method

**2.1. Data Preprocess and Problem Formulation.** We downloaded a total of 50,339 human SDs from the University of California Santa Cruz genome browser (<http://www.genome.ucsc.edu>) [2]. 1,447 human CNVs screened by a tiling array and an SNP genotyping array are obtained from Redon et al. [15]. We used Megablast [30, 31] to align all SDs against each CNV (The parameters of Megablast are set as follows:  $-e$  0.0001,  $-F$  E,  $-W$  34, and  $-M$  1000000). We found that megablast is able to complete the alignment task under this setting within one week, whereas the regular blastn is unable to finish within a reasonable period of time. Although the speed can be theoretically improved by using word size larger than 34 bp, we did not observe significant differences when further enlarging the word size. According to the alignment result, we construct an “alignment matrix” for each CNV (Figure 2). Denote  $n_k$  as the length of the  $k$ th CNV sequence and  $m$  as the number of SDs which can be aligned to the  $k$ th CNV. Let  $A_k = (a_{ij})$  be a binary  $m \times n_k$  matrix. Each element in the matrix  $A_k$  is defined as  $a_{ij} = 1$  if the  $i$ th SD is aligned to the  $j$ th position of the  $k$ th CNV and  $a_{ij} = 0$  otherwise, where  $1 \leq i \leq m$  and  $1 \leq j \leq n_k$ . Note that gaps and mismatches are excluded in  $A_k$ . Theoretically, real duplicons tend to produce segments of consecutive “1s” with higher frequency and longer length in the matrix. On the other hand, segments of 1s due to random or occasional alignments are less frequent and relatively shorter. In the following, we describe an HMM for identifying duplicon regions with sufficient frequency and length.

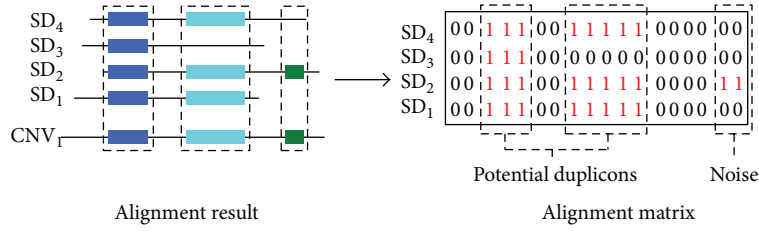


FIGURE 2: The left figure illustrates one alignment result. Fragments with the same color represent the subsequences on CNV<sub>1</sub> and SDs having high similarity. The right figure illustrates the alignment matrix corresponding to the alignment result. In this matrix, the two clusters of 1s are potential duplcons, whereas the remaining parts are probably noise.

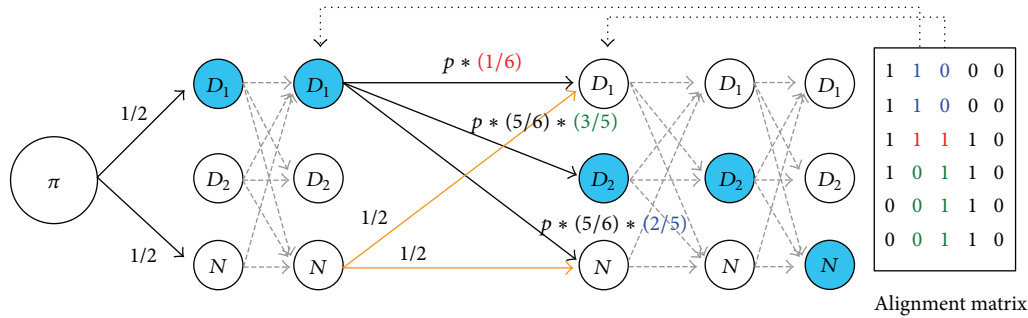


FIGURE 3: An example of state transition probability of our HMM. We take the second and third columns as an instance and highlight the transition probability for  $D_1$  state. Note that  $\omega = 1/6$  and  $\gamma = 2/5$ . The expected Viterbi path in this instance is  $D_1, D_1, D_2, D_2, N$ .

2.2. *Hidden Markov Model.* The HMM is specified by five sets of parameters,  $\lambda = (S, O, \pi, T, E)$ , where  $S$  is the set of states,  $O$  is the set of observation,  $\pi$  is the initial state,  $T$  is the set of state transition probabilities, and  $E$  is the set of emission probabilities. We define  $S = (D_1, D_2, N)$  as our state alphabet set, where  $D_1$  and  $D_2$  represent two duplcon states, and  $N$  is the nonduplcon state. We use two duplcon states in order for distinguishing adjacent duplcons. Our HMM starts at the initial state  $\pi$  with equal transition probability to one duplcon state and the nonduplcon state.

In our HMM, the state transition probabilities  $T$  are designed to approximate the length of known duplcons and reflect the transition likelihood implied by 0/1 patterns of two adjacent columns in the matrix. First, the average length of known duplcons  $L$  is computed from the duplcon library [9]. The probability of transition from one duplcon state to itself (e.g.,  $D_1$  to  $D_1$ ) is set to  $p = 1 - 1/L$ , which corresponds to a geometric distribution with mean  $L$ . In addition, we also compute the frequencies of three 0/1 patterns ( $f_{1,1}$ ,  $f_{0,1}$ , and  $f_{1,0}$ ) in two adjacent columns. For example (see Figure 3),  $f_{1,1}$ ,  $f_{0,1}$ , and  $f_{1,0}$  in the first two columns of the matrix are 3, 0, and 1, respectively. Intuitively,  $f_{1,1}$ ,  $f_{0,1}$ , and  $f_{1,0}$  imply the likelihood of transition to the same duplcon state, the other duplcon state, or nonduplcon state, respectively.

Let  $\omega = f_{1,1}/(f_{1,1} + f_{0,1} + f_{1,0})$  and  $\gamma = f_{0,1}/(f_{0,1} + f_{1,0})$ . For each duplcon state, we define three state transition probabilities: (1) transition to the same duplcon state with probability  $p\omega$ ; (2) transition to the other duplcon state with probability  $(1 - p\omega)\gamma$ ; (3) transition to nonduplcon state with probability  $(1 - p\omega)(1 - \gamma)$ . The transition probability for the

nonduplcon state is set to be equally likely. Figure 3 illustrates an example of our state transition probabilities.

