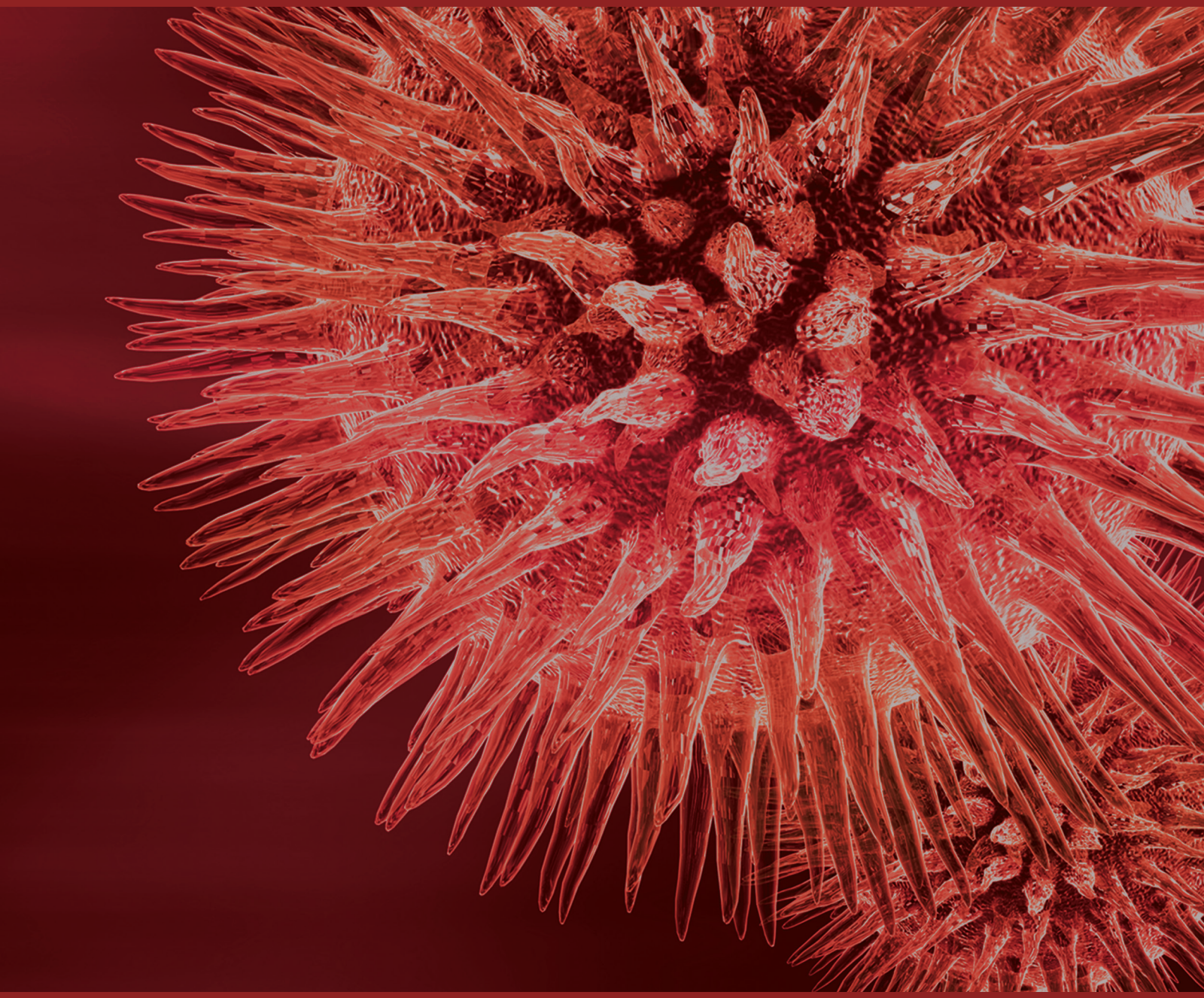
BioMed Research International

Sequence and Structure Analysis of Biological Molecules Based on Computational Methods

Guest Editors: Jia-Feng Yu, Yue-Dong Yang, Xiao Sun, and Ji-Hua Wang



Sequence and Structure Analysis of Biological Molecules Based on Computational Methods

BioMed Research International

Sequence and Structure Analysis of Biological Molecules Based on Computational Methods

Guest Editors: Jia-Feng Yu, Yue-Dong Yang, Xiao Sun, and Ji-Hua Wang

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 565328, 3 pages
http://dx.doi.org/10.1155/2015/565328

Editorial

Sequence and Structure Analysis of Biological Molecules Based on Computational Methods

Jia-Feng Yu,[1] Yue-Dong Yang,[2] Xiao Sun,[3] and Ji-Hua Wang[1,4]

[1] *Shandong Provincial Key Laboratory of Functional Macromolecular Biophysics, Institute of Biophysics, Dezhou University, Dezhou 253023, China*
[2] *Institute for Glycomics, Griffith University, Southport, QLD 4222, Australia*
[3] *State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China*
[4] *College of Physics and Electronic Information, Dezhou University, Dezhou 253023, China*

Correspondence should be addressed to Jia-Feng Yu; jfyu1979@126.com, Yue-Dong Yang; yuedong.yang@gmail.com, Xiao Sun; xsun@seu.edu.cn, and Ji-Hua Wang; jhw25336@126.com

Received 14 April 2015; Accepted 14 April 2015

The number of sequences and structures of biological molecules, such as DNA, RNA, and proteins, is rapidly increasing in databases. According to the statistics in the newest version of GOLD database (https://gold.jgi-psf.org/) [1], 6651 genomes have been completed and published and 51954 genomes are of permanent drafts or incomplete. The protein data bank (PDB, http://www.rcsb.org/) [2] has published 107958 biological macromolecular structures, including 35542 protein sequences, 28142 structures of human sequences, and 7611 nucleic acid containing structures. To discover the underlying mechanism behind the information, it is necessary to develop effective feature parameters for representing the sequence and structure. Although many studies have been proposed for sequence and structure analysis, more and more studies indicated that the problem is far from solved [3, 4]. How to computationally analyze the big data of biologically molecular sequences and structures has been one of the major challenges in bioinformatics. Machine learning and optimization methods are important methods for the analyses. This special issue focuses on recent progress of the computational methods for biological sequences or structures studies. Correspondingly, after a rigorous peer review, eleven papers were selected. We briefly describe these papers as follows.

In "Predicting Homogeneous Pilus Structure from Monomeric Data and Sparse Constraints," K. Xiao et al. developed a new approach to predict pseudoatomic models of *pili* by combining ambiguous symmetric constraints with sparse distance information obtained from experiments. The method was successfully implemented to the reconstruct the gonococcal (GC) pilus from *Neisseria gonorrhoeae*. A global sampling in a wide range implied that a pilus might have more than one but fewer than many possible intact conformations.

In "Evolutionary and Expression Analysis of miR-#-5p and miR-#-3p at the miRNAs/isomiRs Levels," L. Guo et al. explored the potential evolutionary and expression divergence and relationships between miRNAs from different arms of different/same pre-miRNAs according to the arm selection and/or arm switching phenomenon in miRNA world. They found no bias in the numbers but different nucleotide compositions between 5p-miRNA and 3p-miRNA. IsomiR expression profiles from the two arms are always stable, but isomiR expressions in diseased samples are prone to show larger degree of dispersion. miR-#-5p and miR-#-3p have relative independent evolution/expression patterns and datasets of target mRNAs, which might also contribute to the phenomena of arm selection and/or arm switching. Simultaneously, miRNA/isomiR expression profiles may be regulated via arm selection and/or arm switching, and the dynamic miRNAome and isomiRome will adapt to functional and/or evolutionary pressures. A comprehensive analysis and further experimental study at the miRNA/isomiR levels are quite necessary for miRNA study.

In "Strong Ligand-Protein Interactions Derived from Diffuse Ligand Interactions with Loose Binding Sites," L. Marsh presented a fine-grained computational method for numerical integration of total binding free energy arising from diffuse regional interaction of a ligand in multiple conformations using a Markov Chain Monte Carlo. The application to the bacterial multidrug efflux pump AcrB indicated that diffuse binding effects could cause 100-fold binding affinity for some ligands while little role in the binding of other ligands. Analysis of other proteins with large binding pockets indicated that the influence of this process varies greatly, dependent on ligand and protein target. This work may be of broad interest to those studying a variety of biological systems including protein folding, DNA-protein binding, and drug-receptor docking that depend on dispersed interactions.

In "Redesigning Protein Cavities as a Strategy for Increasing Affinity in Protein-Protein Interaction: Interferon-γ Receptor 1 as a Model," J. Černý et al. presented a new strategy for designing high affinity variants of a binding protein through mutating residues at positions lining internal cavities of one of the interacting molecules instead of at the interface. The test on interferon-γ receptor 1 has brought up to sevenfold increase in the binding affinity. Analysis shows that the affinity increase is linked to the restriction of molecular fluctuations in the unbound state of the receptor. This serves as an example of a viable strategy for designing protein variants with increased affinity.

In "Multi-Instance Multilabel Learning with Weak-Label for Predicting Protein Function in Electricigens," J.-S. Wu et al. have applied the state-of-the-art MIML with weak-label learning algorithm MIMLwel for predicting protein functions in two typical real-world electricigens organisms widely used in microbial fuel cells studies. The experimental results validate the effectiveness of MIMLwel algorithm in predicting protein functions with incomplete annotation.

In "Detecting Protein-Protein Interactions with a Novel Matrix-Based Protein Sequence Representation and Support Vector Machines," Z.-H. You et al. developed a novel computational approach to effectively detect the protein interactions based on a novel matrix-based representation of protein sequence by SVM. The sequence information includes the order of amino acids and composition of dipeptide. The prediction was proven highly accurate on the benchmark of yeast PPIs datasets. Thus, it can be a helpful supplement to the missing data and false positive by experimental results from high-throughput techniques.

In "An Improved Method for Completely Uncertain Biological Network Alignment," B. Shen et al. designed an improved method to analyze uncertain biological networks by complete probabilistic biological network alignment. This method has solved the weakness of PBNA by allowing both networks to be probabilistic; thus, it could take full advantage of the uncertain information of biological network. The new method was proven to consistently improve over PBNA in both GO Consistency and Global Network Alignment Score.

In "A Systematic Analysis of Candidate Genes Associated with Nicotine Addiction," M. Liu et al. have performed a systematic analysis on a set of nicotine addiction-related genes (NAGenes) to explore their characteristics at network levels. They found that NAGenes tended to have a more moderate degree, weaker clustering coefficient in the network, and are less central in the network compared to genes related to alcohol addiction or cancer. In addition, an intuitional view was provided to understand their major molecular functions by six clusters from the clustering of these genes with themes in synaptic transmission, signal transduction, metabolic process, and apoptosis. Moreover, it was found that nicotine addiction involves neurodevelopment, neurotransmission activity, and metabolism related biological functions. The systematic network and functional enrichment analysis for nicotine addiction in this study is valuable for understanding the molecular mechanisms underlying nicotine addiction.

In "Constraint Programming based Biomarker Optimization," M. Zhou et al. proposed an algorithm for highly accurate feature selection for biomedical classification problem, while allowing the inclusion of user-input constraints for the optimization process. The experimental results showed that the proposed method provided flexibility of allowing both the well-known disease biomarkers like P53 and the existing feature selection algorithms' results as the constraints, while achieving promising classification performances. This work provided useful tool for efficiently exploring the health big data, consisting of both the bio-OMIC data and huge amount of knowledge accumulated in the literature and other sources.

In "An Improved Opposition-Based Learning Particle Swarm Optimization for the Detection of SNP-SNP Interactions," J. Shang et al. presented an improved method for detecting SNP-SNP interactions by using opposition-based learning particle swarm optimization. The algorithm has ensured the ability of global searching and prevented premature convergence. The application to a dataset of age-related macular degeneration shows the strength of the method on real applications by capturing important features of genetic architecture not previously discovered. The method provides an important tool in detecting SNP-SNP interactions in future.

In "Nucleosome Organization around Pseudogenes in the Human Genome," G. Liu et al. investigated the effect of nucleosome positioning on pseudogene transcription. The results show that, for transcribed pseudogenes, nucleosomes upstream of the start positions and end positions of transcribed pseudogenes are depleted. Interestingly, the same depletion is also observed for nontranscribed pseudogenes. The consistent pattern of sequence-dependent prediction with the assessment by experimental data indicates that sequence-dependent mechanism of nucleosome positioning may play important roles in both the transcription initiation and termination of pseudogenes.

Acknowledgments

We are grateful to the anonymous reviewers whose critical review helped improve the quality of the papers in this special issue. We would like to acknowledge the organizers and committee members of the Sixth National Conference on Bioinformatics and Systems Biology and International

Workshop on Advanced Bioinformatics (held in Nanjing, October 6–9, 2014) for their efforts in providing a forum to discuss sequence and structure analysis of biological molecules based on computational methods, through which this special issue was made possible.

Jia-Feng Yu
Yue-Dong Yang
Xiao Sun
Ji-Hua Wang

References

[1] T. B. Reddy, A. D. Thomas, D. Stamatis et al., "The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification," *Nucleic Acids Research*, vol. 43, no. D1, pp. D1099–D1106, 2015.

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

[3] J.-F. Yu, Z.-Z. Guo, X. Sun, and J.-H. Wang, "A review of the computational methods for identifying the over-annotated genes and missing genes in microbial genomes," *Current Bioinformatics*, vol. 9, no. 2, pp. 147–154, 2014.

[4] N. K. Petty, "Genome annotation: man versus machine," *Nature Reviews Microbiology*, vol. 8, article 762, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 524821, 12 pages
http://dx.doi.org/10.1155/2015/524821

Research Article

An Improved Opposition-Based Learning Particle Swarm Optimization for the Detection of SNP-SNP Interactions

Junliang Shang,[1] Yan Sun,[1] Shengjun Li,[1] Jin-Xing Liu,[1,2] Chun-Hou Zheng,[3] and Junying Zhang[4]

[1] *School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China*
[2] *Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China*
[3] *College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui 230039, China*
[4] *School of Computer Science and Technology, Xidian University, Xi'an 710071, China*

Correspondence should be addressed to Jin-Xing Liu; sdcavell@126.com

Received 25 September 2014; Revised 30 December 2014; Accepted 2 January 2015

Academic Editor: Yuedong Yang

SNP-SNP interactions have been receiving increasing attention in understanding the mechanism underlying susceptibility to complex diseases. Though many works have been done for the detection of SNP-SNP interactions, the algorithmic development is still ongoing. In this study, an improved opposition-based learning particle swarm optimization (IOBLPSO) is proposed for the detection of SNP-SNP interactions. Highlights of IOBLPSO are the introduction of three strategies, namely, opposition-based learning, dynamic inertia weight, and a postprocedure. Opposition-based learning not only enhances the global explorative ability, but also avoids premature convergence. Dynamic inertia weight allows particles to cover a wider search space when the considered SNP is likely to be a random one and converges on promising regions of the search space while capturing a highly suspected SNP. The postprocedure is used to carry out a deep search in highly suspected SNP sets. Experiments of IOBLPSO are performed on both simulation data sets and a real data set of age-related macular degeneration, results of which demonstrate that IOBLPSO is promising in detecting SNP-SNP interactions. IOBLPSO might be an alternative to existing methods for detecting SNP-SNP interactions.

1. Introduction

There is an increasing interest in understanding the underlying genetic architecture of complex diseases, such as cancer, heart disease, diabetes, Crohn's disease, and many others, which represent the major part of current clinical diseases [1, 2]. Research of complex diseases is one of the hottest fields of bioinformatics and genome-wide association studies (GWAS) become routine strategies. With the methods of GWAS, hundreds of thousands of single nucleotide polymorphisms (SNPs) speculated to associate with complex diseases have been identified. Nevertheless, these SNPs have limited effects on predicting the phenotype, and a large fraction of genetic contributions to complex diseases remain unclear. Recent advances make it clear that besides rare SNPs not genotyped in GWAS, the "missing heritability" can be partly explained by nonlinear interactive effects of multiple SNPs, namely, SNP-SNP interactions [3]. Detection of SNP-SNP interactions is therefore a compelling next step in GWAS.

In general, the detection of SNP-SNP interactions is a great challenge [4]. The first challenge is the intensive computational burden imposed by the enormous search space, which prohibits real applications of most existing methods, especially those exhaustive ones. For instance, search space of a 100,000-SNP data set with considered maximum order of 3 is an astronomical number $\sum_{k=1}^{3} C_{100,000}^{k}$. The second challenge is the complexity of genetic architecture of a complex disease. Limited or even no prior knowledge available for a complex disease, such as the order and the effect magnitude of a SNP-SNP interaction, makes it difficult for the development of heuristic methods. The third evaluation measures that determine how well a SNP combination contributes to

the phenotype are limited. Evaluation measures should be efficient in computational cost and insensitive to both SNP combination order and dependency type. Though several evaluation measures have been widely used in the detection of SNP-SNP interactions, developing new evaluation measures that can effectively and efficiently capture SNP-SNP interactions is still a direction.

Though methodological and computational perplexities of the detection of SNP-SNP interactions have been well recognized, the algorithmic development is still ongoing. Exhaustive algorithms, for example, MDR [5], appear promising for small scale data sets. However, for large scale data sets, especially those for GWAS, the detection of SNP-SNP interactions becomes a *needles-in-a-haystack* problem and exhaustive algorithms lose their ability [6, 7]. Heuristic algorithms are popular since they can retain as many informative SNPs as possible while largely reducing computational complexity. For example, Jiang et al. formulated the detection of SNP-SNP interactions from the viewpoint of binary classification and designed *epi*Forest on the basis of the gini importance given by the random forest to select a small set of candidate SNPs [8]. Zhang and Liu proposed a Bayesian partition approach BEAM to find groups of genotypes with large posterior probability [9]. Tang et al. introduced the concept of epistatic module and designed a Gibbs sampling approach *epi*MODE to detect such modules [10]. Wan et al. developed a SNP-SNP interaction detection method SNPRuler based on both predictive rule inference and two-stage design [11]. They also presented another method BOOST, which involves only Boolean values and allows the use of fast logic operations to obtain contingency tables [12]. Besides machine learning methods, entropy based methods are also applied to this field. Chanda et al. developed an interaction index based on entropy theory to prioritize interacting SNPs [13]. They also applied two entropy theoretic measures to three SNP-SNP interaction detection methods: AMBIENCE [14] with a phenotype associated information measure; KWII [15] with the coinformation measure detecting SNP-SNP interactions associated with the binary phenotype; and CHORUS [16] combining these two measures together to identify associations with quantitative traits.