Theoretically, the columns of a real duplcon should have higher frequency of 1s than those of nonduplcon columns. Thus, we define observation  $O = (o_1, o_2, \dots, o_{n_k})$  as the number of 1s in each of the  $n_k$  columns, respectively. The emission probability  $E$  of the  $i$ th duplcon state is designed to reflect the probability of observing  $o_i$  1s, assuming that this position is a real duplcon. First, we estimate the probability of observing a duplcon in one SD from the known duplcon library [9]. That is,  $P_o = C/M$ , where  $C$  is the average copy number of one duplcon and  $M$  is the number of total SDs in the duplcon library. Let  $k$  be the number of 1s in the column and  $n$  the number of SDs in the alignment matrix. The emission probability on the duplcon state is defined as  $P_d = \sum_{i=0}^k \binom{n}{i} P_o^i (1 - P_o)^{n-i}$ , corresponding to a cumulative binomial distribution. And the emission probability on nonduplcon state is defined as  $1 - P_d$ .

The maximum probability path in the HMM starting from  $\pi$  and ending at state  $S_{o_{n_k}} [x]$  is given by

$$P(V | A_k, \lambda) = P(S[x] | \pi) \times P(S_{o_1} [x]) \prod_{i=2}^{n_k} P(S_{o_i} [x] | S_{o_{i-1}} [x]) \times P(S_{o_{n_k}} [x]). \quad (1)$$

This maximum probability path is found by the Viterbi algorithm [32], and all positions are assigned to one of the

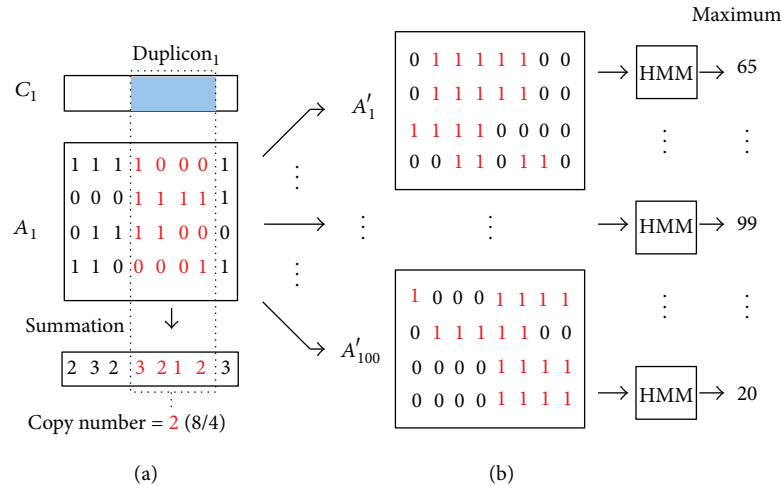


FIGURE 4: (a) An example of computing copy number for a duplison. The average copy number of duplison<sub>1</sub> is  $8/4 = 2$ . (b) Flow of permutation test. Each consecutive 1s in  $A_1$  is randomly relocated to create 100 artificial alignment matrices  $A'_1, A'_2, \dots, A'_{100}$ .

three states. We identify segments with at least 100  $D_1$  or  $D_2$  duplison states as potential duplisons.

**2.3. Permutation Test.** The statistical significance of each potential duplison is assessed by a permutation test. We define “copy number” of a duplison as the average number of SDs aligned to each position of the duplison (Figure 4(a)). The permutation test computes the probability of observing the copy number of a potential duplison from permuted data. Real duplisons tend to have sufficient number of copies, which are less likely to be observed by chance only. In the permutation test, each segment of consecutive 1s in the alignment matrix is randomly relocated to create an artificial matrix (Figure 4(b)). 100 artificial matrices are created separately for each alignment matrix. Then, duplisons of each artificial matrix are identified by applying our HMM. The maximum copy number among all duplisons in each artificial matrix is recorded. For each potential duplison of the original matrix, the  $P$  value is defined as the fraction of artificial matrices for which maximum copy number is larger than that of the potential duplison. Only those duplisons with  $P$  value  $< 0.01$  are retained as our final solution.

For instance, suppose we have 30 copies of a potential duplison observed in alignment matrix  $A_1$ . After permutation test, there are ten maximum copy numbers (from artificial simulations) greater than 30 ( $P$  value =  $0.1 > 0.01$ ). This potential duplison would be eliminated due to its nonsignificant  $P$  value. On the contrary, if there is no maximum copy number of artificial duplisons in  $A_1$  greater than 30, the duplison ( $P$  value =  $0 < 0.01$ ) is assessed as a potential true duplison.

**2.4. Gene Ontology Analysis.** We retrieve known genes annotated by Ensembl (<http://www.ensembl.org>). Duplisons overlapped with these known genes are included in our

analysis. In order to investigate the functional bias of these duplisons, we identified over- and underrepresented functions defined by gene ontology (GO) term analysis (<http://www.geneontology.org>). For each GO subcategory (level 2 and level 3) of biological process, cellular component, and molecular function, we compute the numbers of all genes and all duplisons that fall into each subcategory. The statistical significance of over- or underrepresentation in any GO subcategory is computed by chi-square test.  $P$  values are corrected using Bonferroni correction for multiple testing. The subcategories with  $P < 0.05$  are investigated in our analysis.

**2.5. Hierarchical Clustering and Phylogenetic Analysis of Duplisons.** A binary “phylogenetic profile” was constructed based on the extent of shared duplisons for each duplication segment composed of ten or more duplisons. The duplication segment is defined as the chimpanzee SDs and CNVs (chimpanzee specific, human specific, and human/chimpanzee shared) in which the segments are aligned by our duplisons with sequence identities  $\geq 95\%$  and length  $\geq 100$  bp. If a duplison is present within a duplication segment, we assigned “1” for that duplison in duplication segment, otherwise assigned “0,” generating a binary phylogenetic profile for each duplication segment. If there is no shared duplison among two duplication segments, these two segments are considered to have no related evolutionary history. A duplication group is a cluster of duplication segments grouped based on the amount of shared duplisons. Complex duplication segments were then clustered into several duplication groups by hierarchical clustering on the basis of the similarity of their phylogenetic profiles. ClustalW is used to generate phylogenetic clusters of these profiles (<http://www.ebi.ac.uk/Tools/clustalw2/index.html>). Each clade in the phylogenetic tree stands for a duplication group in our analysis.