Recently, many swarm intelligence based algorithms have been proposed for the detection of SNP-SNP interactions [17–31]. Among them, particle swarm optimization (PSO) appears promising and some related works have been reported [22–30]. Yang et al. [22] used the binary PSO with odds ratio as the fitness function (OR-BPSO) to evaluate the risk of breast cancer. Based on the OR-BPSO, Chang et al. [23] proposed the odds ratio-based discrete binary PSO (OR-DBPSO) for the detection of SNP-SNP interactions with the quantitative phenotype. Chuang et al. [24] proposed a chaotic PSO (CPSO) that identifies the best SNP combination for breast cancer association studies. For enhancing the reliability of the PSO in the identification of the best SNP-SNP interaction associated with breast cancer, they also developed an improved PSO (IPSO) [25] and proved that the IPSO is highly reliable than the OR-BPSO. More recently, they used the gauss chaotic map PSO (Gauss-PSO) to detect the best association with breast cancer [26]. Experimental results revealed that the Gauss-PSO was able to identify higher difference values between cases and controls than both the PSO and the CPSO. Yang et al. [27] developed a double-bottom chaotic map PSO (DBM-PSO) that overcomes the respective disadvantages of the PSO and the CPSO. Then, DBM-PSO is successfully applied to determine gene-gene interactions based on *chi*-square test [28]. Hwang et al. [29] proposed a complementary-logic PSO (CLPSO) to increase the efficiency of significant model identification in case-control study. Wu et al. [30] applied PSO to analyze the SNP-SNP interactions associated with hypertension. However, these methods, almost all of which are developed by a group except the one proposed by Wu et al. [30], only focus on finding the best genotype-genotype of a SNP-SNP interaction among possible genotypes of SNP combinations, but not the SNP-SNP interactions among possible SNP combinations. Obviously, the limited sample size of SNP data affects their computational accuracies of fitness functions and hence hinders their further applications. Furthermore, these methods are experimented on very small scale data sets (<30 SNPs) of certain complex diseases, performance of which on various kinds of large scale data sets are still unclear.

In this study, we proposed an improved opposition-based learning particle swarm optimization (IOBLPSO) with mutual information as its fitness function to detect SNP-SNP interactions. IOBLPSO is the first PSO based method to find SNP-SNP interactions among possible SNP combinations. Highlights of IOBLPSO are the introduction of three strategies, that is, opposition-based learning (OBL), dynamic inertia weight, and a postprocedure. Among them, OBL is the core, which is presented in the stage of updating particle experiences and common knowledge of swarm, not only for enhancing the global explorative ability, but also for avoiding premature convergence. Dynamic inertia weight is computed before the stage of updating particle velocities to allow particles to cover a wider search space when the considered SNP is likely to be a random SNP and to converge on promising regions of the search space while capturing a highly suspected SNP. The postprocedure is used as the final stage for carrying out a deep search in highly suspected SNP sets. Experiments of IOBLPSO are performed on lots of simulation data sets under the evaluation measures of both detection power and computational complexity. Results demonstrate that IOBLPSO is promising for the detection of all simulation models of SNP-SNP interactions. IOBLPSO is also applied on data set of age-related macular degeneration (AMD). Results show the strength of IOBLPSO on real applications and capture important features of genetic architecture of AMD that have not been described previously, which provide new clues for biologists on the exploration of AMD associated SNPs. IOBLPSO might be an alternative to existing methods for the detection of SNP-SNP interactions.

2. Methods

2.1. Particle Swarm Optimization (PSO). The PSO, proposed by Kennedy and Eberhart [32], is a member of the family of swarm intelligence algorithms, which mimic the collective

behaviors of organisms based on information sharing, like ants and birds, which can jointly perform many complex tasks though each individual is very limited in its capability. The PSO is a stylized representation of the movement of birds (viewed as particles) in a flock, where each particle uses its own experience and the common knowledge gained by the entire swarm to find an optimal position [29].

In PSO, the position of a particle represents a possible solution. In each generation, the position of each particle is adjusted according to its updated velocity and is estimated by a fitness function for providing a good search direction. Whether the velocity of each particle is updated depends on three variables: its previous velocity, its individual experience, and the common knowledge of the swarm. Specifically, the individual experience of each particle is updated while fitness value of its current position is higher than that of its previous experience; the common knowledge of the swarm is updated by the one of individual experiences of all particles with the highest fitness value while such value is higher than that of their previous common knowledge. This feedback strategy leads the swarm to gradually converge to an optimal solution [25–29].

Owing to its high capability and good generality in solving complex problems, the PSO has become a widely adopted swarm intelligence algorithm. However, it still has several defects, for example, premature convergence, stagnation phenomenon, and slow convergence speed in the later evolution period, which imply that the PSO should be further improved, especially for a specific complex problem, for example, the detection of SNP-SNP interactions. In general, the PSO consists of 4 stages: (1) initializing particles, (2) evaluating particles using fitness function, (3) updating particle velocities and positions, and (4) updating particle experiences and common knowledge of swarm. These stages are detailed and described in the following section.

2.2. IOBLPSO: An Improved Opposition-Based Learning Particle Swarm Optimization for the Detection of SNP-SNP Interactions. The flowchart of IOBLPSO is shown in Figure 1, where its highlights are with grey background. Below we describe IOBLPSO in detail from 6 stages.

(1) Mapping SNPs and Initializing Particles. At present, the popular way of mapping SNPs is to collect them as a matrix, where a row represents genotypes of an individual and a column represents a SNP. Genotypes of a SNP are coded as $\{0, 1, 2\}$, corresponding to homozygous common genotype (e.g., *AA*, *BB*), heterozygous genotype (e.g., *Aa*, *aA*, *Bb*, *bB*), and homozygous minor genotype (e.g., *aa*, *bb*). The label of an individual is a binary phenotype being either 0 (control) or 1 (case).

Based on above numerical mapping, the position of the p_{th} particle at iteration t can be represented as $\text{Position}_t(p) = (\text{SNP}^t_{p1}, \ldots, \text{SNP}^t_{pk}, \ldots, \text{SNP}^t_{pK})$, where $p \in \{1, 2, \ldots, P\}$, $k \in \{1, 2, \ldots, K\}$, $t \in \{1, 2, \ldots, T\}$, P is the number of particles, K is the considered order of SNP-SNP interactions, T is the number of iterations, SNP^t_{pk} is the index of the selected k_{th} SNP of the p_{th} particle at iteration t, $\text{SNP}^t_{pk} \in \{1, 2, \ldots, M\}$, and M is the number of SNPs in the data set. The velocity of the p_{th} particle at iteration t is represented as $\text{Velocity}_t(p) = (v^t_{p1}, \ldots, v^t_{pk}, \ldots, v^t_{pK})$, where v^t_{pk} is the velocity of SNP SNP^t_{pk} and $v^t_{pk} \in [1 - M, M - 1]$. Similarly, the individual experience of the p_{th} particle, that is, the position of the p_{th} particle with the highest fitness value until iteration t, can be denoted as $pbest_t(p) = (p\text{SNP}^t_{p1}, \ldots, p\text{SNP}^t_{pk}, \ldots, p\text{SNP}^t_{pK})$, and the common knowledge of swarm, that is, the best position of all particles with the highest fitness value until iteration t, is denoted as $gbest_t = (g\text{SNP}^t_1, \ldots, g\text{SNP}^t_k, \ldots, g\text{SNP}^t_K)$.

Before the first iteration, $\text{Position}_1(p)$, $\text{Velocity}_1(p)$, $pbest_1(p)$, and $gbest_1$ are randomly initialized in their respective domains.

(2) Updating Dynamic Inertia Weight. Inertia weight is used to control the impact of the previous velocity of a particle on its current velocity. A large inertia weight facilitates the global exploration and thus enables the method to execute a search over various regions, while a small inertia weight facilitates the local exploitation, which searches a promising region [27]. In order to effectively balance the global exploration and the local exploitation, a dynamic inertia weight is introduced to IOBLPSO, which can be defined as

$$W^t_{pk} = \frac{\max(\text{count}_t) - \text{count}_t\left[p\text{SNP}^t_{pk}\right]}{\max(\text{count}_t) - \min(\text{count}_t)}, \quad (1)$$

where $\text{count}_t = (\text{ct}^t_1, \ldots, \text{ct}^t_m, \ldots, \text{ct}^t_M)$ and ct^t_m is a counter that counts the number of SNP m presented in *pbest* from iteration 1 to iteration t. This strategy allows particles to cover a wider search space while the considered SNP is likely to be a random SNP and to converge on promising regions of the search space while capturing a highly suspected SNP.

(3) Evaluating Particles Using Fitness Function. Fitness function of the IOBLPSO plays an important role on deciding which SNP combination is the SNP-SNP interaction and measuring how much the effect of a captured SNP-SNP interaction to the phenotype is. In the IOBLPSO, mutual information is applied as its fitness function, since it is well developed and can measure multivariate dependence without complex modeling. Mutual information has been widely used as a promising measure for feature selection and here is defined as

$$MI(X; Y) = H(X) + H(Y) - H(X, Y), \quad (2)$$

where $H(X)$ is the entropy of X; X, representing a SNP combination, is the general expression of $\text{Position}_t(p)$, $pbest_t(p)$, and $gbest_t$; $H(Y)$ is the entropy of the phenotype Y; $H(X, Y)$ is the joint entropy of both X and Y. It is clear that higher mutual information value, namely, fitness value, indicates stronger association between the phenotype and the SNP combination.

(4) Updating Particle Velocities and Positions. IOBLPSO executes a search for SNP-SNP interactions by continuously

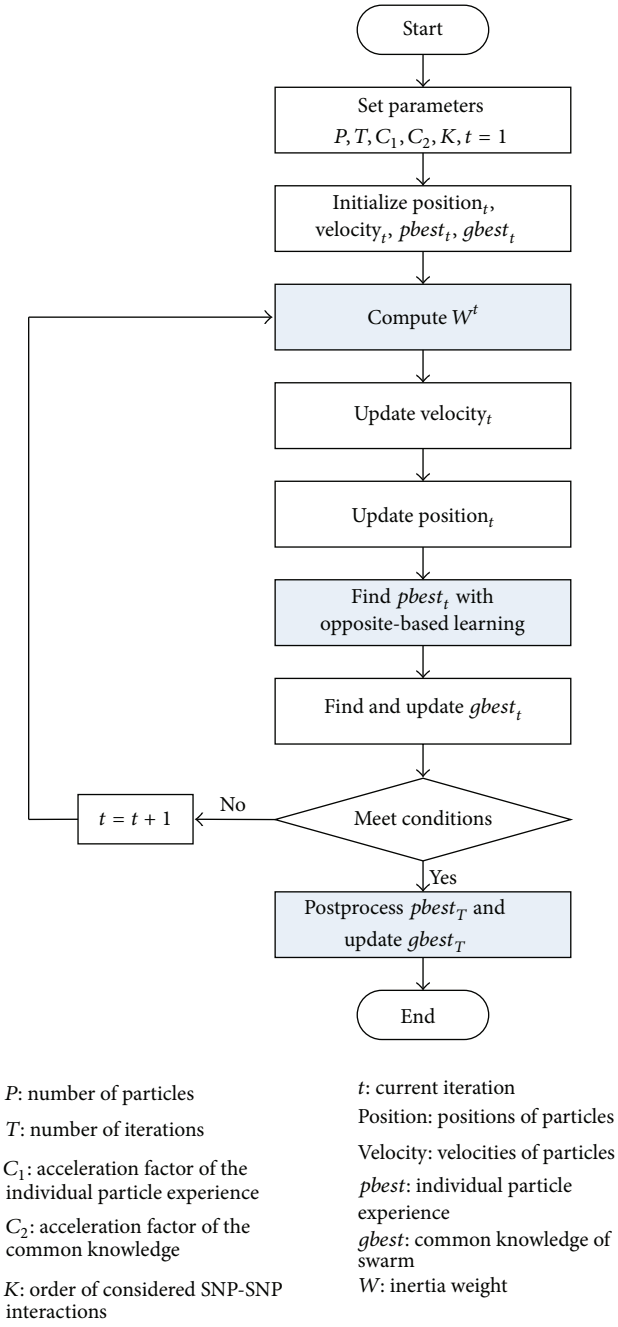


Figure 1: The flowchart of IOBLPSO. Three components with grey background are highlights of IOBLPSO.

updating particle velocities and particle positions in all iterations. The velocity of SNP^t_{pk} is updated using the following equations:

$$\begin{aligned} \widetilde{v}^{t+1}_{pk} &= W^t_{pk} \cdot v^t_{pk} + C_1 \cdot r_1 \cdot \left(p\mathrm{SNP}^t_{pk} - \mathrm{SNP}^t_{pk}\right) \\ &\quad + C_2 \cdot r_2 \cdot \left(g\mathrm{SNP}^t_{k} - \mathrm{SNP}^t_{pk}\right), \\ v^{t+1}_{pk} &= \begin{cases} \widetilde{v}^{t+1}_{pk} & \widetilde{v}^{t+1}_{pk} \in [1-M, M-1] \\ \mathrm{rand}\,(1-M, M-1) & \widetilde{v}^{t+1}_{pk} \notin [1-M, M-1], \end{cases} \end{aligned} \tag{3}$$

where C_1 and C_2, controlling how far a particle moves in a single iteration, are acceleration factors and r_1 and r_2 are random values in $(0, 1)$. To obtain a valid velocity, a random value is sampled in $[1 - M, M - 1]$ while $\widetilde{v}^{t+1}_{pk}$ exceeds its domain. Based on v^{t+1}_{pk}, the position of SNP^t_{pk} can be updated by the following two equations:

$$\begin{aligned} \overline{\mathrm{SNP}}^{t+1}_{pk} &= \mathrm{SNP}^t_{pk} + v^{t+1}_{pk}, \\ \mathrm{SNP}^{t+1}_{pk} &= \begin{cases} \mathrm{int}\left(\overline{\mathrm{SNP}}^{t+1}_{pk}\right) & \overline{\mathrm{SNP}}^{t+1}_{pk} \in [1, M] \\ \mathrm{int}\,(\mathrm{rand}\,(1, M)) & \overline{\mathrm{SNP}}^{t+1}_{pk} \notin [1, M]. \end{cases} \end{aligned} \tag{4}$$

Because of SNP_{pk}^{t+1} being a SNP index, an integer between 1 and M is randomly sampled if $\overline{\mathrm{SNP}}_{pk}^{t+1}$ exceeds its domain. Such random sampling strategies on updating of both v_{pk}^{t+1} and SNP_{pk}^{t+1} help to increase the diversity of the search, the more possibility of jumping out local optima and getting into global optima.

(5) Updating Particle Experiences and Common Knowledge of Swarm. Another strategy introduced to IOBLPSO is the OBL. The basic principle of OBL is the consideration of a solution and its corresponding opposite solution simultaneously to approximate the global optima [33]. In the IOBLPSO, if the solution is $\mathrm{Position}_t(p)$, its corresponding opposite solution can be defined as

$$\mathrm{Position}'_t(p) = 1 + M - \mathrm{Position}_t(p). \tag{5}$$

By comparing fitness values of $\mathrm{Position}_t(p)$, $\mathrm{Position}'_t(p)$, and $pbest_t(p)$, the individual experience of the p_{th} particle at iteration $t+1$, that is, $pbest_{t+1}(p)$, is updated to the one among them with highest fitness value, which can be written as

$$pbest_{t+1}(p) = \begin{cases} \mathrm{Position}_t(p) & MI(\mathrm{Position}_t(p); Y) = \mathrm{Val} \\ \mathrm{Position}'_t(p) & MI(\mathrm{Position}'_t(p); Y) = \mathrm{Val} \\ pbest_t(p) & MI(pbest_t(p); Y) = \mathrm{Val}, \end{cases} \tag{6}$$

where $\mathrm{Val} = \max(MI(\mathrm{Position}_t(p); Y), MI(\mathrm{Position}'_t(p); Y), \text{and } MI(pbest_t(p); Y))$. From this equation, it can be seen that the employed OBL strategy facilitates IOBLPSO not only expanding the search space and enhancing the global explorative ability, but also accelerating the convergence and avoiding premature convergence.

Similarly, whether the common knowledge of the swarm at iteration $t+1$, for example, $gbest_{t+1}$, is updated or maintained as $gbest_t$ depends on fitness values of individual experiences of all particles at iteration $t+1$ and can be defined as

$$gbest_{t+1} = \begin{cases} pbest_{t+1}(p) & MI(pbest_{t+1}(p); Y) > MI(gbest_t; Y) \\ gbest_t & MI(pbest_{t+1}(p); Y) \le MI(gbest_t; Y). \end{cases} \tag{7}$$

(6) Deep Searching with a Postprocedure. A postprocedure is provided when completing the iteration process to carry out a deep search of SNP-SNP interactions in a highly suspected SNP sets. First, all SNPs are descending sorted according to their counters in count_T, and the specified number of top SNPs (By default, 10) are selected into the highly suspected SNP sets. Second, IOBLPSO conducts an exhaustive search within the highly suspected SNP sets to determine whether fitness value of one or more SNP combinations is higher than that of $gbest_T$. If indeed detected, $gbest_T$ is updated by the best one among them. $gbest_T$ is therefore the final result of IOBLPSO.

3. Results and Discussion

3.1. Simulation Data. Six commonly used models of SNP-SNP interactions with their orders being equal to 2 (i.e., $K = 2$) are exemplified for the study [7, 9, 10, 34, 35]. Model 1 and Model 2 are models displaying both marginal effects and interactive effects, and others show no marginal effects but interactive effects. Specifically, the penetrance in Model 1 increases only when both SNPs have at least one minor allele [9, 10]; Model 2 assumes that the minor allele in one SNP has the marginal effect; however the effect is inversed while minor alleles in both SNPs are present [9]; Model 3 and Model 4 are directly cited from the reference [35]; Model 5 is a ZZ model [34]; and Model 6 is an XOR model [35]. Model 3~Model 6 are exemplified here since they provide a high degree of complexity to challenge ability of a method in detecting SNP-SNP interactions [7]. For each model, 50 data sets are generated by the simulator EpiSIM [36], each containing 2000 cases and 2000 controls genotyped with 100 SNPs. For each data set, random SNPs are set independently with MAFs chosen from $[0.05, 0.5]$ uniformly and detailed parameters of ground-truth SNPs are recorded in Figure 2, where ground-truth SNPs refer to the causative SNPs that truly associated with the phenotype, in other words, the SNPs in models added into the simulation data sets.

3.2. Evaluation Measure. Detection power is one of the generally used evaluation measures in the field of the detection of SNP-SNP interactions, and various forms of detection power have been proposed depending on what is desired to measure [4, 7, 9–11, 21, 31, 37]. In this study, two types of detection power are introduced, namely, Power 1 and Power 2.

Power 1 [4, 7, 9–11, 21, 31] is defined as the proportion of data sets in which all ground-truth SNPs are detected with no false positives, which can be written as

$$\mathrm{Power\ 1} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{8}$$

where N is the number of data sets with the same parameter settings (here, $N = 50$), and $x_i \in \{0, 1\}$ is the detection tag; that is, if 2 ground-truth SNPs in data set i are detected with no false positives, $x_i = 1$; otherwise, $x_i = 0$. Though Power 1 seems not practical since false positives are inevitable for any statistical tests and fewer false positives result in larger false negatives, we still introduce it because it is advantageous in practical applications and might be of interest to biologists due to false positives implying wasted experimental effort to validate the results.

Sometimes, allowing some small Type-I error rate is more reasonable; thus Power 2 [4, 7, 21] is introduced here, which is defined as an average proportion of ground-truth SNPs in the top 2 detected SNPs, and can be written as

$$\mathrm{Power\ 2} = \frac{1}{2 \cdot N}\sum_{i=1}^{N} y_i, \tag{9}$$

where y_i is the number of ground-truth SNPs in the top 2 SNPs identified in data set i.

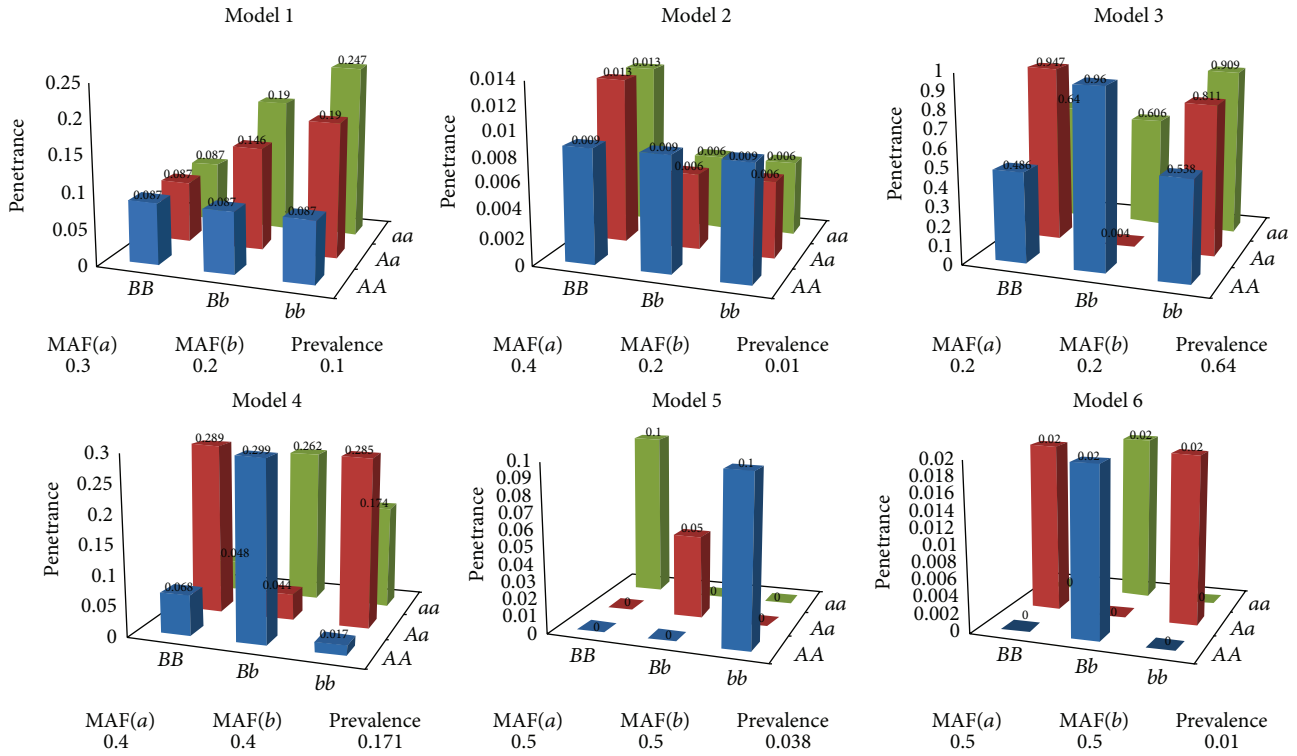


FIGURE 2: The simulation models of SNP-SNP interactions. In the figure, penetrance is the probability of the occurrence of a disease given a particular genotype; prevalence is the proportion of individuals that occur a disease; MAF(a) and MAF(b) are, respectively, minor allele frequencies of a and b.

Computational complexity is also considered. We measure running time in the same computational environment to assess realistic applicability of compared methods.

3.3. Performance of IOBLPSO on Simulation Data. To demonstrate the validity of IOBLPSO, its detection power is evaluated by comparison with several typical SNP-SNP interaction detection methods, that is, BOOST [12], AntEpiSeeker [20], SNPRuler [11], and TEAM [38]. These machine learning methods are recently proposed, claimed to facilitate large scale data sets, and their packages are online freely available [7]. Besides these methods, two modified PSO methods for SNP-SNP interaction detection, namely, DBM-PSO [27] focusing on finding the best genotype-genotype of a SNP-SNP interaction among possible genotypes of SNP combinations and the PSO focusing on finding SNP-SNP interactions among possible SNP combinations, are also compared.

In the study, parameters of each method are generally set as default. Only a few are changed according to suggestions in order to balance result accuracy and computational cost. For BOOST, interaction threshold is set to 10, that is, results of BOOST are the SNP-SNP interactions whose likelihood ratio test statistic values >10 with 4 degrees of freedom. For AntEpiSeeker, the numbers of ants and iterations is set to 500 and 10, respectively. For TEAM, permutation number is set to 100. For a fair comparison, parameter settings of PSO based methods are the same. Specifically, the number of particles P and the number of iterations T are respective set to 100 and 100; both acceleration factors C_1 and C_2 are set to 2 [39]; the inertia weight W is set to 0.65. It is believed that performance of IOBLPSO mainly depends on parameters (P, T). Hence we further examine the influence of these parameters on detection power with (25, 100), (50, 100), (100, 25), (100, 50), and (100, 100).

Detection power of compared methods on simulation data sets is reported in Figure 3. Detection power of IOBLPSO and the PSO with different numbers of particles is shown in Figure 4, and that with different numbers of iterations is shown in Figure 5. The average running time of the methods on simulation data sets is recorded in Table 1. From Figures 3, 4, and 5 and Table 1, we have the following observations.

It is seen that IOBLPSO outperforms compared methods on all cases regardless of models, the numbers of particles, and iterations. Specifically, no matter, according to Power 1 or Power 2, detection power of IOBLPSO on all models and (P, T) settings is comparable and sometimes superior to that of compared methods, which might be the result of introducing three effective strategies into IOBLPSO: OBL expanding the search space and enhancing the global explorative ability, dynamic inertia weight guiding the particles to more promising regions, and postprocedure carrying out a deep search in highly suspected SNP sets; with the numbers of particles or iterations grow, detection power of both IOBLPSO and the PSO increase quickly, especially IOBLPSO; IOBLPSO identifies almost all ground-truth SNPs

Table 1: Average running time (seconds) of compared methods on simulation data sets. Experiments are conducted with Intel Xeon 2.00 GHz CPUs and 6 GB of RAM running Microsoft Windows XP Professional x64 Edition 2003 Service Pack 2 for computational complexity analysis.

Methods	BOOST	AntEpiSeeker	SNPRuler	TEAM	DBM-PSO	ISO	IOBLPSO
Running time	0.36	1146.60	1.56	13.14	68.73	13.40	20.75

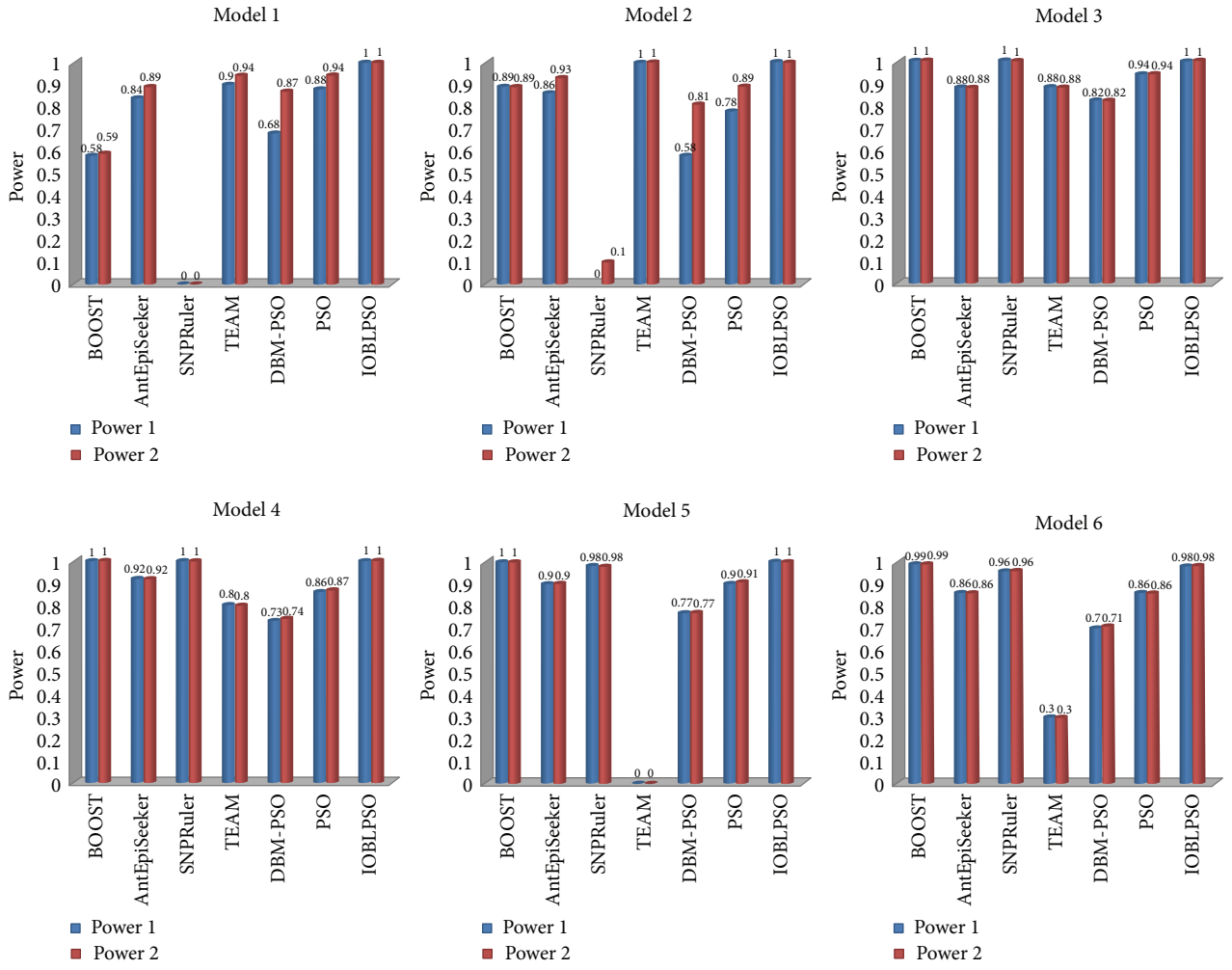


Figure 3: Detection power of compared methods on simulation data sets.

on all models with the parameter setting (100, 100), even with (25, 100) or (100, 25); IOBLPSO has perfect detection power on Model 1 and Model 2; that is to say, compared with other PSO based methods, IOBLPSO needs less particles and/or iterations to obtain higher detection power, implying that IOBLPSO can handle large scale data sets for GWAS and its scalability is better than others; for Model 1 and Model 2, Power 1 and Power 2 of IOBLPSO reach a prefect level, Power 1 and Power 2 of other methods have different values since these two models display not only interaction effects but also marginal effects, leading to compared methods sometimes only identifying several ground-truth SNPs, but not SNP-SNP interactions; similarly, for each method on Model 3~ Model 6, Power 1 and Power 2 of each compared method are almost always equal because single ground-truth SNPs show no main effects; in terms of computational complexity, though IOBLPSO is not the fast one among all compared methods, it can finish the work at affordable time costs; more importantly, its time costs can be estimated and controlled by setting the numbers of particles and iterations freely under the premise of ensuring sufficient accuracy.

3.4. Application to Real AMD Data. In the study, potential of IOBLPSO can also be verified by analyzing a real AMD data set [40], which contains 103.611 SNPs genotyped with 96 cases and 50 controls. AMD, which refers to pathological changes in the central area of the retina, is the most important cause of irreversible visual loss in elderly populations and is considered as a complex disease whereby multiple SNP-SNP interactions interact with environmental factors to the disease [4, 10]. We run IOBLPSO on AMD data set 20 times with different combinations of the number of particles

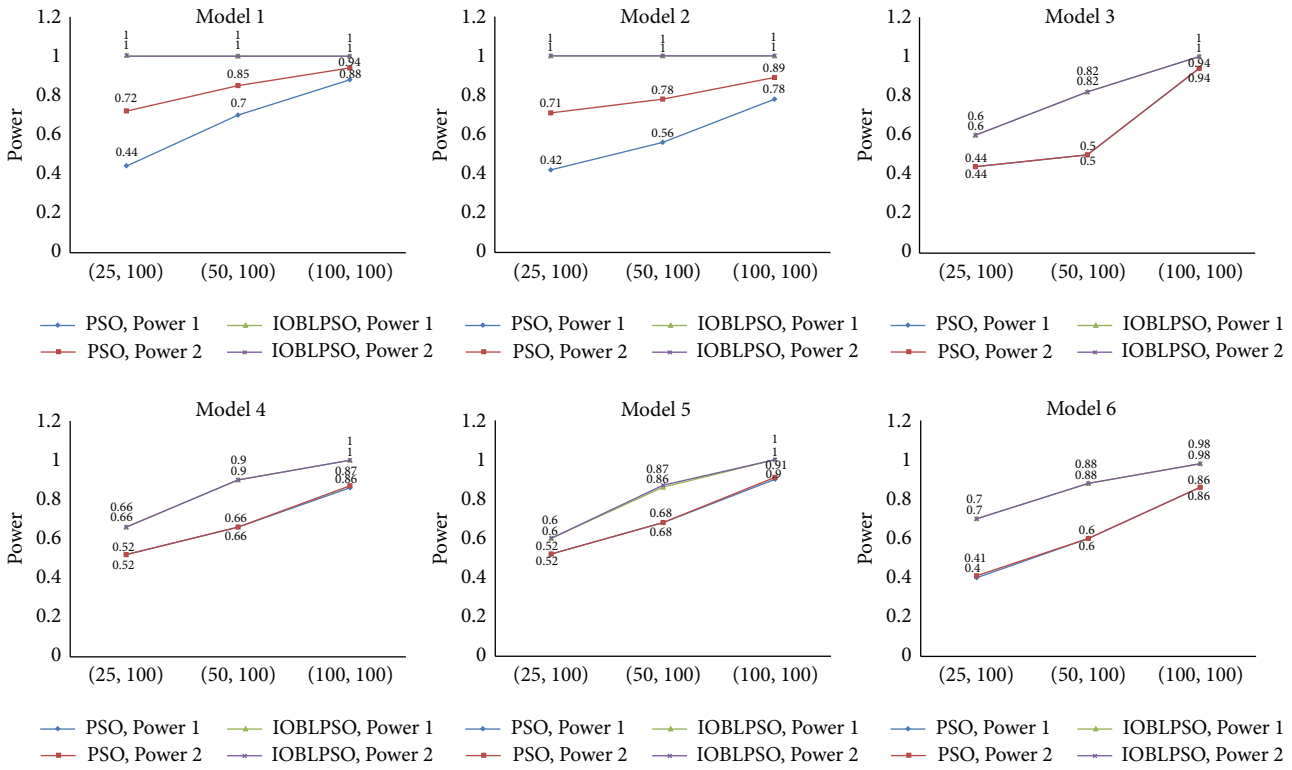


Figure 4: Detection power of IOBLPSO and the PSO with different numbers of particles. The numbers of particles are set to 25, 50, and 100, while the number of iterations is equal to 100.