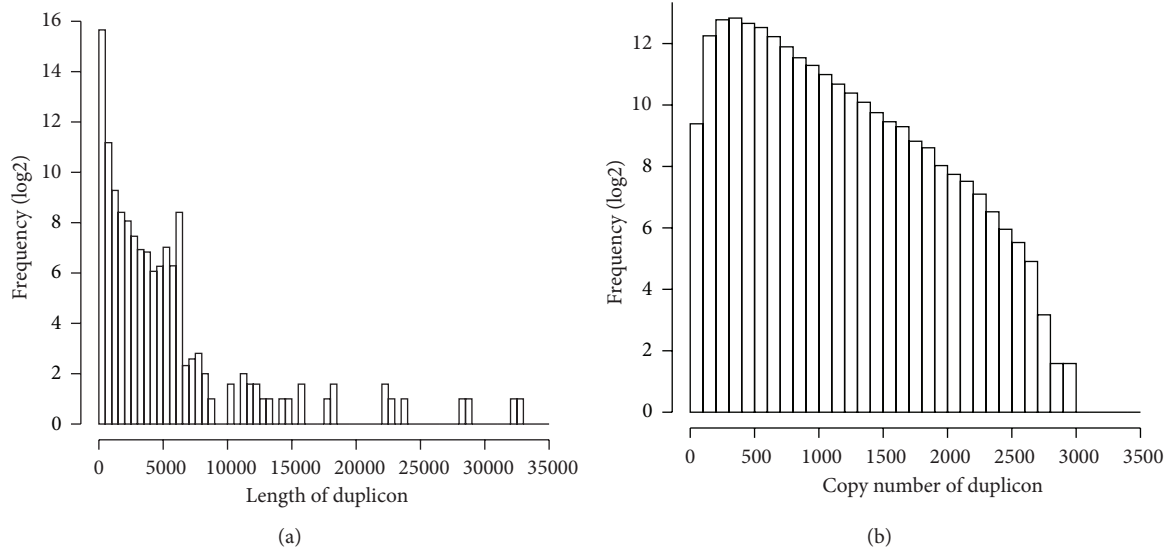


FIGURE 5: (a) The distribution of lengths of our duplicons. (b) The distribution of copy numbers of our duplicons.

### 3. Results and Discussion

**3.1. Novel Duplicons Annotated by Our Pipeline.** The binary and source code of the entire pipeline have been encapsulated via bash script and are available at <http://www.cs.ccu.edu.tw/~ythuang/Tool/HMMDupFinder/>. We downloaded a total of 50,339 human SDs from the University of California Santa Cruz genome browser (<http://www.genome.ucsc.edu>) [2]. 1,447 human CNVs screened by a tiling array and an SNP genotyping array are obtained from Redon et al. [15]. We used Megablast [30, 31] to align all SDs against each CNV and created 1,447 alignment matrices (see Section 2). We design and implement a HMM and run the HMM on alignment matrices for annotating duplicons. A total of 102,405 initial duplicons were found by the HMM. After filtration by a permutation test ( $P < 0.01$ ) and removal of identical duplicons, 56,377 unique duplicons were retained. These duplicons are spread among 1,095 CNVs. On average, each CNV contains approximately 54 unique duplicons. There are 963 CNVs (88%) having two or more identical duplicons within the genomic region, and 2,994 duplicons appear twice or more in the same CNV. ~71% of our duplicons are novel compared with known duplicons in [9]. Table 1 lists numbers of duplicons on each chromosome. Figure 5 illustrates the distribution of length and copy number of all duplicons. The average length of our duplicons is 425 bp, which is shorter than that of duplicons annotated by A-Bruijn graph method (~4,651 bp) [9, 24]. This is because A-Bruijn graph methods chain duplicons in proximity or across repeats, whereas our HMM will distinguish adjacent duplicons (see Method). On the other hand, the average copy number of our duplicons is 644, which is much larger than that of previous study (~6 copies) [24]. This is not unexpected since our method assessed the statistical significance of each duplicon by a permutation test on the copy number. Therefore, duplicons without sufficient copy number are discarded. Nevertheless, even with a more

stringent criterion, we still identified many duplicons with long length (>10,000 bp) and with high frequency of copies (>2,000 copies).

**3.2. Mosaic Structure is Common in Human and Chimpanzee.** Our duplicons were annotated by CNVs and SDs in human. The distribution of these duplicons within CNVs and SDs in other primates is still unclear. Therefore, we downloaded chimpanzee and human SDs identified by self-comparison of the chimpanzee assembly and alignment of shotgun sequences [10]. These SDs were classified into three categories: 219 chimpanzee specific SDs (i.e., chimpanzee SDs that do not overlap with any human SDs), 618 human specific SDs (i.e., human SDs that do not overlap with any chimpanzee SDs), and 658 human/chimpanzee shared SDs. Our duplicons were BLAST aligned to SDs. Table 2 lists the number (and percentage) for each type of SDs containing our duplicons. The results indicated that our duplicons also appeared in majority of chimpanzee specific SDs (which are not included in our annotation process). In fact, over 98% of SDs in all three categories contained our duplicons. Furthermore, each SD includes an average of 24~43 duplicons, regardless of chimpanzee specific or human specific SDs. Consequently, these results suggest that the mosaic structure composed of duplicons is not only limited to human SDs but is also common in chimpanzee SDs.