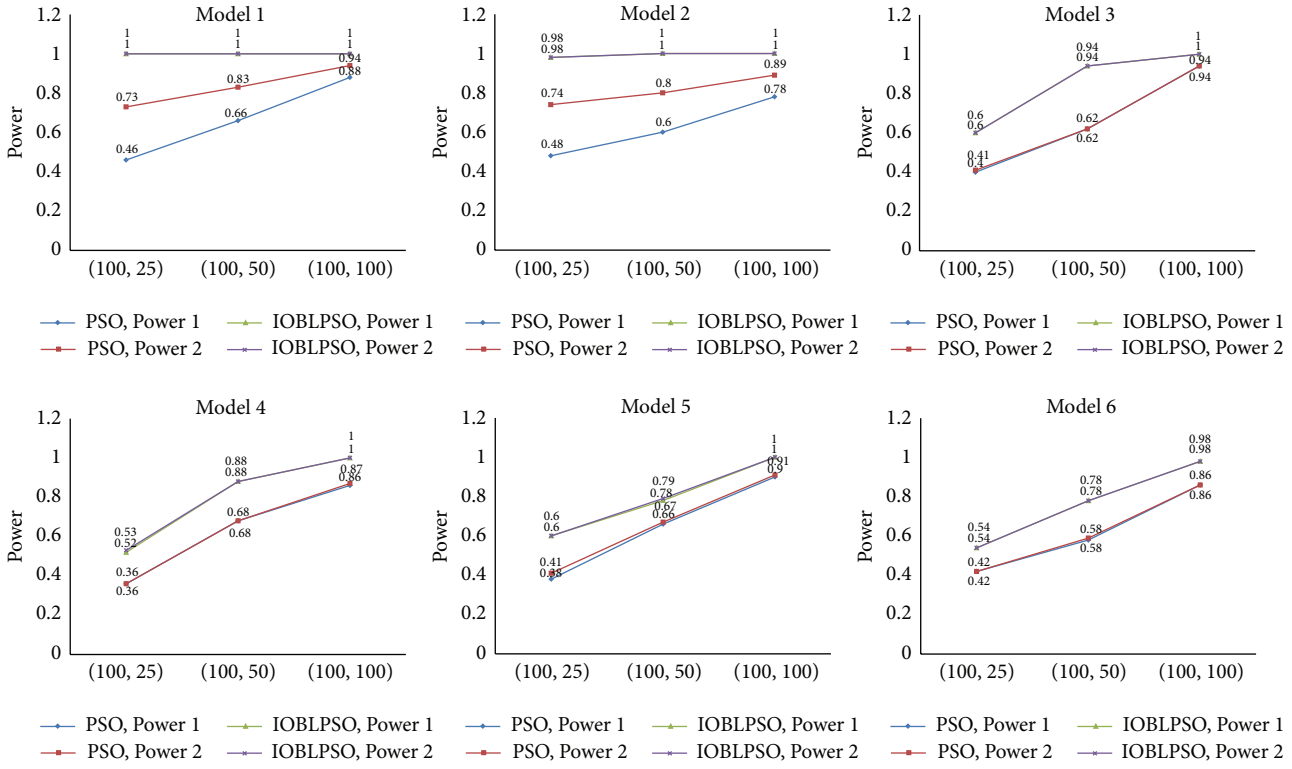


Figure 5: Detection power of IOBLPSO and the PSO with different numbers of iterations. The numbers of iterations are set to 25, 50, and 100, while the number of particles is equal to 100.

Table 2: Detected SNP-SNP interactions associated with AMD. *P* values of detected SNP-SNP interactions before Bonferroni correction, as well as their linkage disequilibrium (LD) correlation coefficients r^2, are also recorded. The SNPs in different SNP-SNP interactions have low LD.

SNP	Gene	Chromosome	Times	Mutual information value		*P* value	LD (r^2)
				Individual	Interaction		
rs380390	*CFH*	1	1	**0.1412**	**0.2955**	5.0006$e-08$	0.0089
rs1374431	*N/A*	2		0.0198			
rs380390	*CFH*	1	2	**0.1412**	0.2949	5.7995$e-08$	0.0015
rs2402053	*N/A*	7		0.0476			
rs1329428	*CFH*	1	2	0.1218	0.2853	3.0012$e-08$	0.0019
rs9328536	*MED27*	9		0.0563			
rs380390	*CFH*	1	2	**0.1412**	0.2777	1.4359$e-07$	0.0050
rs10512174	*ISCA1*	9		0.0844			
rs380390	*CFH*	1	2	**0.1412**	0.2775	1.9451$e-07$	0.0018
rs718263	*NCALD*	8		0.0471			
rs380390	*CFH*	1	1	**0.1412**	0.2760	1.9798$e-07$	0.0019
rs223607	*N/A*	6		0.0079			
rs380390	*CFH*	1	**8**	**0.1412**	0.2752	**1.1381e−09**	0.0031
rs1363688	*N/A*	5		0.0949			
rs380390	*CFH*	1	1	**0.1412**	0.2678	6.4999$e-07$	0.0013
rs210758	*N/A*	4		0.0131			
rs380390	*CFH*	1	1	**0.1412**	0.2641	2.4273$e-07$	0.0040
rs10507949	*N/A*	13		0.0948			

P (10000, 20000) and the number of iterations T (500, 1000), each running 5 times. The order of SNP-SNP interactions K is set to 2 since the small sample size of 146 individuals is insufficient for secure detection of any higher order SNP-SNP interactions. Both acceleration factors C_1 and C_2 are set to 2. Detected SNP-SNP interactions associated with AMD are listed in Table 2, where their mutual information values of individual SNPs and SNP-SNP interactions are recorded.

It has been widely accepted that rs380390 and rs1329428 are believed to be significantly associated with AMD [10]. These two SNPs are in an intron of the *CFH* gene in chromosome 1. There are biologically plausible mechanisms for the involvement of *CFH* in AMD and at least 100 mutations in *CFH* have been proven to increase the risk of AMD and other disorders. *CFH* is a regulator that activates the alternative pathway of the complement cascade, the mutations in which can lead to an imbalance in normal homeostasis of the complement system. This phenomenon is thought to account for substantial tissue damage in AMD [41]. In the IOBLPSO, these two SNPs are detected as members of SNP-SNP interactions, especially the rs380390. Almost all SNP-SNP interactions include rs380390, since it has the strongest main effect, leading to its combinations with other SNPs displaying strong interaction effects. This phenomenon indicates that IOBLPSO is sensitive to those SNPs displaying strong main effects.

The SNP-SNP interaction (rs380390, rs1374431), also reported by [42, 43], has the strongest interaction effect. Rs1374431 is located in a noncoding region between genes *LOC644301* and *KIAA1715*. *KIAA1715* is usually found in adult brain regions. Although no evidences were reported with this gene related to AMD, it may be a plausible candidate gene associated with AMD [42, 43]. Another SNP-SNP interaction (rs380390, rs2402053) has the second highest mutual information value. The SNP rs2402053 is in the intergenic region between genes *TFEC* and *TES* in chromosome 7q31 [44]. It is worth noting that mutations in some genes on 7q31-7q32 are revealed in patients with retinal disorders [45]. Therefore, rs2402053 may be a new genetic factor contributing to the underlying mechanism of AMD [46–50].

It is interesting that the SNP-SNP interaction (rs380390, rs1363688) was successfully detected 8 times by the IOBLPSO and by other methods [4, 21, 51], though it has moderate interaction effect. However, in terms of *P* value, the interaction (rs380390, rs1363688) is the most statistically significant one among all detected SNP-SNP interactions, which might be the reason of it being frequently detected. This fact implies that IOBLPSO is capable of capturing SNP-SNP interactions with statistically significant *P* values, though its fitness function is the mutual information. The SNP-SNP interaction (rs1329428, rs9328536) [52, 53], rs10507949 [4], and rs10512174 [51] also have been identified in AMD association studies, but their functions are still unclear. Other SNPs, that is, rs210758, rs223607, and rs718263, are the first time being identified. Further studies with the use of large-scale case-control samples are needed to confirm whether these SNPs have true associations with AMD. We hope that, from these results, some clues could be provided for the exploration of causative factors of AMD.

4. Conclusions

Detection of SNP-SNP interactions is believed to be important in understanding underlying mechanism of complex

diseases. In this study, we proposed an improved opposition-based learning particle swarm optimization, or IOBLPSO, to detect SNP-SNP interactions. To the best of our knowledge, IOBLPSO is the first PSO based method to detect SNP-SNP interactions among possible SNP combinations. Highlights of IOBLPSO are the introduction of three strategies: OBL, dynamic inertia weight, and a postprocedure. Among them, OBL is the core, which is presented in the stage of updating particle experiences and common knowledge of swarm, not only for enhancing the global explorative ability, but also for avoiding premature convergence. Dynamic inertia weight is computed before the stage of updating particle velocities to allow particles to cover a wider search space while the considered SNP is likely to be a random SNP and to converge on promising regions of the search space while capturing a highly suspected SNP. The postprocedure is introduced as the final stage for carrying out a deep search in highly suspected SNP sets. Experiments of IOBLPSO are performed on lots of simulation data sets under the evaluation measures of detection power and computational complexity. Results demonstrate that IOBLPSO is promising for the detection of all simulation models of SNP-SNP interactions. IOBLPSO is also applied on a real AMD data set, results of which not only show the strength of IOBLPSO on real applications, but also capture important features of genetic architecture of AMD that have not been described previously. These features might provide new clues for biologists on the exploration of AMD associated genetic factors.

IOBLPSO might be an alternative to existing methods for detecting SNP-SNP interactions and has several merits. First, IOBLPSO is easy to be implemented, and its time costs can be estimated and controlled. Second, OBL and other two strategies help to improve the performance of IOBLPSO. Third, mutual information is effective in measuring SNP-SNP interactions. Fourth, compared with other methods, IOBLPSO needs less particles and/or iterations to obtain higher detection power, implying that IOBLPSO can handle large scale data sets for GWAS and its scalability is better than others. Though IOBLPSO is a beneficial exploration in the detection of SNP-SNP interactions, it still has several limitations; for example, multiple SNP-SNP interactions in a data set are not considered simultaneously; IOBLPSO is sensitive to those SNPs that display strong main effects. Furthermore, recent advancements in sequencing technology have enabled the sequencing of the whole-exome or even whole-genome of a cohort. The rare or de novo mutations resulting from these experiments should be considered. For example, Wu et al. recently proposed a bioinformatics method called SPRING for prioritizing candidate mutations [54]. It is therefore interesting to consider the problem of interactive effects of such de novo mutations. Limitations of IOBLPSO, as well as this new research hotspot, will inspire us to continue working in the future.

Conflict of Interests

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

Acknowledgments

This work was supported by the Scientific Research Reward Foundation for Excellent Young and Middle-age Scientists of Shandong Province (BS2014DX004), the Science and Technology Planning Project of Qufu Normal University (xkj201410), the Scientific Research Foundation of Qufu Normal University (BSQD20130119), the Shandong Provincial Natural Science Foundation (ZR2013FL016), the China Postdoctoral Science Foundation Funded Project (2014M560264), the Shenzhen Municipal Science and Technology Innovation Council (JCYJ20140417172417174), the Award Foundation Project of Excellent Young Scientists in Shandong Province (BS2014DX005), the Project of Shandong Province Higher Educational Science and Technology Program (J13LN31), the Scientific Research Foundation of Qufu Normal University (XJ201226), and the Innovation and Entrepreneurship Training Project for College Students of Qufu Normal University (2014A096).

References

[1] L. R. Cardon and J. I. Bell, "Association study designs for complex diseases," *Nature Reviews Genetics*, vol. 2, no. 2, pp. 91–99, 2001.

[2] N. Risch and K. Merikangas, "The future of genetic studies of complex human diseases," *Science*, vol. 273, no. 5281, pp. 1516–1517, 1996.

[3] B. Maher, "The case of the missing heritability," *Nature*, vol. 456, no. 7218, pp. 18–21, 2008.

[4] J. Shang, J. Zhang, Y. Sun, X. Dai, and Y. Zhang, "EpiMiner: a three-stage co-information based method for detecting and visualizing epistatic interactions," *Digital Signal Processing*, vol. 24, pp. 1–13, 2014.

[5] M. D. Ritchie, L. W. Hahn, N. Roodi et al., "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *The American Journal of Human Genetics*, vol. 69, no. 1, pp. 138–147, 2001.

[6] 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.

[7] J. Shang, J. Zhang, Y. Sun, D. Liu, D. Ye, and Y. Yin, "Performance analysis of novel methods for detecting epistasis," *BMC Bioinformatics*, vol. 12, article 475, 2011.

[8] R. Jiang, W. Tang, X. Wu, and W. Fu, "A random forest approach to the detection of epistatic interactions in case-control studies," *BMC Bioinformatics*, vol. 10, no. 1, article S65, 2009.

[9] Y. Zhang and J. S. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nature Genetics*, vol. 39, no. 9, pp. 1167–1173, 2007.

[10] W. Tang, X. Wu, R. Jiang, and Y. Li, "Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy," *PLoS Genetics*, vol. 5, no. 5, Article ID e1000464, 2009.

[11] X. Wan, C. Yang, Q. Yang, H. Xue, N. L. S. Tang, and W. Yu, "Predictive rule inference for epistatic interaction detection in genome-wide association studies," *Bioinformatics*, vol. 26, no. 1, pp. 30–37, 2010.

[12] X. Wan, C. Yang, Q. Yang et al., "BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control

studies," *The American Journal of Human Genetics*, vol. 87, no. 3, pp. 325–340, 2010.

[13] P. Chanda, L. Sucheston, A. Zhang, and M. Ramanathan, "The interaction index, a novel information-theoretic metric for prioritizing interacting genetic variations and environmental factors," *European Journal of Human Genetics*, vol. 17, no. 10, pp. 1274–1286, 2009.

[14] P. Chanda, L. Sucheston, A. Zhang et al., "AMBIENCE: a novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes," *Genetics*, vol. 180, no. 2, pp. 1191–1210, 2008.

[15] P. Chanda, A. Zhang, D. Brazeau et al., "Information-theoretic metrics for visualizing gene-environment interactions," *The American Journal of Human Genetics*, vol. 81, no. 5, pp. 939–963, 2007.

[16] P. Chanda, L. Sucheston, S. Liu, A. Zhang, and M. Ramanathan, "Information-theoretic gene-gene and gene-environment interaction analysis of quantitative traits," *BMC Genomics*, vol. 10, article 509, 2009.

[17] C. S. Greene, J. M. Gilmore, J. Kiralis, P. C. Andrews, and J. H. Moore, "Optimal use of expert knowledge in ant colony optimization for the analysis of epistasis in human disease," in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pp. 92–103, Springer, Berlin, Germany, 2009.

[18] C. S. Greene, B. C. White, and J. H. Moore, "Ant colony optimization for genome-wide genetic analysis," in *Ant Colony Optimization and Swarm Intelligence*, pp. 37–47, Springer, Berlin, Germany, 2008.

[19] R. Rekaya and K. Robbins, "Ant colony algorithm for analysis of gene interaction in high-dimensional association data," *Revista Brasileira de Zootecnia*, vol. 38, no. 1, pp. 93–97, 2009.

[20] Y. Wang, X. Liu, K. Robbins, and R. Rekaya, "AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm," *BMC Research Notes*, vol. 3, article 117, 2010.

[21] J. Shang, J. Zhang, X. Lei, Y. Zhang, and B. Chen, "Incorporating heuristic information into ant colony optimization for epistasis detection," *Genes and Genomics*, vol. 34, no. 3, pp. 321–327, 2012.

[22] C.-H. Yang, H.-W. Chang, Y.-H. Cheng, and L.-Y. Chuang, "Novel generating protective single nucleotide polymorphism barcode for breast cancer using particle swarm optimization," *Cancer Epidemiology*, vol. 33, no. 2, pp. 147–154, 2009.

[23] H.-W. Chang, C.-H. Yang, C.-H. Ho, C.-H. Wen, and L.-Y. Chuang, "Generating SNP barcode to evaluate SNP-SNP interaction of disease by particle swarm optimization," *Computational Biology and Chemistry*, vol. 33, no. 1, pp. 114–119, 2009.

[24] 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, no. 4, pp. 336–342, 2012.

[25] 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, no. 5, Article ID e37018, 2012.

[26] L.-Y. Chuang, Y.-D. Lin, H.-W. Chang, and C.-H. Yang, "SNP-SNP interaction using Gauss chaotic map particle swarm optimization to detect susceptibility to breast cancer," in *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS '14)*, pp. 2548–2554, Waikoloa, Hawaii, USA, January 2014.

[27] 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, no. 1, pp. 260–279, 2012.

[28] C.-H. Yang, Y.-D. Lin, L.-Y. Chuang, and H.-W. Chang, "Double-bottom chaotic map particle swarm optimization based on chi-square test to determine gene-gene interactions," *BioMed Research International*, vol. 2014, Article ID 172049, 10 pages, 2014.

[29] M.-L. Hwang, Y.-D. Lin, L.-Y. Chuang, and C.-H. Yang, "Determination of the SNP-SNP interaction between breast cancer related genes to analyze the disease susceptibility," *International Journal of Machine Learning and Computing*, vol. 4, no. 5, pp. 468–473, 2014.

[30] S.-J. Wu, L.-Y. Chuang, Y.-D. Lin et al., "Particle swarm optimization algorithm for analyzing SNP-SNP interaction of renin-angiotensin system genes against hypertension," *Molecular Biology Reports*, vol. 40, no. 7, pp. 4227–4233, 2013.

[31] M. Aflakparast, H. Salimi, A. Gerami, M.-P. Dubé, S. Visweswaran, and A. Masoudi-Nejad, "Cuckoo search epistasis: a new method for exploring significant genetic interactions," *Heredity*, vol. 112, no. 6, pp. 666–674, 2014.

[32] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, pp. 1942–1948, December 1995.

[33] H. R. Tizhoosh, "Opposition-based learning: a new scheme for machine intelligence," in *International Conference on Computational Intelligence for Modelling, Control and Automation, International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA/IAWTIC '05)*, vol. 1, pp. 695–701, Vienna, Austria, 2005.

[34] W. N. Frankel and N. J. Schork, "Who's afraid of epistasis?" *Nature genetics*, vol. 14, no. 4, pp. 371–373, 1996.

[35] W. Li and J. Reich, "A complete enumeration and classification of two-locus disease models," *Human Heredity*, vol. 50, no. 6, pp. 334–349, 2000.

[36] J. Shang, J. Zhang, X. Lei, W. Zhao, and Y. Dong, "EpiSIM: simulation of multiple epistasis, linkage disequilibrium patterns and haplotype blocks for genome-wide interaction analysis," *Genes & Genomics*, vol. 35, no. 3, pp. 305–316, 2013.

[37] X. Jiang, R. E. Neapolitan, M. M. Barmada, S. Visweswaran, and G. F. Cooper, "A fast algorithm for learning epistatic genomic relationships," in *Proceedings of the AMIA Annual Symposium*, p. 341, 2010.

[38] X. Zhang, S. Huang, F. Zou, and W. Wang, "TEAM: efficient two-locus epistasis tests in human genome-wide association study," *Bioinformatics*, vol. 26, no. 12, pp. i217–i227, 2010.

[39] 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.

[40] R. J. Klein, C. Zeiss, E. Y. Chew et al., "Complement factor H polymorphism in age-related macular degeneration," *Science*, vol. 308, no. 5720, pp. 385–389, 2005.

[41] M. K. M. Adams, J. A. Simpson, A. J. Richardson et al., "Can genetic associations change with age? CFH and age-related macular degeneration," *Human Molecular Genetics*, vol. 21, no. 23, pp. 5229–5236, 2012.

[42] B. Han, M. Park, and X.-W. Chen, "DASSO-MB: detection of epistatic interactions in genome-wide association studies using Markov blankets," in *Proceedings of the IEEE International*

Conference on Bioinformatics and Biomedicine (BIBM '09), pp. 148–153, November 2009.

[43] B. Han, M. Park, and X.-W. Chen, "A Markov blanket-based method for detecting causal SNPs in GWAS," *BMC Bioinformatics*, vol. 11, no. 3, article S5, 2010.

[44] E. S. Tobias, A. F. L. Hurlstone, E. MacKenzie, R. Mcfarlane, and D. M. Black, "The *TES* gene at 7q31.1 is methylated in tumours and encodes a novel growth-suppressing LIM domain protein," *Oncogene*, vol. 20, no. 22, pp. 2844–2853, 2001.

[45] S. J. Bowne, L. S. Sullivan, S. H. Blanton et al., "Mutations in the inosine monophosphate dehydrogenase 1 gene (IMPDH1) cause the RP10 form of autosomal dominant retinitis pigmentosa," *Human Molecular Genetics*, vol. 11, no. 5, pp. 559–568, 2002.

[46] B. Han, X.-W. Chen, Z. Talebizadeh, and H. Xu, "Genetic studies of complex human diseases: characterizing SNP-disease associations using Bayesian networks," *BMC Systems Biology*, vol. 6, no. 3, article S14, 2012.

[47] B. Han, X.-W. Chen, and Z. Talebizadeh, "FEPI-MB: identifying SNPs-disease association using a Markov Blanket-based approach," *BMC Bioinformatics*, vol. 12, p. S3, 2011.

[48] B. Han, X.-W. Chen, and Z. Talebizadeh, "A fast markov blankets method for epistatic interactions detection in genome-wide association studies," in *Proceedings of the 9th International Workshop on Data Mining in Bioinformatics (BIOKDD '10)*, 2010.

[49] S. Lee, M.-S. Kwon, I.-S. Huh, and T. Park, "CUDA-LR: CUDA-accelerated logistic regression analysis tool for gene-gene interaction for genome-wide association study," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW '11)*, pp. 691–695, November 2011.

[50] J. Fontanarosa and Y. Dai, "A block-based evolutionary optimization strategy to investigate gene-gene interactions in genetic association studies," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW '10)*, pp. 330–335, Hong Kong, December 2010.

[51] X. Guo, Y. Meng, N. Yu, and Y. Pan, "Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering," *BMC Bioinformatics*, vol. 15, no. 1, article 102, 2014.

[52] M.-S. Kwon, M. Park, and T. Park, "IGENT: efficient entropy based algorithm for genome-wide gene-gene interaction analysis," *BMC Medical Genomics*, vol. 7, article S6, 2014.

[53] K. Kim, M.-S. Kwon, S. Y. Lee, J. Namkung, M. D. Li, and T. Park, "GxG-Viztool: a program for visualizing gene-gene interactions in genetic association analysis," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW '12)*, pp. 838–843, 2012.

[54] J. Wu, Y. Li, and R. Jiang, "Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies," *PLoS Genetics*, vol. 10, no. 3, Article ID e1004237, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 910515, 5 pages
http://dx.doi.org/10.1155/2015/910515

Research Article

Constraint Programming Based Biomarker Optimization

Manli Zhou,[1,2] Youxi Luo,[1,3] Guoquan Sun,[1] Guoqin Mai,[1] and Fengfeng Zhou[1]

[1]*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China*
[2]*Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Beijing 100049, China*
[3]*School of Science, Hubei University of Technology, Wuhan, Hubei 430068, China*

Correspondence should be addressed to Fengfeng Zhou; fengfengzhou@gmail.com

Received 28 September 2014; Accepted 13 December 2014

Academic Editor: Antonello Merlino

Efficient and intuitive characterization of biological big data is becoming a major challenge for modern bio-OMIC based scientists. Interactive visualization and exploration of big data is proven to be one of the successful solutions. Most of the existing feature selection algorithms do not allow the interactive inputs from users in the optimizing process of feature selection. This study investigates this question as fixing a few user-input features in the finally selected feature subset and formulates these user-input features as constraints for a programming model. The proposed algorithm, fsCoP (feature selection based on constrained programming), performs well similar to or much better than the existing feature selection algorithms, even with the constraints from both literature and the existing algorithms. An fsCoP biomarker may be intriguing for further wet lab validation, since it satisfies both the classification optimization function and the biomedical knowledge. fsCoP may also be used for the interactive exploration of bio-OMIC big data by interactively adding user-defined constraints for modeling.

1. Introduction

Biological big data is being accumulated at an accelerated speed, facilitated by the rapid invention and development of bio-OMIC data production technologies [1, 2]. Interactive exploration technology is widely used to mine knowledge from various big data areas [3] and may be useful to rapidly and accurately detect phenotype-associated biomarkers from the huge amount of bio-OMIC data [4]. This is usually formulated as the feature selection problem [5, 6].

Various algorithms have been proposed to choose a few from a large number of features, by optimizing a phenotypic measurement. The principle of parsimony prefers a minimum number of features for an accurate representation of the data [7]. Detailed introduction may be found for both general feature selection algorithms [8] and phenotype-associated biomarker detection algorithms [9] from the literature. Considering millions or more of bio-OMIC features for each sample, although the exhaustive search guarantees the detection of optimal feature subset, its computational requirement exceeds the capacity of any high-performance computing systems under the current parallel computing architecture. So all the existing feature selection algorithms screen for the suboptimal solutions based on some heuristic rules.

Heuristic feature selection algorithms may be grouped as two classes based on how they generate the finally chosen features. The class I wrapper or group optimization algorithms evaluate a feature subset by testing its classification performance with a learning algorithm. The features are selected by heuristic rules or randomly, and only the feature subset with the best classification performance will be kept for further investigation, for example, forward stepwise selection [10] and ant colony optimization [11]. The class II filtering or individual ranking algorithms measure each feature's correlation with the class labels and rank the features by their measurement. A heuristic assumption is that the combination of top-ranked K features should produce a good classification performance, where K is an arbitrarily chosen integer. They are usually much faster than the class I algorithms but lack model robustness due to the ignorance of feature interdependence [12]. It is also difficult to determine how many features should be chosen from the ordered feature list.

This work proposes a constraint programming based interactive feature selection algorithm, fsCoP, for efficient

exploration of the bio-OMIC big data. An interactive feature selection problem requires a fast and accurate detection of features and the integration of user-input features in the final result. The majority of existing feature selection algorithms do not consider how to make sure a given feature subset appears among the finally selected features. fsCoP fixes the user-input features in the result by formulating them as constraints of the programming model. Our data show that features chosen by fsCoP perform well similar to or much better than the existing feature selection algorithms in classification, even with the constraints of fixed features from both literature and other algorithms.

2. Materials and Methods

2.1. Dataset Downloading and Preprocessing. Two microarray-based gene expression profiling datasets are downloaded from the NCBI GEO database [13]. Both datasets GSE5406 [14] and GSE1869 [15] profiled ischemic cardiomyopathy samples and their controls on the Affymetrix Human Genome U133A Array (HG-U133A) platform. The transcriptomes are normalized using the RMA algorithm [16]. The gene expression profiles of ischemic cardiomyopathy samples and the nonfailing controls are kept for binary classification study in this work.

2.2. Feature Selection Based on Constraint Programming (fsCoP). This work proposes a constraint programming based feature selection algorithm, allowing the user to determine a few features in the finally chosen feature subset. The prefixed features may be the biomarkers known to be associated with the phenotype in the literature or the features selected by other feature selection algorithms. This model is proposed to answer the biological questions like whether a few genes together with the ischemic cardiomyopathy associated ACE2 (angiotensin-converting enzyme-2) may constitute an accurate model for the disease early detection. The majority of the existing feature selection algorithms do not have the integrating component for fixing a few features in the final feature subset. Let FixedSubset be the set of features to be fixed in the final result and let c be the class number. Class j has n_j samples, where $j = 1, 2, \ldots, c$. The programming model is defined as follows:

$$\min_{w_i, \xi_k} \quad \left\{ \sum_{i=1}^{p} w_i + \lambda \left(\sum_{j=1}^{c} \sum_{k=1}^{n_j} \xi_k \right) \right\}, \tag{1}$$

$$\text{s.t.} \quad \sum_{i=1}^{p} \left| s_{ki}^{j} - m_i^{j} \right| w_i - \xi_k < \sum_{i=1}^{p} \left| s_{ki}^{j} - \tilde{m_i}^{j} \right| w_i, \quad \text{for } k \in \left\{1, 2, \ldots, n_j\right\}, \quad j \in \{1, 2, \ldots, c\}, \tag{2}$$

$$w_f \geq \text{MinWeight}, \quad \text{for } f \in \text{FixedSubset}, \tag{3}$$

$$0 \leq w_i \leq 1, \tag{4}$$

$$\xi_k \geq 0. \tag{5}$$

The average value of the ith feature is denoted as m_i^j for the samples in class j. Formula (2) makes that the centroid of class j is the closest centroid to the samples of the class j. Each prefixed feature has the weight no smaller than MinWeight. Each feature has a weight $w_i \in [0, 1]$, where only features with positive weights are selected by the algorithm.

2.3. Classification Performance Measurements. A binary classification model is trained over the datasets of positive and negative samples, whose numbers are P and N, respectively. The classification performance is usually measured by the sensitivity Sn = TP/(TP+FN) and specificity Sp = TN/(TN+FP), where TP, FN, TN, and FP are the numbers of true positives, false negatives, true negatives, and false positives. The overall classification performances may be measured by the overall accuracy Acc = (TP + TN)/(TP + FN + TN + FP) and balanced overall accuracy Avc = (TP + TN)/(TP + FN + TN+FP). Matthew's Correlation Coefficient is also calculated to measure how well a classification model is, and it is defined as MCC = (TP×TN − FP×FN)/sqrt((TP + FP) × (TP + FN) × (TN+FP)×(TN+FN)), where sqrt(x) is the squared root of x.

Fivefold cross validation (5FCV) strategy is used to train the model and calculate how well a model performs. Fluctuation may occur for different seeds of the random number generator. So 30 runs of the 5FCV experiments are carried out with different random seeds.

2.4. Comparison with Four Feature Selection Algorithms. The proposed feature selection algorithm fsCoP is compared with two ranking algorithms, that is, t-test (TRank) [17] and Wilcoxon test (WRank) [18], and two other widely used algorithms, that is, prediction analysis of microarrays (PAM) [19] and regularized random forest (RRF) [20].

The ultimate goal of the proposed model is to select a subset of features with accurate classification performance. The performance of a given feature subset is measured by five widely used classification algorithms, including support vector machine (SVM) [21], Naive Bayesian [22], decision tree (DTree) [23], Lasso [24], and K-nearest neighbor [25]. The classification model with the best Matthew's Correlation Coefficients is kept for the comparison study.

This work uses the default parameters of all the investigated algorithms implemented in the statistical software R/Rstudio version 3.1.1 released on July 10, 2014 [26, 27]. A classification model is usually obtained by trying multiple classification algorithms [28, 29]. So this work compares the feature selection algorithms based on the highest MCC values of the five aforementioned classification algorithms.

3. Results and Discussion

3.1. Constrains from the Literature. The angiotensin-converting enzyme-2 (ACE2) at the location Xp22.2 of the human genome HG19 is chosen to be fixed in the algorithm fsCoP, denoted as fsCoP(ACE2). ACE2 was observed to be differentially expressed between ischemic and nonischemic cardiomyopathy and may play a role in transducing the signal of heart failure pathophysiology [15]. The expression

Table 1: Performance comparison of the algorithm fsCoP. fsCoP has no prefixed features, and the model fsCoP(ACE2) has two predetermined features.

GSE5406					
fsCoP	Sn	Sp	Acc	Avc	MCC
SVM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
NBayes	1.000	0.998	1.000	0.999	0.999
DTree	0.992	0.800	0.967	0.896	0.848
Lasso	0.999	0.900	0.987	0.950	0.939
KNN	1.000	0.871	0.983	0.936	0.923
fsCoP(ACE2)	Sn	Sp	Acc	Avc	MCC
SVM	1.000	0.996	0.999	0.998	0.998
NBayes	**1.000**	**0.998**	**1.000**	**0.999**	**0.999**
DTree	0.993	0.796	0.967	0.894	0.847
Lasso	1.000	0.907	0.988	0.953	0.944
KNN	0.999	0.860	0.982	0.930	0.916
GSE1869					
fsCoP	Sn	Sp	Acc	Avc	MCC
SVM	1.000	0.955	0.983	0.978	0.965
NBayes	1.000	0.972	0.990	0.986	0.979
DTree	0.907	0.000	0.567	0.453	NaN
Lasso	0.960	0.989	0.971	0.974	0.943
KNN	**1.000**	**0.994**	**0.998**	**0.997**	**0.996**
fsCoP(ACE2)	Sn	Sp	Acc	Avc	MCC
SVM	1.000	0.939	0.977	0.970	0.953
NBayes	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
DTree	0.987	0.000	0.617	0.493	NaN
Lasso	0.990	0.967	0.981	0.978	0.962
KNN	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table 2: Running time of fsCoP and fsCoP(ACE2) on GSE5406. All the running times are calculated in seconds and column "repeat" gives the number of repeats of each model with different random seed.

Repeat	fsCoP	Avg (fsCoP)	fsCoP(ACE2)	Avg (fsCoP(ACE2))
5	11.95	2.39	11.78	2.36
10	23.83	2.38	23.96	2.40
50	120.01	2.40	117.79	2.36
100	240.23	2.40	236.75	2.37

level of ACE2 is detected by two probe sets (219962_at and 222257_s_at) in the Affymetrix microarray platform U133A (GPL96). These two features will be fixed in the feature subset fsCoP(ACE2), and the performances of the five classification algorithms are compared using the selected features by fsCoP and fsCoP(ACE2).

Firstly, fsCoP and fsCoP(ACE2) achieve similarly good performance on the two investigated datasets, that is, GSE5406 and GSE1869. Table 1 shows that, except the decision tree algorithm on the dataset GSE1869, there are no greater than 0.021 differences in MCC between the two versions of fsCoP. The greatest difference occurs for the NBayes classification algorithm on the dataset GSE1869, where fsCoP(ACE2) (1.000 in MCC) improves fsCoP (0.979).

Secondly, if only the best classification algorithm is chosen for each subset of selected features, fsCoP(ACE2) also performs well similar with fsCoP. SVM(fsCoP) only improves NBayes(fsCoP(ACE2)) by 0.001 in MCC. The other classification performance measurements also show that this is a minor improvement, with the maximal difference being 0.002 in specificity (Sp). The comparison of the best classification models between the two datasets in Table 1 also shows that NBayes(fsCoP(ACE2)) even performs 0.002 better than KNN(fsCoP) on the dataset GSE1869.

fsCoP runs fast similar with or without fixing a few features. The running time of the algorithm fsCoP with or without fixing user-selected features is compared between fsCoP and fsCoP(ACE2). Since fsCoP runs very fast, we repeat the model testing for multiple times with different random seeds, as in Table 2. The data suggests that fsCoP(ACE2) runs slightly faster than fsCoP for most of the times, except for the case of 10 repeats.

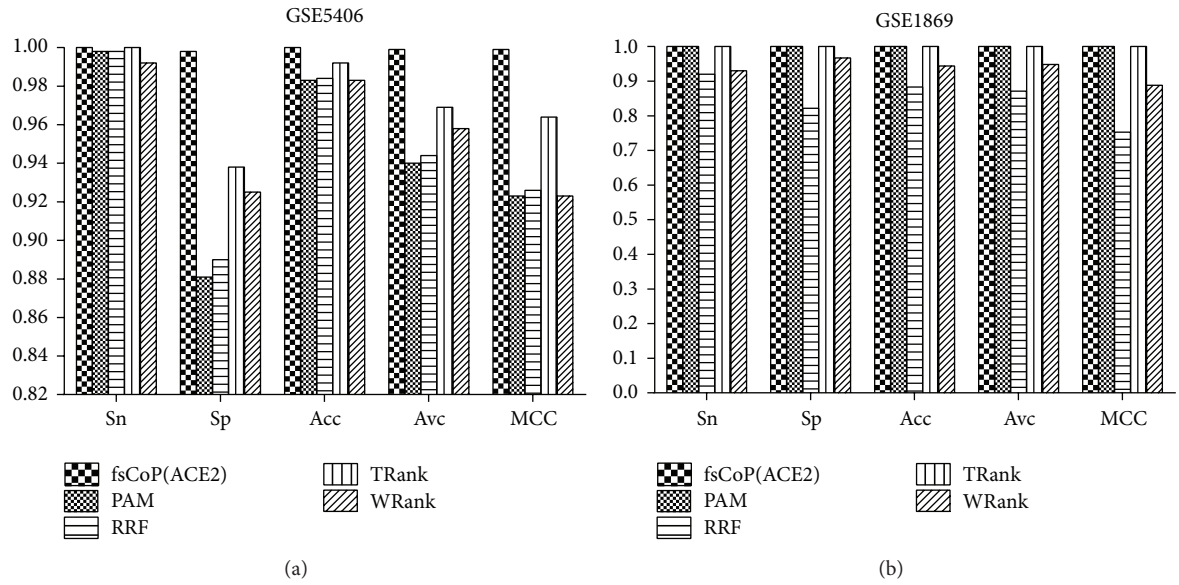


Figure 1: Classification performance comparison of the five feature selection algorithms on the datasets: (a) GSE5406 and (b) GSE1869. The histograms give the detailed values of the classification performance measurements, that is, Sn, Sp, Acc, Avc, and MCC.