Similarly, we compare the distribution of duplicons within CNVs between human and chimpanzee. 353 and 438 CNVs in the genomes of 30 humans and 30 chimpanzees were obtained from Perry et al. [17], respectively. These CNVs were also classified into 288 chimpanzee specific CNVs, 207 human specific CNVs, and 296 human/chimpanzee shared CNVs. As shown in Table 2, all of chimpanzee specific CNVs also contain our duplicons, indicating that these duplicons

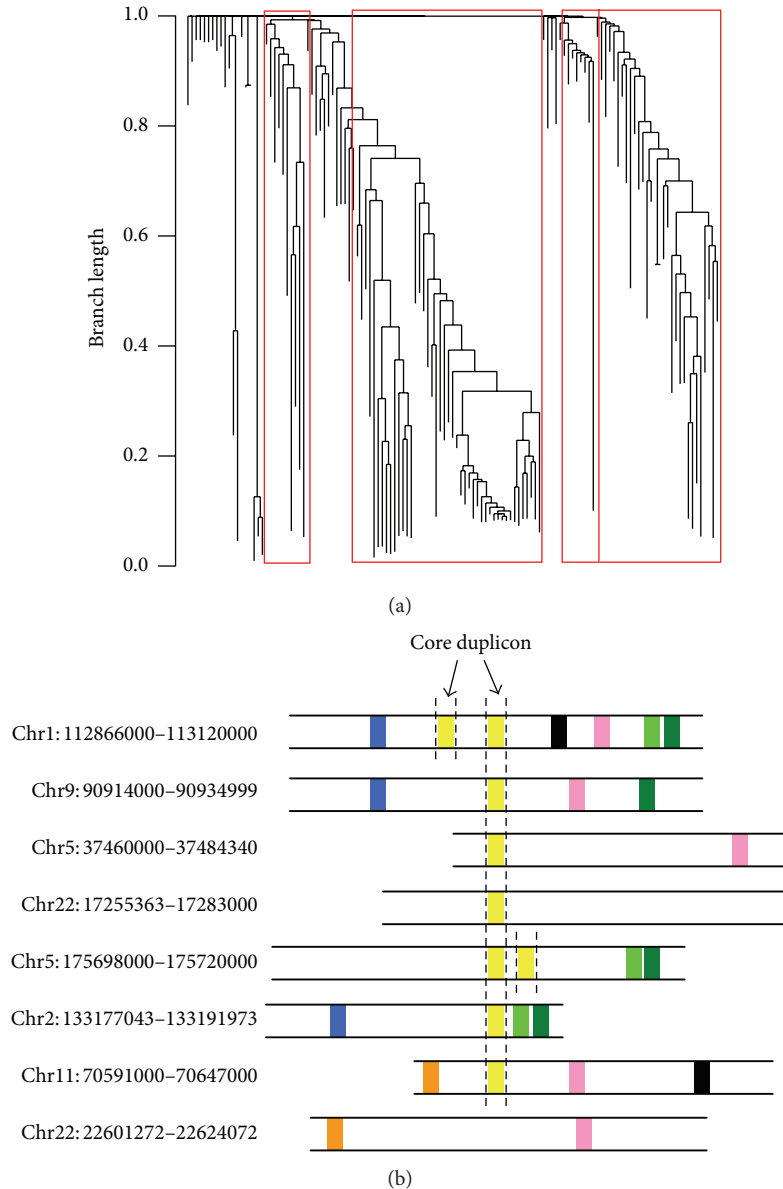


FIGURE 6: (a) Chimpanzee specific SDs are clustered by running Neighbor-Joining algorithm on their phylogenetic profiles constructed by duplicons. Four clades are revealed in this phylogenetic tree. (a) A cluster of chimpanzee specific SDs with shared duplicons. Different colors denote distinct duplicons. A core duplication shared by a majority of these SDs is highlighted by vertical dash lines.

are not limited to human CNVs. Overall, the majority of CNVs in three categories includes our duplicons, and each CNV contains approximately 16~22 duplicons. This phenomenon shows that duplicons are also common in chimpanzee CNVs. Compared with the results on SDs, the average numbers of duplicons on each CNV or SD are also quite similar. Consequently, the mosaic structure of juxtaposed duplicons may be common within SDs and CNVs in hominoid.

*3.3. Phylogenetic Analysis and Identification of Core Duplicons.* A number of studies suggested that secondary duplications

may have occurred recently among existing duplications, and these recent duplications tend to share more duplicons in common [24]. Thus, we reconstruct phylogenetic history of these SDs and CNVs using a representation of duplicons called phylogenetic profile [24]. A phylogenetic profile is created for each SD and CNV based on the presence or absence of each duplication (see Method). For each group of human specific, chimpanzee specific, and human/chimpanzee shared SDs and CNVs from [17], a phylogenetic tree is reconstructed by running the Neighbor-Joining algorithm on their phylogenetic profiles constructed by duplicons [33]. That is, the branch length reflects the degree of SDs/CNVs having the same duplicons in common.

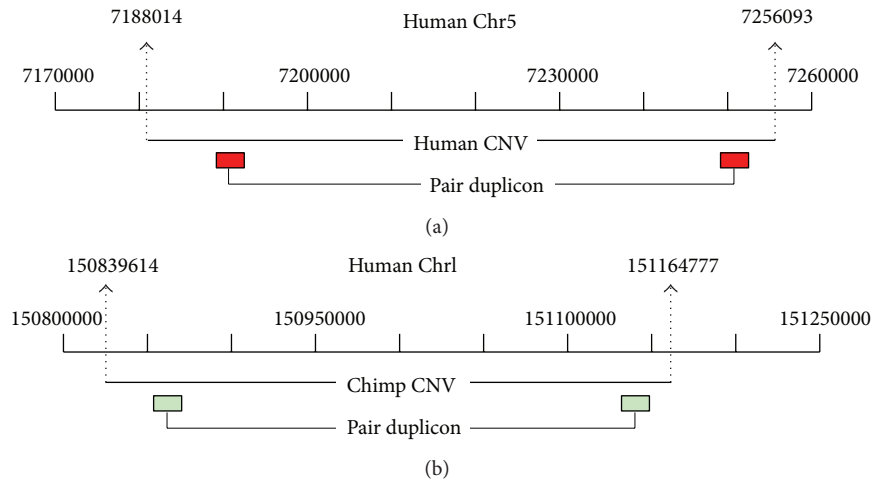


FIGURE 7: (a) The human CNV is flanked by two identical duplicons at both ends; (b) the chimpanzee CNV is flanked by two identical duplicons at both ends.

Figure 6(a) illustrates one phylogenetic tree reconstructed via duplicon profiles for chimpanzee specific SDs, where the other phylogenetic results can be found in Supplementary Figures 1–6 (see Supplementary Material available online at <http://dx.doi.org/10.1155/2013/264532>). Together, these results suggested that many of these SDs and CNVs share common ancestry of duplications, which are probably owing to recurrent duplications from a few seeding duplication blocks.

A large fraction of recent duplications have been shown to be centered around a small subset of “core duplicons” [24]. The structure of core duplicons with flanking duplicons is speculated to drive the rapid expansion of SDs widespread in hominoid genomes. The phylogenetic clustering of SDs or CNVs with common ancestry can be further used for identifying these core duplicons, which are shared by majority of SDs/CNVs in the same clade. A core duplicon is defined as a duplicon shared by >67% of SDs/CNVs in the same clade [24]. Figure 6(b) illustrates one core duplicon found in a clade. A total of 639 core duplicons were found. In summary, our analysis shows that many SDs and CNVs in human and chimpanzee have a nonrandom clustering structure of common duplicons and ancestry, and a number of core duplicons with flanking duplicons may trigger further duplications leading to novel SDs or CNVs.

**3.4. Comparison of Duplicon Densities in CNVs and Non-CNV Regions.** Duplicons identified by our pipeline were based on a subset of known CNVs in the human genome. As novel CNVs were reported by new sequencing projects, the power of our method can be estimated by observing the density of our duplicons in other newly annotated CNVs and non-CNV regions. Coordinates of 21,678 human CNVs are obtained from the Database of Genomic Variants (<http://projects.tcag.ca/variation>). Overlapping CNVs are merged and the 1,447 training CNVs used for annotating our duplicons are excluded. Non-CNV regions are defined as the genomic regions in between these known CNV regions.

Note that non-CNV regions may still contain some CNVs not annotated. We first align all duplicons against the entire human genome and compute the duplicon density in CNV and non-CNV regions. Since core duplicons tend to be shared by more CNVs than noncore duplicons, each duplicon is assigned a weight reflecting its frequency in the training CNVs. The weighted density in one genomic region is defined as the summation of total weights of duplicons aligned to this region divided by the region length.

Table 3 lists average densities of all CNVs and non-CNV regions separately for each chromosome. The average densities in CNVs and non-CNV regions in the entire genome are 4.307 and 1.767, respectively. The density is significantly higher in CNV than non-CNV regions ( $P < 10^{-5}$ ; two-tailed Welch’s  $t$  test). Although our duplicons are annotated from a subset of CNVs in the human genome, the results show that these duplicons also pervasively appear in other known CNV regions. And core duplicons are indeed more common in all CNVs. In non-CNV regions, there could be some CNVs still uncovered, because we still found a few genomic regions with high density.

**3.5. NAHR Mediated by Flanking Duplicons.** A number of studies have noted that genomic regions flanked by duplicated sequences are susceptible to NAHR [1, 9, 15, 21, 22, 29]. These regions are often hotspots of genomic instability that was prone to recurrent CNVs. A recent analysis of a fosmid clone indicated that a CNV is flanked by a pair of duplicons [9]. Figures 7(a) and 7(b) illustrate one human CNV and one chimpanzee CNV with flanking duplicons annotated by our pipeline. As a consequence, we are interested in the distribution of duplicons that locate in flanking regions of CNVs. A pair of duplicons is defined as flanking a CNV if it appears within 25% regions from two ends of the CNV and the similarity (and length) is >90%.

We first investigated 1,097 human CNVs with duplicons annotated by our pipeline [15]. Among them, 1,035 (94%) CNVs have two or more duplicons within their genomic

TABLE 1: The total number of duplicons of each chromosome.

Chr.	No. of dup.	Chr.	No. of dup.	Chr.	No. of dup.	Chr.	No. of dup.
1	6047	7	5329	13	216	19	2346
2	3607	8	2192	14	889	20	408
3	2142	9	4049	15	2621	21	143
4	1847	10	2039	16	5659	22	1430
5	2537	11	1873	17	3266	X	3078
6	1681	12	1989	18	599	Y	390

TABLE 2: The distribution of duplicons on human/chimpanzee SDs and CNVs. The number of hits stands for the number of SDs/CNVs containing our duplicons. The percentage of hits is shown in brackets. The last column is the average number of duplicons and the percentage of base pair in one SD or CNV.

Data set	Total no.	No. of hits (%)	Average no.
Chimpanzee-specific SDs	219	219 (100%)	43
Human-specific SDs	618	603 (98%)	31
Human/chimp-shared SDs	658	654 (99%)	24
Chimpanzee-specific CNVs	288	288 (100%)	16
Human-specific CNVs	207	206 (99%)	23
Human/chimp-shared CNVs	296	252 (85%)	22

TABLE 3: The average densities of duplicons in CNV and non-CNV regions on each chromosome.

Chr.	CNV	Non-CNV	Chr.	CNV	Non-CNV
1	1.95	1.48	13	1.58	1.51
2	2.13	1.83	14	2.59	1.56
3	2.53	2.11	15	1.80	1.27
4	2.68	2.44	16	1.25	0.92
5	3.03	2.17	17	0.74	0.90
6	2.67	1.93	18	2.51	1.68
7	1.97	1.84	19	0.93	0.44
8	2.77	2.14	20	1.27	1.37
9	2.11	1.50	21	1.50	0.65
10	1.77	1.75	22	0.62	0.22
11	2.90	1.77	X	3.21	3.04
12	1.92	1.97	Y	2.77	0.89

region. 815 out of 1,097 human CNVs (74%) were found to have paired duplicons flanking 25% of both ends. We also analyzed 791 human and chimpanzee CNVs from Perry et al. [17]. Our results indicated that 519 human/chimpanzee CNVs (66%) are also flanked by paired duplicons. Interestingly, each of these CNVs contains averagely ~11 paired duplicons, which could be hotspots of NAHR. This implies that further NAHR occurred within these CNVs may create different breaking points, leading to a complex duplication-within-duplication structure. Thus, these genomic regions may be prone to recurrent CNVs. However, it should be noted that our analysis is based on predefined CNV boundaries, which have been shown to be overestimated [34]. Thus, the requirement of 25% from both ends may eliminate many paired duplicons within real CNV boundaries. Nevertheless, our results provided evidence that there are many paired duplicons within or

surrounding a CNV region. As a consequence, boundaries of these complex CNVs may be hard to delineate, since NAHR may reoccur in different breaking points.