3.2. Comparison of fsCoP(ACE2) with the Existing Feature Selection Algorithms. A further comparison of fsCoP(ACE2) with the other existing feature selection algorithms is conducted for the best classification algorithms on each of the selected features, as shown in Figure 1. First of all, fsCoP(ACE2) performs the best (100%) in sensitivity (Sn) with the classification algorithm NBayes on both datasets, as in Figures 1(a) and 1(b). SVM(TRank) achieves the same sensitivities for both datasets, and KNN(PAM) also achieves 100% in Sn on the dataset GSE1869. NBayes(fsCoP(ACE2)) achieves 0.998 in specificity (Sp) on the dataset GSE5406, and no other feature selection algorithms reach the same specificity level. Figure 1(a) suggests that the second best feature selection algorithm may be TRank, which achieves 0.964 in MCC on the dataset GSE5406.

3.3. Constraints from the Existing Feature Selection Algorithms. Except for the features selected by TRank, fsCoP improves all the other three feature selection algorithms. fsCoP(A) is defined to be feature list selected by fsCoP, with the fixed features selected by Algorithm A. Figure 2 shows that fsCoP(TRank) achieves the same classification performance as TRank, and, for the three other feature selection algorithms, fsCoP() achieves higher averaged values and smaller standard deviations for all the five classification performance measurements. The most significant improvement of fsCoP is observed for the RRF algorithm, with 0.0916 in Sp improvement. So besides the integration of known biomarkers from the literature, fsCoP may also be used to further refine the feature subset selected by the existing feature selection algorithms. Better classification performance with smaller fluctuation may be obtained stably by fsCoP, compared with the algorithms.

After the further refining by fsCoP, features selected by all the four feature selection algorithms achieve 100% in the classification sensitivity, while maintaining at least 92% in specificity. And at least 0.95 in MCC is achieved for all the four cases.

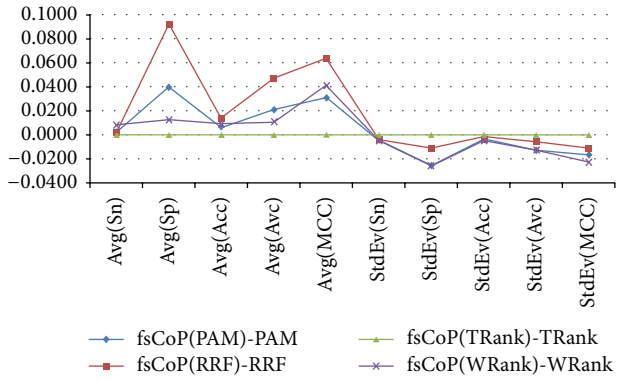


Figure 2: Improvements of fsCoP compared with the four investigated feature selection algorithms, by fixing the features selected by each algorithm. The average "Avg()" and standard deviation "StdEv()" of the five classification performance measurements, that is, Sn, Sp, Acc, Avc, and MCC, are calculated over the 30 runnings of 5-fold cross validations of a given feature subset.

Conflict of Interests

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

Authors' Contribution

Manli Zhou and Youxi Luo contributed equally to this work.

Acknowledgments

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040400), Shenzhen Peacock Plan (KQCX20130628112914301), Shenzhen Research Grants (ZDSY20120617113021359, CXB201104220026A, and JCYJ20130401170306884), the MOE Humanities Social Sciences Fund (no. 13YJC790105) and Doctoral Research Fund of HBUT (no. BSQD13050), and Key Laboratory of Human-Machine-Intelligence Synergic Systems, Chinese Academy of Sciences. Computing resources were partly provided by the Dawning supercomputing clusters at SIAT CAS. The constructive comments from the two anonymous reviewers are greatly appreciated.

References

[1] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, "Ten years of next-generation sequencing technology," *Trends in Genetics*, vol. 30, no. 9, pp. 418–426, 2014.

[2] N. Chen, W. Sun, X. Deng et al., "Quantitative proteome analysis of HCC cell lines with different metastatic potentials by SILAC," *Proteomics*, vol. 8, no. 23-24, pp. 5108–5118, 2008.

[3] J. Heer and S. Kandel, "Interactive analysis of big data," *XRDS: Crossroads*, vol. 19, no. 1, pp. 50–54, 2012.

[4] H. Ge, A. J. M. Walhout, and M. Vidal, "Integrating "omic" information: a bridge between genomics and systems biology," *Trends in Genetics*, vol. 19, no. 10, pp. 551–560, 2003.

[5] S. Datta and V. Pihur, "Feature selection and machine learning with mass spectrometry data," *Methods in Molecular Biology*, vol. 593, pp. 205–229, 2010.

[6] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[7] D. A. Bell and H. Wang, "A formalism for relevance and its application in feature subset selection," *Machine Learning*, vol. 41, no. 2, pp. 175–195, 2000.

[8] I. Iguyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[9] S. Baek, C.-A. Tsai, and J. J. Chen, "Development of biomarker classifiers from high-dimensional data," *Briefings in Bioinformatics*, vol. 10, no. 5, pp. 537–546, 2009.

[10] S. Colak and C. Isik, "Feature subset selection for blood pressure classification using orthogonal forward selection," in *Proceedings of the IEEE 29th Annual Northeast Bioengineering Conference*, pp. 122–123, IEEE, March 2003.

[11] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," in *Feature Extraction, Construction and Selection*, pp. 117–136, Springer, 1998.

[12] I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," *Pattern Recognition*, vol. 43, no. 1, pp. 5–13, 2010.

[13] T. Barrett, S. E. Wilhite, P. Ledoux et al., "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, no. D1, pp. D991–D995, 2013.

[14] S. Hannenhalli, M. E. Putt, J. M. Gilmore et al., "Transcriptional genomics associates FOX transcription factors with human heart failure," *Circulation*, vol. 114, no. 12, pp. 1269–1276, 2006.

[15] M. M. Kittleson, K. M. Minhas, R. A. Irizarry et al., "Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure," *Physiological Genomics*, vol. 21, pp. 299–307, 2005.

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

[17] P. Baldi and A. D. Long, "A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, no. 6, pp. 509–519, 2001.

[18] W.-M. Liu, R. Mei, X. Di et al., "Analysis of high density expression microarrays with signed-rank call algorithms," *Bioinformatics*, vol. 18, no. 12, pp. 1593–1599, 2002.

[19] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6567–6572, 2002.

[20] H. T. Deng and G. Runger, "Feature selection via regularized trees," in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN '12)*, pp. 1–8, Brisbane, Australia, June 2012.

[21] S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, vol. 12, no. 6, pp. 783–789, 1999.

[22] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence: 1995*, pp. 338–345, Morgan Kaufmann, 1995.

[23] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[24] D. Ghosh and A. M. Chinnaiyan, "Classification and selection of biomarkers in genomic data using LASSO," *Journal of Biomedicine and Biotechnology*, vol. 2005, no. 2, pp. 147–154, 2005.

[25] M.-L. Zhang and Z.-H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification," in *Proceedings of the IEEE International Conference on Granular Computing*, vol. 2, pp. 718–721, Beijing, China, July 2005.

[26] R: A Language and Environment for Statistical Computing, http://www.r-project.org/.

[27] J. S. Racine, "RSTUDIO: a platform-independent IDE for R and sweave," *Journal of Applied Econometrics*, vol. 27, no. 1, pp. 167–172, 2012.

[28] F. Zhou and Y. Xu, "cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data," *Bioinformatics*, vol. 26, no. 16, pp. 2051–2052, 2010.

[29] P. Guo, Y. Luo, G. Mai et al., "Gene expression profile based classification models of psoriasis," *Genomics*, vol. 103, no. 1, pp. 48–55, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 168358, 14 pages
http://dx.doi.org/10.1155/2015/168358

Research Article

Evolutionary and Expression Analysis of miR-#-5p and miR-#-3p at the miRNAs/isomiRs Levels

Li Guo,[1] Jiafeng Yu,[2] Hao Yu,[1] Yang Zhao,[1] Shujie Chen,[1] Changqing Xu,[1] and Feng Chen[1]

[1] *Department of Epidemiology and Biostatistics, School of Public Health, Nanjing Medical University, Nanjing 211166, China*
[2] *Shandong Provincial Key Laboratory of Functional Macromolecular Biophysics, Institute of Biophysics, Dezhou University, Dezhou 253023, China*

Correspondence should be addressed to Li Guo; gl8008@163.com

Received 12 June 2014; Revised 27 September 2014; Accepted 29 September 2014

Academic Editor: Yuedong Yang

We mainly discussed miR-#-5p and miR-#-3p under three aspects: (1) primary evolutionary analysis of human miRNAs; (2) evolutionary analysis of miRNAs from different arms across the typical 10 vertebrates; (3) expression pattern analysis of miRNAs at the miRNA/isomiR levels using public small RNA sequencing datasets. We found that no bias can be detected between the numbers of 5p-miRNA and 3p-miRNA, while miRNAs from miR-#-5p and miR-#-3p show variable nucleotide compositions. IsomiR expression profiles from the two arms are always stable, but isomiR expressions in diseased samples are prone to show larger degree of dispersion. miR-#-5p and miR-#-3p have relative independent evolution/expression patterns and datasets of target mRNAs, which might also contribute to the phenomena of arm selection and/or arm switching. Simultaneously, miRNA/isomiR expression profiles may be regulated via arm selection and/or arm switching, and the dynamic miRNAome and isomiRome will adapt to functional and/or evolutionary pressures. A comprehensive analysis and further experimental study at the miRNA/isomiR levels are quite necessary for miRNA study.

1. Introduction

MicroRNAs (miRNAs) have been widely studied as a class of well-conserved negative regulatory molecules. They play an important role in biological processes by regulating gene expression at the posttranscriptional level [1, 2]. As endogenous small noncoding RNAs (ncRNAs) (~22 nt), miRNAs are generated from the cleavage of primary miRNAs (pri-miRNAs) and precursor miRNAs (pre-miRNAs) by Drosha and Dicer cleavage [3–5]. miRNA may be generated from 5p or 3p arm of pre-miRNA, and the selection is believed to be influenced by hydrogen-bonding selection [6]. Based on the typical miRNA genesis, one arm can produce abundant active mature miRNAs, while another arm can produce rare and inactive miRNAs* (miRNA star, also ever named passenger strand). However, increasing evidence indicates that both arms can generate mature miRNAs under specific developmental stages or species [7–13]. Indeed, many pre-miRNAs have been reported to yield two kinds of mature miRNAs, although the two products, miR-#-5p and miR-#-3p, may vary in expression levels. The term given to this dynamic selection and expression is "arm switching" [8, 14]. Evolutionary analysis demonstrates that both miR-#-5p and miR-#-3p are conserved, although the nondominant miRNA sequences are not well-conserved with dominant miRNA sequences [15]. Increasing reports indicate that the nondominantly expressed miRNA sequences may act as potential regulatory molecules with unexpectedly abundant expression levels [16–18].

Although the typical miRNA is annotated and studied as a single sequence, accumulating evidence suggests that multiple sequences with varied $5'$ and/or $3'$ ends or varied lengths have been detected from the miRNA locus. The annotated or canonical miRNA is only one specific member of the multiple sequences. These multiple sequences are termed miRNA variants, also named isomiRs [19–23]. The miRNA isoforms are mainly derived from imprecise cleavage

by Drosha/Dicer and 3′ addition events through miRNA processing and maturation processes. RNA editing and single nucleotide polymorphisms (SNPs) also contribute to the generation of these multiple isomiRs [22]. The occurrence of multiple isomiRs is quite common, and each miRNA locus can be associated with these various miRNA isoforms [9, 19, 21, 23–30]. Despite the fact that both miR-#-5p and miR-#-3p are generated from the pre-miRNA and can form miRNA:miRNA duplex through nucleotide complementary base pairing, the two miRNA loci may yield various isomiR expression profiles and patterns [31].

This study aimed to explore the potential evolutionary and expression divergences and relationships between miRNAs from different arms of different/same pre-miRNAs. First, we characterized the origins and nucleotide compositions of all the annotated human miRNAs. Second, we performed evolutionary analysis on the common miRNAs among 10 typical vertebrates and then analyzed the non-dominant miRNAs based on the pre-miRNAs. Finally, the expression analysis was performed in samples from female patients using published small RNA sequencing datasets. Because gender difference can affect isomiR expression profiles [32], and common variation affects various diseases and medically relevant characteristics in a sex-dependent manner [33], we selected female patients to analyze miRNA/isomiR expression profiles to avoid potential effects from gender difference. miRNA expression patterns were mainly estimated at the miRNA/isomiR levels, especially between homologous miRNAs and between miR-#-5p and miR-#-3p. This study provides insights on the arm selection and/or arm switching in miRNAs from the evolutionary and expression angles, which would partly be informative to understanding the dynamic miRNAome and isomiRome and to characterizing miRNA and isomiR expression profiles. Study from the isomiR level may be a necessary way to understand miRNA, especially for those isomiRs from ever termed passenger stand, which will contribute to further explore miRNA biogenesis and function.

2. Materials and Methods

2.1. Source Data and Primary Analysis. According to the evolutionary taxa and numbers of known miRNA genes, 10 vertebrate species were selected: *Petromyzon marinus* (pma, Agnathostomata), *Danio rerio* (dre, Pisces), *Xenopus tropicalis* (xtr, Amphibia), *Anolis carolinensis* (aca, Lepidosauria), *Gallus gallus* (gga, Aves), *Equus caballus* (eca, Mammalia, Laurasiatheria), *Bos taurus* (bta, Mammalia, Ruminantia), *Monodelphis domestica* (mdo, Mammalia, Metatheria), *Mus musculus* (mmu, Mammalia, Rodentia), and *Homo sapiens* (hsa, Mammalia, Primates, Hominidae). All the pre-miRNAs and miRNAs were retrieved from the miRBase database (Release 20.0, http://www.mirbase.org/) [34].

Location information of miRNA on pre-miRNAs was obtained according to the annotations in the miRBase database. Specifically, miRNA generated from 5p arm of pre-miRNA was named miR-#-5p (# indicated the detailed miRNA name, such as miR-100), and miRNA generated from 3p arm of pre-miRNA was named miR-#-3p. If there is no existing annotation, the detailed location distributions were determined using self-developed scripts. Many miRNAs may be generated from multicopy pre-miRNAs, and herein we only presented the detailed isomiR expression profiles based on location of the first pre-miRNA. In the study, miR-#-5p and miR-#-3p were defined as miRNA pairs generated from the 5p and 3p arm of pre-miRNA, respectively, and 5p-miRNA and 3p-miRNA were defined as the miRNAs generated from 5p or 3p arm of different pre-miRNAs.

2.2. Evolutionary Analysis of miRNAs in Ten Test Vertebrates. Known annotated miRNAs from ten vertebrates were comprehensively surveyed for common miRNA members using self-developed scripts. These miRNAs were further classified based on the unit of miRNA gene family because many miRNAs could belong to the same gene family based on homologous sequences with high sequence similarity. Those pre-miRNAs that were not comprehensively annotated (miR-#-5p or miR-#-3p was not simultaneously annotated based on limited studies), unannotated miRNA sequences, were predicted and obtained from consensus sequences using pre-miRNAs and known human miRNAs. The main reasons were as follows: (1) human miRNAs have been widely studied, and most miR-#-5p and miR-#-3p are reported and annotated; (2) most miRNAs are phylogenetically well-conserved across different animal species, and well-conserved consensus sequences are easily obtained using sequence alignment analysis; (3) although the miR-#-5p and miR-#-3p show different levels of evolutionary divergence, both of them are conserved; (4) according to the known miRNA sequences and pre-miRNAs, the detailed miR-#-5p and miR-#-3p sequences can be collected. The shared miRNAs were aligned using Clustal X 2.0 multiple sequence alignment [35]. Nucleotide divergence was analyzed using MEGA 5.10 software [36] and DnaSP 5.10.01 software [37]. Simultaneously, nucleotide diversity (π), haplotype diversity (Hd), and average number of nucleotide differences (k) for the miRNAs from different animal species were calculated using DnaSP software as special miRNA populations [38]. Evolutionary patterns were estimated based on nucleotide divergence across the ten animal species using percentage of nucleotide substitutions (transition and transversion) and insertions/deletions in each position. The reference nucleotide was denoted as human miRNA. Based on the potential length difference between miRNAs in different species, we only analyzed the core sequences and not the terminus nucleotides with deficiency (these nucleotides were mostly derived from length differences). Nucleotide divergence patterns were further estimated between 5p-miRNA and 3p-miRNA and between miR-#-5p and miR-#-3p.

In order to track the evolutionary history of pre-miRNAs and miRNAs from the different arms, especially between homologous miRNAs, phylogenetic trees of pre-miRNAs were reconstructed using the neighbor-net method [39] in SplitsTree 4.10 [40], and networks of miRNAs were defined based on Jukes-Cantor model and Network 4.6.1.1 (http://www.fluxus-engineering.com/) using the median-joining (MJ) method. Also, the free energies of some pre-miRNAs were estimated through the RNAfold WebServer (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) [41, 42].

2.3. Analysis of the miRNA/isomiR Expression Levels Using Public Sequencing Datasets. In order to understand the expression patterns of miR-#-5p and miR-#-3p pairs, we analyzed them at the miRNA/isomiR levels using small RNA sequencing datasets generated by The Cancer Genome Atlas (TCGA) pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions constituting the TCGA research network can be found at http://cancergenome.nih.gov/. Available small RNA sequencing datasets associated with the three kinds of women's diseases including breast cancer (BRCA), ovarian serous cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), and their respective control samples were selected to investigate miRNA expression patterns at the miRNA/isomiR levels (see Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/168358). We also conducted expression analysis in the three kinds of women's diseases dataset of some miRNAs (especially homologous miRNAs) identified from our evolutionary analysis. All of these high-throughput sequencing datasets were generated on Illumina HiSeq sequencing platform.

Reads per million (RPM) were used to estimate the relative expression levels, and relative expression rate (percentage) in the miRNA locus was used to assess the isomiR expression patterns across different samples. In order to track relative expression levels of miRNA/isomiR and reduce potential sequencing errors/mapping procedures, only those abundant miRNAs/isomiRs were selected to perform the analysis using larger sample sizes. The abundant expression and larger sample sizes could reduce error. Further, functional analysis was performed between miR-#-5p and miR-#-3p and between canonical miRNA sequences and their 5′ isomiRs (with the novel 5′ ends and seed sequences). According to the seed sequences, target mRNAs were predicted and obtained from TargetScan program (http://www.targetscan.org/).