*3.6. Comparison with Duplicon Library.* We compared sequences of our duplicons with those in the duplicon library [9], which contains 10,291 duplicon sequences. Our duplicons were BLAST aligned against each duplicon sequence in the library (we considered the alignment results with sequence identities  $\geq 95\%$  and length  $\geq 100$  bp). In total, 16,819 (30%) of our duplicons were overlapped with 2,359 (23%) of the duplicon library. It has been shown that ~24% of CNVs are overlapped with SDs [15]. Thus, the difference between our duplicons and duplicon library is probably due to the fact that our duplicons were annotated based on CNVs, whereas duplicons in the library were identified solely based on SDs. However, it should be noted that duplicons with insignificant copy numbers were filtered by our permutation test. Thus, the difference between our duplicons and the duplicon library is not unexpected.

We further compare the distribution of duplicons on chimpanzee specific SDs and CNVs from [17]. These chimpanzee SDs and CNVs are not included in both studies and thus can observe distribution of these duplicons on nonhuman duplications. Table 4 summarizes the differences between our duplicons and the duplication library. There are 1,048 duplicons in the duplication library overlapped with chimp-specific SDs. Of these, 681 duplicons (65%) are also overlapped with our duplicons. On the other hand, there are 3,310 duplicons annotated by our HMM overlapped with chimp-specific SDs. Of these, 2,554 (82%) are also overlapped with duplicons in the library. In the analysis of CNVs, 1,510 duplicons in the library are located in chimp-specific CNVs. Of these, 886 (59%) duplicons are also overlapped with our

TABLE 4: Comparison of duplicons annotated by HMM and the duplication library. The numbers of (1) duplicons overlapped with each other, (2) duplicons overlapped with chimp-specific SDs, and (3) duplicons overlapped with chimp-specific CNVs are listed for each set of duplicons.

	Our duplicons	Duplib
Total No. of duplicons	56377	10291
No. of duplicons satisfying (1)	16819	2359
No. of duplicons satisfying (2)	3110	1048
No. of duplicons satisfying (1) and (2)	2554	681
Percentage	15% (2554/16819)	29% (681/2359)
Percentage	82% (2554/3110)	65% (681/1048)
No. of duplicons satisfying (3)	2645	1510
No. of duplicons satisfying (1) and (3)	2209	886
Percentage	13% (2209/16819)	38% (886/2359)
Percentage	84% (2209/2645)	59% (886/1510)

TABLE 5: GO analysis of biological process at levels 2 and 3. *P* values are computed by chi-square test with Bonferroni correction.

GO term	GO category	<i>P</i> value	Obs./exp.
	Level 2		
GO:0000003	Metabolic process	$4.36 \times 10^{-9}$	0.75
GO:0001906	Multicellular organismal process	$8.22 \times 10^{-8}$	1.36
GO:0002376	Biological adhesion	$1.10 \times 10^{-6}$	0.28
GO:0008152	Cellular process	$3.53 \times 10^{-6}$	1.16
GO:0009987	Developmental process	$4.35 \times 10^{-6}$	1.33
GO:0010926	Positive regulation of biological process	$4.70 \times 10^{-3}$	1.38
GO:0016032	Regulation of biological process	$1.90 \times 10^{-2}$	0.86
GO:0022414	Locomotion	$2.80 \times 10^{-2}$	0.44
	Level 3		
GO:0048856	Anatomical structure development	$1.40 \times 10^{-13}$	1.71
GO:0051239	Regulation of multicellular organismal process	$6.28 \times 10^{-13}$	2.33
GO:0043170	Macromolecule metabolic process	$1.53 \times 10^{-9}$	0.65
GO:0009058	Biosynthetic process	$2.64 \times 10^{-9}$	0.57
GO:0002682	Regulation of immune system process	$1.10 \times 10^{-8}$	2.70
GO:0019222	Regulation of metabolic process	$1.74 \times 10^{-8}$	0.53
GO:0007275	Multicellular organismal development	$8.82 \times 10^{-8}$	1.53
GO:0048518	Positive regulation of biological process	$5.32 \times 10^{-7}$	1.68
GO:0007154	Cell communication	$4.69 \times 10^{-6}$	0.65
GO:0001816	Cytokine production	$4.98 \times 10^{-6}$	2.89
GO:0051656	Establishment of organelle localization	$6.84 \times 10^{-6}$	4.29
GO:0045321	Leukocyte activation	$1.35 \times 10^{-5}$	2.44
GO:0032879	Regulation of localization	$3.90 \times 10^{-5}$	2.14
GO:0044238	Primary metabolic process	$1.62 \times 10^{-4}$	0.77
GO:0001775	Cell activation	$1.92 \times 10^{-4}$	2.21
GO:0055114	Oxidation reduction	$2.17 \times 10^{-4}$	0.15
GO:0048583	Regulation of response to stimulus	$5.14 \times 10^{-4}$	2.24
GO:0051050	Positive regulation of transport	$6.46 \times 10^{-4}$	2.84
GO:0007155	Cell adhesion	$1.08 \times 10^{-3}$	0.34
GO:0032898	Neurotrophin production	$6.88 \times 10^{-3}$	18.9
GO:0060033	Anatomical structure regression	$1.81 \times 10^{-2}$	9.47
GO:0008283	Cell proliferation	$2.39 \times 10^{-2}$	0.45

TABLE 6: GO analysis of molecular function at levels 2 and 3. *P* values are computed by chi-square test with Bonferroni correction.

GO term	GO category	<i>P</i> value	Obs./exp.
Level 2			
GO:0003824	Catalytic activity	$1.53 \times 10^{-33}$	1.78
GO:0005488	Binding	$1.41 \times 10^{-27}$	0.62
GO:0005215	Transporter activity	$4.72 \times 10^{-14}$	2.08
GO:0030528	Transcription regulator activity	$3.31 \times 10^{-5}$	0.37
GO:0015457	Auxiliary transport protein activity	$4.18 \times 10^{-3}$	3.92
GO:0005198	Structural molecule activity	$1.01 \times 10^{-2}$	0.40
Level 3			
GO:0022857	Transmembrane transporter activity	$7.32 \times 10^{-31}$	3.10
GO:0004133	Glycogen debranching enzyme activity	$4.35 \times 10^{-30}$	71.2
GO:0016740	Transferase activity	$4.10 \times 10^{-28}$	2.52
GO:0022892	Substrate-specific transporter activity	$1.64 \times 10^{-25}$	2.83
GO:0043167	Ion binding	$3.89 \times 10^{-24}$	0.12
GO:0003676	Nucleic acid binding	$1.42 \times 10^{-14}$	0.23
GO:0000166	Nucleotide binding	$6.41 \times 10^{-11}$	0.15
GO:0016491	Oxidoreductase activity	$8.19 \times 10^{-9}$	2.31
GO:0005515	Protein binding	$2.84 \times 10^{-6}$	0.67
GO:0016787	Hydrolase activity	$3.77 \times 10^{-6}$	1.61
GO:0016787	Transcription factor activity	$1.97 \times 10^{-5}$	0.07
GO:0016787	Channel regulator activity	$9.61 \times 10^{-4}$	4.97
GO:0016787	Bacterial binding	$9.69 \times 10^{-4}$	8.21
GO:0016787	Cell surface binding	$3.80 \times 10^{-3}$	5.93
GO:0016787	Peptide binding	$4.77 \times 10^{-2}$	1.90
GO:0016787	Signal transducer activity	$9.11 \times 10^{-2}$	1.35

TABLE 7: GO analysis of cellular component at levels 2 and 3. *P* values are computed by chi-square test with Bonferroni correction.

GO term	GO category	<i>P</i> value	Obs./exp.
Level 2			
GO:0032991	Macromolecular complex	$4.14 \times 10^{-15}$	1.97
GO:0044422	Organelle part	$8.08 \times 10^{-9}$	1.59
GO:0005576	Extracellular region	$3.09 \times 10^{-5}$	0.35
Level 3			
GO:0043234	Protein complex	$3.42 \times 10^{-20}$	2.36
GO:0044422	Organelle part	$2.22 \times 10^{-9}$	1.65
GO:0044446	Intracellular organelle part	$5.95 \times 10^{-9}$	1.64
GO:0044463	Cell projection part	$1.17 \times 10^{-8}$	5.17
GO:0042995	Cell projection	$2.09 \times 10^{-5}$	2.48
GO:0016020	Membrane	$8.17 \times 10^{-5}$	0.65
GO:0044425	Membrane part	$8.70 \times 10^{-5}$	0.62
GO:0032311	Angiogenin-PR1 complex	$5.55 \times 10^{-4}$	21.5
GO:0043227	Membrane-bounded organelle	$2.04 \times 10^{-3}$	0.71
GO:0032994	Protein-lipid complex	$3.68 \times 10^{-3}$	7.18
GO:0034358	Plasma lipoprotein particle	$3.68 \times 10^{-3}$	7.18

duplicons. Among our 2,645 duplicons located within chimp-specific CNVs, 2,209 (84%) are overlapped with duplicons in their library.

These results suggested that duplicons identified by both approaches all appear partially in chimp-specific SDs and

CNVs. However, given the higher percentage of our duplicons intersected with both chimp-specific SDs/CNVs and duplicons in the library (82% and 84% versus 65% and 59%), we concluded that duplicons found by our approach are more conservative. This may be due to the requirement of

sufficient copy number in our HMM and permutation test, whereas duplicon copies in the library are not validated with a statistical approach.

In terms of efficiency, it is worth mentioning that our HMM is quite efficient compared with the *A*-Bruijn graph algorithm, which requires 29 gigabytes of memory from 32 gigabyte computational cluster [24]. Our HMM can finish the computation within hours on a standard workstation. Consequently, novel duplicons can be efficiently annotated when more CNVs and SDs in other primate genomes are available.

**3.7. Functional Implication of Duplicons.** Our duplicons are smaller subunits within human CNVs. The functional analysis of these duplicons may provide new insight into functional bias not found in previous CNV analysis. We examined the functional bias of our duplicons in gene ontology (GO) categories and compared results with previous analysis of human CNVs. A total of 3,904 genes annotated by Ensembl are overlapped with our duplicons. Tables 5, 6, and 7 list the GO categories (at levels 2 and 3) with over- or underrepresentation of our duplicons ( $P < 0.05$ ; chi-square tests with Bonferroni correction).

For functions related to biological process, we found that eight function categories at level two were significantly biased to our duplicons. At level three, 22 of the 184 GO functions were over- or underrepresented with our duplicons (Table 5). In general, regulation of multicellular organismal process and of biological process is significantly enriched. The highly enriched GO categories overlapped partially with those identified in a previous analysis of CNVs [15], such as regulation of immune system process and regulation of response to stimulus. In contrast to previous analysis, cell adhesion was found to be underrepresented in duplicons. In addition, categories of neurophysiological processes and sensory perception enriched for CNVs are not found to be significantly enriched in duplicons. On the other hand, cell proliferation, oxidation reduction, and metabolic process are found to be significantly underrepresented among duplicons. The impoverishment of these functions probably reflects that purifying selection is against duplicons on dosage of these genes.

In terms of molecular functions, six GO terms at level two and 16 GO terms at level three are over- or underrepresented (Table 6). Specifically, duplicons are overrepresented in catalytic activity, transporter activities, and auxiliary transport protein activity. On the other hand, majority of binding activities, including ion binding, nucleic acid binding, and nucleotide binding are, underrepresented. These results suggest that distinct levels of evolutionary constraint on duplicons vary among functional categories.

## Acknowledgments

The authors thank the reviewers for their valuable comments. Shian-Zu Wu and Yao-Ting Huang were supported in part by NSC Grants 101-2221-E-194-MY3 and 101-2627-B-194-002. Trees-Juen Chuang was supported in part by NSC Grant 99-2628-B-001-008-MY3.