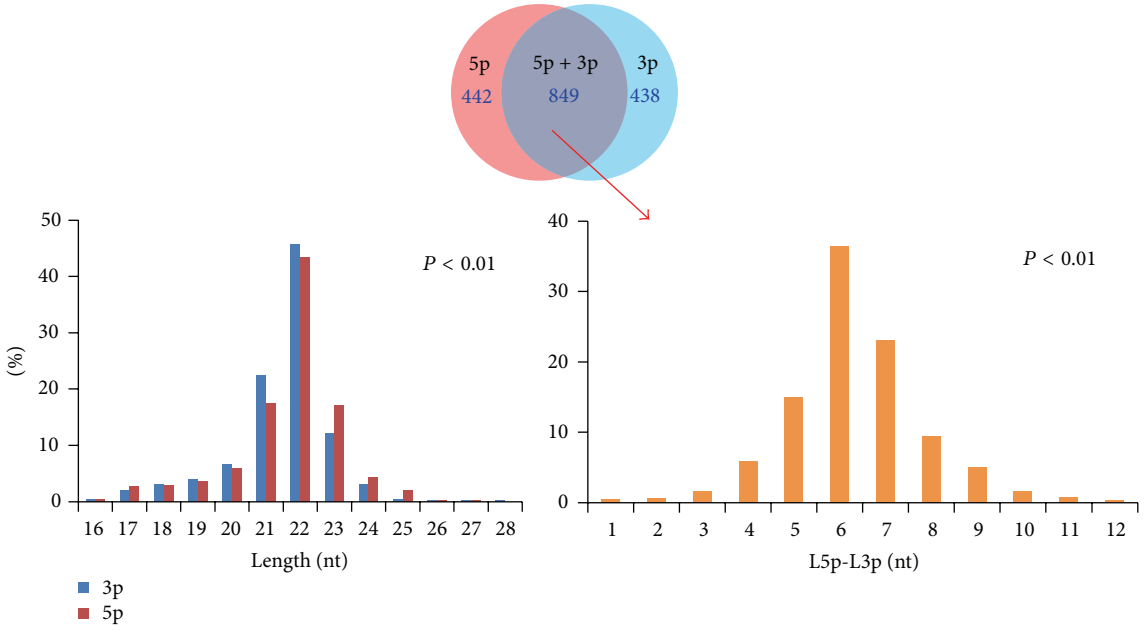
2.4. Statistical Analysis. Data were evaluated using paired t-test (length distributions between miR-#-5p and miR-#-3p), Student's t-test (length distributions between 5p-miRNA and 3p-miRNA), Chi-square test (nucleotide compositions between different miRNAs from 5p or 3p), Wilcoxon signed-rank test (nucleotide divergence pattern between miR-#-5p and miR-#-3p), and Spearman correlation test (nucleotide divergence between miR-#-5p and miR-#-3p and homologous miRNAs). Differences were considered statistically significant if the P value is less than 0.05. All tests were two-tailed and conducted using Stata software (Version 11.0).

3. Results

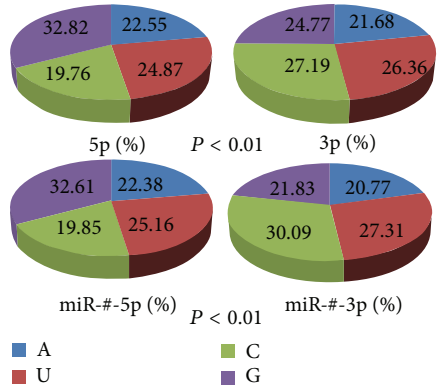
3.1. Primary Analysis of Human miR-#-5p/miR-#-3p and 5p-miRNA/3p-miRNA. There were 2,578 annotated human mature miRNAs in the miRBase database (Release 20.0). A total of 1,291 miRNAs were characterized from the 5p arms of pre-miRNAs, while the others were characterized from the 3p arms. Of these, 849 pairs were identified as miR-#-5p and miR-#-3p from the same pre-miRNAs. Both 5p-miRNA and 3p-miRNA or miR-#-5p and miR-#-3p had different length distributions (5p-miRNA, 21.67 ± 0.04, 3p-miRNA, 21.51 ± 0.04, $t = -2.68$, $P < 0.01$; miR-#-5p, 22.08 ± 0.04, miR-#-3p, 21.72 ± 0.04, $t = 6.01$, $P < 0.01$, Figure 1(a)). 5p-miRNA and 3p-miRNA showed different nucleotide compositions ($\chi^2 = 400.02$, $P < 0.01$, Figure 1(b) and Table 1). Guanine (G) was more predominant in 5p-miRNA (more than 32.82%) than in 3p-miRNA (24.77%). The predominant nucleotide in 3p-miRNA was cytosine (C) (27.19%), which was present at 19.76% in 5p-miRNA (Figure 1(b)). The presence of G, including double (GG), triple (GGG), and fourfold (GGGG) nucleotides, showed larger divergence between miRNAs from different arms (Figure 1(b) and Table 1). Similarly, the nucleotide composition was varied between miR-#-5p and miR-#-3p (Figure 1(b) and Table 1). Significant differences in the continuous nucleotide compositions could be detected between 5p-miR and 3p-miRNA and between miR-#-5p and miR-#-3p (Table 1). Compared to the total nucleotide compositions, nucleotides in each position along miRNA also showed significant difference between 5p-miR and 3p-miRNA and between miR-#-5p and miR-#-3p ($\chi^2 = 656.70$, $P < 0.01$, Figure 1(c); $\chi^2 = 813.57$, $P < 0.01$, Figure 1(d)), although the nucleotides 2–8, termed "seed sequences" of the miRNAs, did not display nucleotide bias.

3.2. Evolutionary Patterns of miR-#-5p/miR-#-3p and 5p-miRNA/3p-miRNA across Species. There were 31 miRNAs gene families (contain 43 miRNA members) shared by the 10 test animal species (Table S2). They may be composed of two or more members with high sequence similarity, but these members were not always shared by the 10 species. The common miRNA might have different number of pre-miRNAs (also termed multicopy pre-miRNAs) in different species and even have different number of homologous miRNAs (Figure S1).

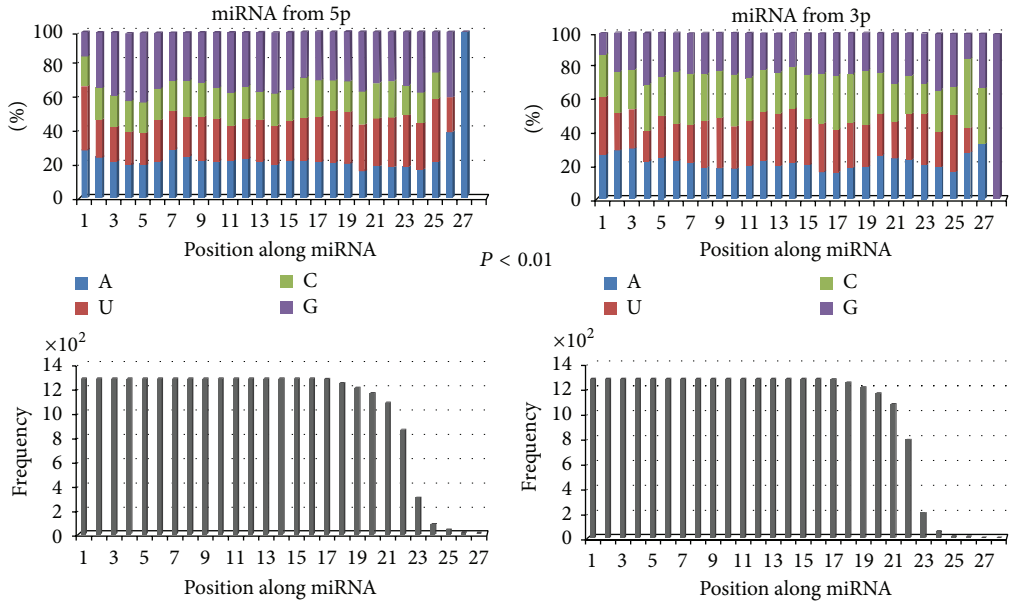
Although miRNAs are regarded as phylogenetically well-conserved small ncRNAs, different miRNAs, including homologous miRNAs, may show various evolutionary patterns (Figure S1) [15, 38]. Analysis of miR-#-5p and miR-#-3p revealed diverse variations in nucleotide composition (Figure S1). Compared to the dominant miRNAs, another strands showed higher levels of nucleotide diversity, haplotype diversity, and average number of nucleotide difference (Table 2). For example, let-7a-5p was highly conserved across the ten species, but let-7a-3p was associated with variation in the nucleotides. Generally, the dominant miRNAs were well-conserved, especially in the "seed sequences" (nucleotides 2–8), while nondominant miRNAs might display more variation in nucleotide composition (Figures S1C and S1D). Although both of them were reported as functional miRNAs existing at abundant levels in one or more species, 55.81% of miR-#-5p and miR-#-3p showed different levels of nucleotide divergence (Figure 2(a) and Table S3). The scatter plot analysis of the shared 43 miRNA genes revealed that both miR-#-5p and miR-#-3p were conserved (Figure 2(a)), with most sites showing minimal variation in nucleotide composition. Herein, 20 dominant miRNAs were identified as 5p-miRNA from 5p arm, and others (23 miRNAs) were identified as 3p-miRNA from 3p arm. We also analyzed the functional regions (seed sequences) of miRNAs, and only 4



(a)



(b)



(c)

Figure 1: Continued.

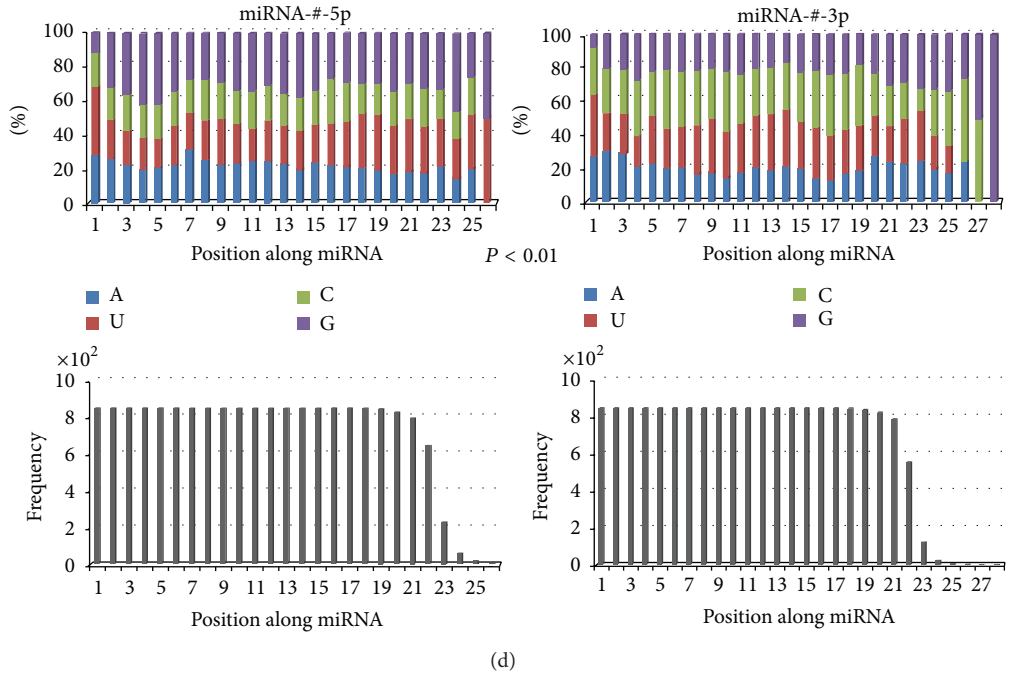


Figure 1: Primary analysis of human miRNAs according to their locations on pre-miRNAs. (a) Location and length distributions of human miRNAs. There are 849 pairs of miRNAs that are characterized as miR-#-5p and miR-#-3p from the same pre-miRNAs. Despite the fact that 22 nt is predominant length, the length distributions of 5p-miRNA and 3p-miRNA are highly variable ($t = -2.68$, $P < 0.01$). The frequency distribution of D-value (difference value) of miR-#-5p and miR-#-3p indicates that the two arms of pre-miRNA are likely to generate different miRNAs with different lengths ($t = 6.01$, $P < 0.01$). (b) Nucleotide compositions between 5p-miRNA and 3p-miRNA, miR-#-5p and miR-#-3p. Guanine (G) is the most predominant nucleotide in 5p-miRNA and miR-#-3p (more than 32%), while moderate distributions of the four nucleotides can be detected in 3p-miRNA and miR-#-3p. The two kinds of miRNAs are likely to have different nucleotide compositions ($\chi^2 = 400.02$, $P < 0.01$). (c) Difference in nucleotide compositions based on the position along miRNA is detected between 5p-miRNA and 3p-miRNA ($\chi^2 = 656.70$, $P < 0.01$). The frequency distributions of nucleotides in each position are also presented here. (d) The difference in nucleotide compositions based on position along miRNA is detected between miRNA-#-5p and miRNA-#-3p ($\chi^2 = 813.57$, $P < 0.01$). The frequency distributions of nucleotides in each position are also presented here.

Table 1: Frequency of nucleotide compositions between all human miRNAs from different arms.

Nucleotides	5p-miRNA (%)	3p-miRNA (%)	miR-#-5p (%)	miR-#-3p (%)	χ^2, P
AA	18.38	18.13	17.98[c]	17.09	
UU	22.21	24.39	22.59	25.68	
CC	17.39	30.64	17.80	36.06	
GG	42.02	26.84	41.63	21.17	
Total	**100**	**100**	**100**	**100**	21.31, 0.01
AAA	16.26[a]	17.77	14.39[d]	16.21	
UUU	20.83	23.41	21.36	24.98	
CCC	15.38	33.80	15.40	40.90	
GGG	47.53	25.02	48.85	17.91	
Total	**100**	**100**	**100**	**100**	42.74, 0.00
AAAA	17.72[b]	19.87	14.97[e]	17.95	
UUUU	21.44	22.03	20.38	22.76	
CCCC	8.97	34.99	8.60	45.83	
GGGG	51.86	23.11	56.05	13.46	
Total	**100**	**100**	**100**	**100**	81.10, 0.00

The percentage is estimated based on frequency in all the 5p- or 3p-miRNAs, all the miR-#-5p or miR-#-3p. [a]A significant difference in the triple repetitive nucleotides can be detected between 5p-miRNA and 3p-miRNA ($\chi^2 = 14.82$, $P < 0.01$), [b]a significant difference in the four repetitive nucleotides can be detected between 5p-miRNA and 3p-miRNA ($\chi^2 = 26.71$, $P < 0.01$), [c]a significant difference in the double repetitive nucleotides can be detected between miR-#-5p and miR-#-3p ($\chi^2 = 13.21$, $P < 0.01$), [d]a significant difference in the triple repetitive nucleotides can be detected between miR-#-5p and miR-#-3p ($\chi^2 = 26.89$, $P < 0.01$), and [e]a significant difference in the four repetitive nucleotides can be detected between miR-#-5p and miR-#-3p ($\chi^2 = 52.17$, $P < 0.01$).

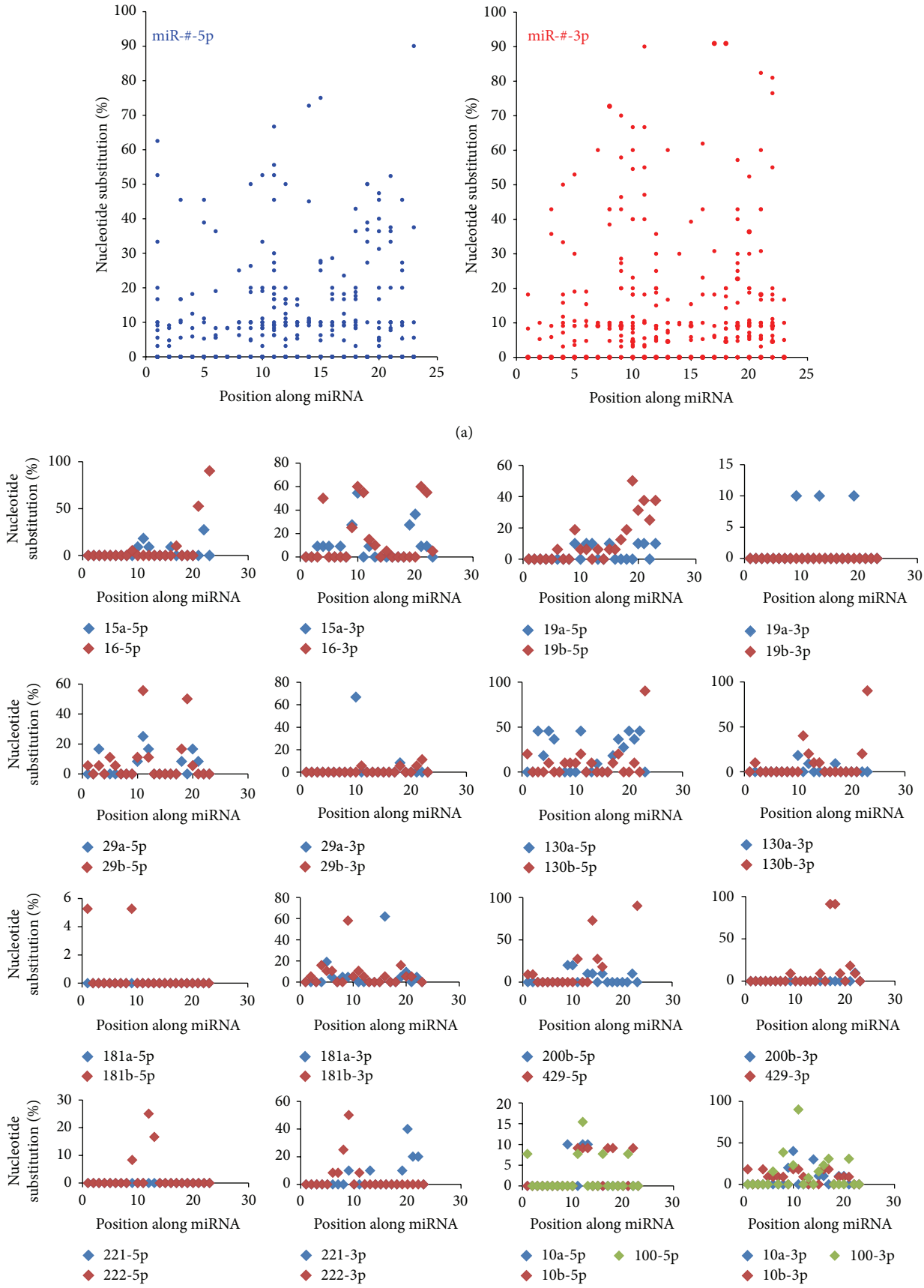
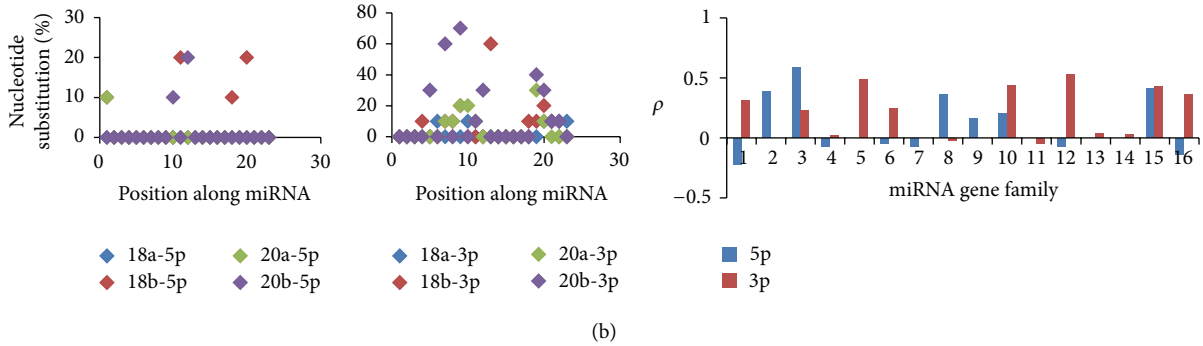


Figure 2: Continued.