## References

- [1] A. J. Sharp, D. P. Locke, S. D. McGrath et al., "Segmental duplications and copy-number variation in the human genome," *American Journal of Human Genetics*, vol. 77, no. 1, pp. 78–88, 2005.
- [2] J. A. Bailey, Z. Gu, R. A. Clark et al., "Recent segmental duplications in the human genome," *Science*, vol. 297, no. 5583, pp. 1003–1007, 2002.
- [3] X. She, Z. Jiang, R. A. Clark et al., "Shotgun sequence assembly and recent segmental duplications within the human genome," *Nature*, vol. 431, no. 7011, pp. 927–930, 2004.
- [4] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 85–97, 2006.
- [5] T. J. Aitman, R. Dong, T. J. Vyse et al., "Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans," *Nature*, vol. 439, no. 7078, pp. 851–855, 2006.
- [6] E. Gonzalez, H. Kulkarni, H. Bolivar et al., "The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility," *Science*, vol. 307, no. 5714, pp. 1434–1440, 2005.
- [7] A. B. Singleton, M. Farrer, J. Johnson et al., " $\alpha$ -synuclein locus triplication causes Parkinson's disease," *Science*, vol. 302, no. 5646, p. 841, 2003.
- [8] A. Rovelet-Lecrux, D. Hannequin, G. Raux et al., "APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy," *Nature Genetics*, vol. 38, no. 1, pp. 24–26, 2006.
- [9] Z. Jiang, R. Hubley, A. Smit, and E. E. Eichler, "DupMasker: a tool for annotating primate segmental duplications," *Genome Research*, vol. 18, no. 8, pp. 1362–1368, 2008.
- [10] Z. Cheng, M. Ventura, X. She et al., "A genome-wide comparison of recent chimpanzee and human segmental duplications," *Nature*, vol. 437, no. 7055, pp. 88–93, 2005.
- [11] X. She, Z. Cheng, S. Zöllner, D. M. Church, and E. E. Eichler, "Mouse segmental duplication and copy number variation," *Nature Genetics*, vol. 40, no. 7, pp. 909–914, 2008.
- [12] A. S. Lee, M. Gutiérrez-Arcelus, G. H. Perry et al., "Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies," *Human Molecular Genetics*, vol. 17, no. 8, pp. 1127–1136, 2008.
- [13] T. J. Nicholas, Z. Cheng, M. Ventura, K. Mealey, E. E. Eichler, and J. M. Akey, "The genomic architecture of segmental duplications and associated copy number variants in dogs," *Genome Research*, vol. 19, no. 3, pp. 491–499, 2009.
- [14] F.-C. Chen, Y.-Z. Chen, and T.-J. Chuang, "CNVdb: a database of copy number variations across vertebrate genomes," *Bioinformatics*, vol. 25, no. 11, pp. 1419–1421, 2009.
- [15] R. Redon, S. Ishikawa, K. R. Fitch et al., "Global variation in copy number in the human genome," *Nature*, vol. 444, no. 7118, pp. 444–454, 2006.
- [16] G. H. Perry, J. Tchinda, S. D. McGrath et al., "Hotspots for copy number variation in chimpanzees and humans," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 21, pp. 8006–8011, 2006.
- [17] G. H. Perry, F. Yang, T. Marques-Bonet et al., "Copy number variation and evolution in humans and chimpanzees," *Genome Research*, vol. 18, no. 11, pp. 1698–1710, 2008.
- [18] D. Komura, F. Shen, S. Ishikawa et al., "Genome-wide detection of human copy number variations using high-density DNA

- oligonucleotide arrays,” *Genome Research*, vol. 16, no. 12, pp. 1575–1584, 2006.
- [19] T. S. Price, R. Regan, R. Mott et al., “SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data,” *Nucleic Acids Research*, vol. 33, no. 11, pp. 3455–3464, 2005.
- [20] K. Chen, J. W. Wallis, M. D. McLellan et al., “BreakDancer: an algorithm for high-resolution mapping of genomic structural variation,” *Nature Methods*, vol. 6, no. 9, pp. 677–681, 2009.
- [21] J. Sebat, B. Lakshmi, J. Troge et al., “Large-scale copy number polymorphism in the human genome,” *Science*, vol. 305, no. 5683, pp. 525–528, 2004.
- [22] A. J. Sharp, Z. Cheng, and E. E. Eichler, “Structural variation of the human genome,” *Annual Review of Genomics and Human Genetics*, vol. 7, pp. 407–442, 2006.
- [23] H. Xiao, N. Jiang, E. Schaffner, E. J. Stockinger, and E. Van Der Knaap, “A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit,” *Science*, vol. 319, no. 5869, pp. 1527–1530, 2008.
- [24] Z. Jiang, H. Tang, M. Ventura et al., “Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution,” *Nature Genetics*, vol. 39, no. 11, pp. 1361–1368, 2007.
- [25] C. L. Kahn and B. J. Raphael, “A parsimony approach to analysis of human segmental duplications,” *Pacific Symposium on Biocomputing*, vol. 14, pp. 126–137, 2009.
- [26] J. A. Bailey and E. E. Eichler, “Primate segmental duplications: crucibles of evolution, diversity and disease,” *Nature Reviews Genetics*, vol. 7, no. 7, pp. 552–564, 2006.
- [27] E. E. Eichler, M. L. Budarf, M. Rocchi et al., “Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity,” *Human Molecular Genetics*, vol. 6, no. 7, pp. 991–1002, 1997.
- [28] P. A. Pevzner, H. Tang, and G. Tesler, “De novo repeat classification and fragment assembly,” *Genome Research*, vol. 14, no. 9, pp. 1786–1796, 2004.
- [29] P. M. Kim, H. Y. K. Lam, A. E. Urban et al., “Analysis of copy number variants and segmental duplications in the human genome: evidence for a change in the process of formation in recent evolutionary history,” *Genome Research*, vol. 18, no. 12, pp. 1865–1874, 2008.
- [30] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.
- [31] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, “A greedy algorithm for aligning DNA sequences,” *Journal of Computational Biology*, vol. 7, no. 1-2, pp. 203–214, 2000.
- [32] L. R. Rabiner, “Tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [33] M. A. Larkin, G. Blackshields, N. P. Brown et al., “Clustal W and Clustal X version 2.0,” *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.
- [34] S. A. McCarroll and D. M. Altshuler, “Copy-number variation and association studies of human disease,” *Nature Genetics*, vol. 39, no. 1, pp. S37–S42, 2007.