(b)

Figure 2: Scatter plots of miR-#-5p and miR-#-3p (a); homologous miRNAs and distribution of Spearman correlation coefficients (b). (a) Scatter plots of nucleotide substitution rates in the common miRNAs and their nondominant strands based on each position along miRNA (the conservation level was estimated based on nucleotide substitution rates along miRNA). (b) Scatter plots of nucleotide substitution between homologous miRNAs (miRNA gene family). miR-#-5p and miR-#-3p were compared and analyzed. The last figure indicates the distribution of Spearman correlation coefficient (ρ) across different miRNA gene families.

Table 2: Nucleotide diversity (π), haplotype diversity (Hd), and average number of nucleotide differences (k) of different miRNA populations.

miRNA	miRNA-#-5p			miR-#-3p		
	π	Hd	k	π	Hd	k
let-7a	0.00	0.00	—	0.18 ± 0.01	0.86 ± 0.05	3.69
Total (let-7 family, 76, 45)	0.12 ± 0.01	0.92 ± 0.01	2.60	0.25 ± 0.01	0.93 ± 0.01	4.90
miR-30b	0.00	0.00	—	0.14 ± 0.04	0.83 ± 0.13	2.86
miR-30c	0.00	0.00	—	0.11 ± 0.03	0.80 ± 0.08	2.09
miR-30d	0.04 ± 0.01	0.58 ± 0.16	0.91	0.12 ± 0.03	0.77 ± 0.13	2.35
Total (part mir-30 family)	0.08 ± 0.01	0.52 ± 0.08	1.83	0.28 ± 0.02	0.93 ± 0.02	5.26

These parameters are estimated according to Figure S1.

pairs (9.30%) indicated difference (Table S3). The difference in average percentages from all the miR-#-5p and miR-#-3p was not significant ($Z = -1.642$, $P > 0.05$), and similar result could be detected based on the dominant miRNA ($Z = -1.55$, $P > 0.05$). Furthermore, although homologous miRNAs displayed close sequence, functional, and evolutionary relationships, no significant correlations were detected between most of homologous miRNAs (Figure 2(b) and Table S4).

Phylogenetic trees and networks were reconstructed using pre-miRNAs and miRNAs from Figure S1, respectively (Figure 3). The phylogenetic tree of let-7a was split into three clusters, and each cluster contained pre-miRNAs from different animal species (Figure 3(a)). Compared to the tree of the single miRNA gene of let-7a, the phylogenetic tree of homologous mir-30b, mir-30c, and mir-30d could be split (Figure 3(b)). mir-30d showed larger genetic distance with mir-30b and mir-30c. The pma-mir-30b and pma-mir-30c were clustered with mir-30d, which indicates that these should be members of pma-mir-30d (Figure S1 and Figure 3). The evolutionary networks of miR-#-5p and miR-#-3p showed various patterns (Figures 3(c) and 3(d)). Different types of sequences (termed miRNA haplotypes) were classified with different frequencies. For example, let-7a-5p was highly conserved across the ten animal species, and only one specific sequence was identified. However, let-7a-3p was associated with high nucleotide variation and showed a complex evolutionary network (Figure S1A and Figure 3(c)). Compared to let-7, both evolutionary networks of miR-30-5p and miR-30-3p showed clear module networks based on miRNA members (Figure 3(d)).

3.3. Expression Analysis of miR-#-5p/miR-#-3p at the miRNA/isomiR Levels. We analyzed available miRNA datasets of 2,144 patients or volunteers with women's diseases (BRCA, OV, or UCEC) and their relevant controls (Table S1). Following evolutionary analysis, several miRNAs were selected to perform expression analysis using these sequencing datasets. Generally, in the miRNA locus, only several isomiRs were dominantly expressed (Figure 4 and Tables S6, S7, and S8). Homologous miRNAs were likely to show similar isomiR expression pattern, such as miR-30a and miR-30e (Figure 4). Dominant miRNAs and their multiple isomiRs were present at abundant expression levels, while most of nondominant strands were not abundant. Abundantly expressed isomiRs were always near the most dominant isomiR sequence. Specifically, their 5′ or 3′ ends either were the same or differ at 1-2 nucleotides (Figure 4 and Tables S6, S7, and S8). The standard deviation (SD) of the average percentage of each isomiR showed diverse distributions (Figure 5 and Figures S2, S3, and S4). Different miRNAs showed different types of isomiRs with diverse expression distribution and SD (Figures 4 and 5 and Figures S2, S3, and S4). Abundantly expressed isomiRs were likely to

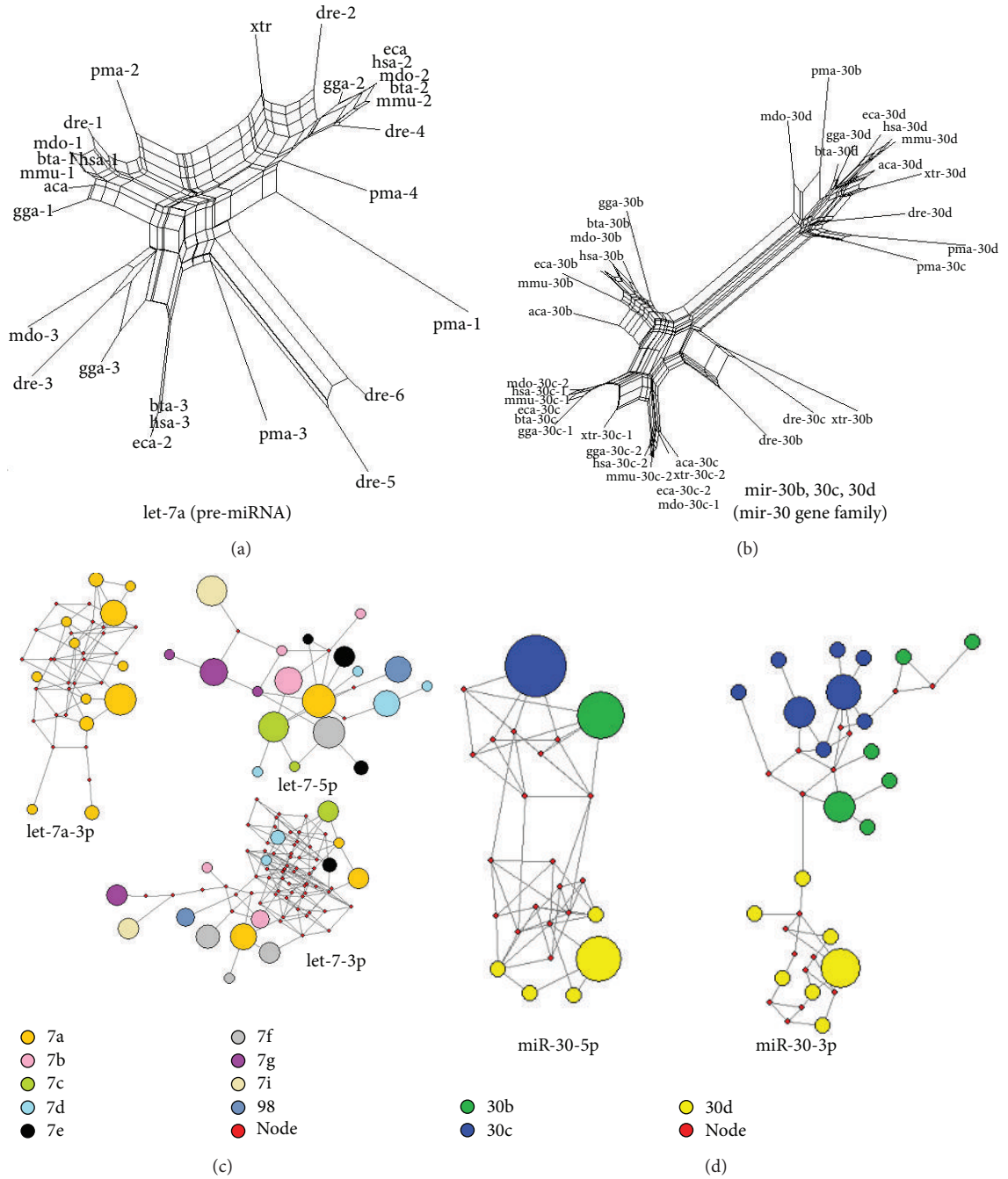


FIGURE 3: Examples of evolutionary patterns of different miRNAs. (a) Phylogenetic tree of let-7a. The tree is reconstructed using all the miRNA precursors, including multicopy pre-miRNAs. Multicopy pre-miRNAs are likely to be located in different clusters. (b) The phylogenetic tree of several members of mir-30 gene family (mir-30b, mir-30c, and mir-30d). The three members are split, and miR-30d shows a larger genetic distance with other members. (c) MJ networks of let-7a-3p, let-7-5p, and let-7-3p. Let-7a-3p is associated with nucleotide variation, although the other strand, let-7a-5p, is highly conserved. Similarly, let-7-5p from let-7 family also shows a simple network with several median vectors compared to let-7a-3p. Both let-7a-3p and let-7-3p show complex evolutionary networks with more median vectors. Members in let-7 gene family do not show clear module networks. The size of the circle shows the frequency of the miRNA haplotype (the specific miRNA sequence). (d) MJ networks of miR-30-5p and miR-30-3p from known miR-30b, miR-30c, and miR-30d sequences. Different miRNA members are likely to cluster together.

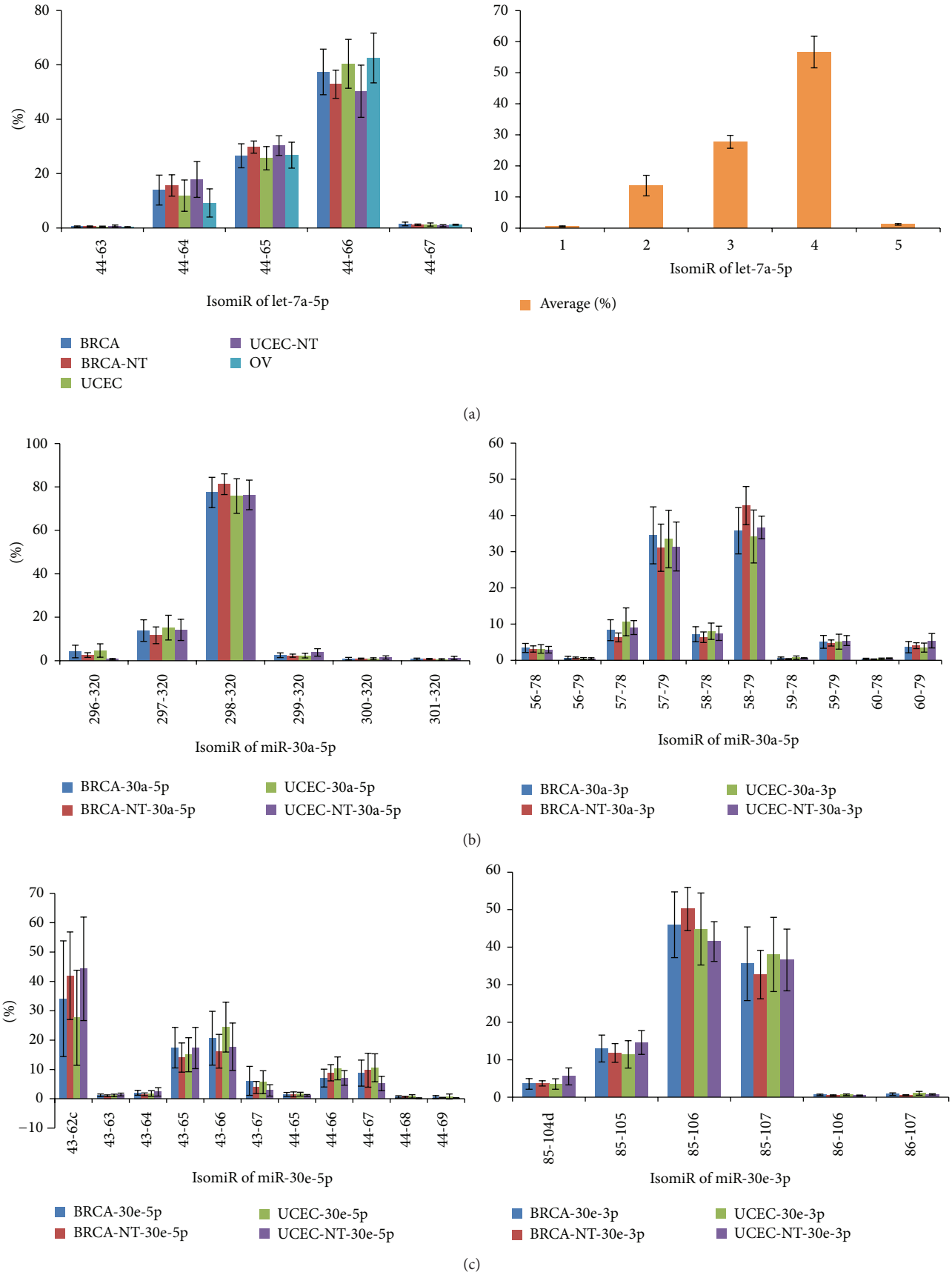
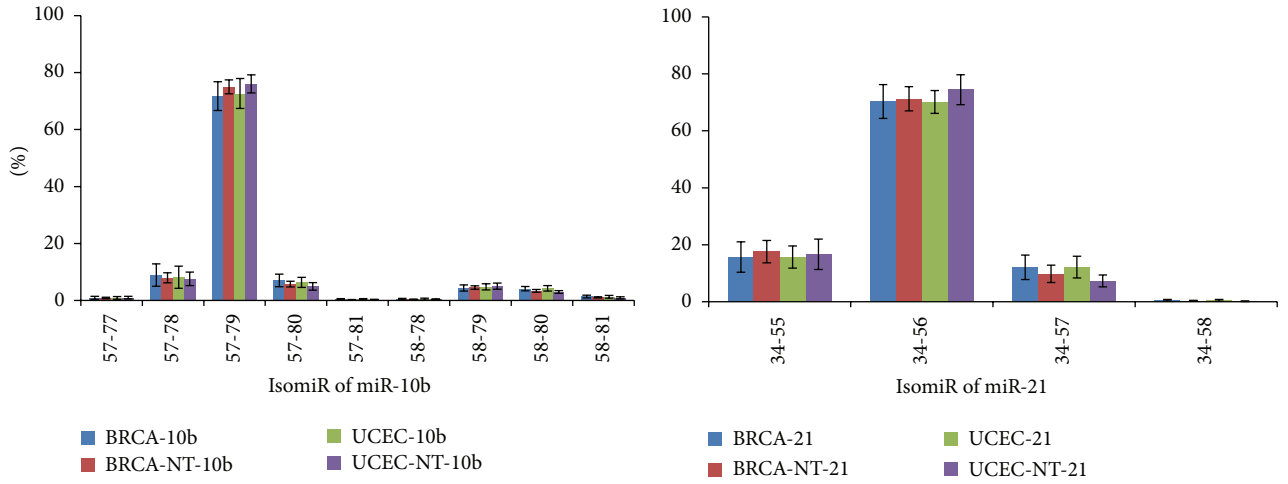


Figure 4: Continued.



(d)

Figure 4: IsomiR expression patterns across different samples. IsomiR is presented here based on the location on chromosome (the detailed location distributions can be found in Tables S6, S7, and S8). The percentage shows the relative expression levels in the miRNA locus. The mean and standard deviation are presented in the figure. BRCA-NT or UCEC-NT shows normal samples that match tumor samples. (a) IsomiR expression patterns of let-7a-5p across the five kinds of samples. Similar distributions can be found across the different samples. The right bar chart indicates distribution of the mean percentage and standard deviation of the five kinds of samples. ((b)-(c)) IsomiR expression patterns of homologous miR-30a and miR-30e. Both of them can generate two kinds of abundant products (miR-#-5p and miR-#-3p). The two arms may show various isomiR expression patterns, but homologous miRNAs are likely to show similar expression patterns. (d) IsomiR expression patterns of miR-10b and miR-21.

be detected larger SD (Figure 4 and Figures S2 and S3), and similar SD distributions could be found between diseased and normal samples (Figure 5 and Figure S4). Generally, at the isomiR level, the average percentages of samples from disease patients would be involved in larger divergence than control samples, and similar results can be detected based on all miRNAs (Figure 5 and Figure S4).

3.4. Functional Analysis of miR-#-5p/miR-#-3p at the miRNA/isomiR Levels. Although miR-#-5p and miR-#-3p had different sequences and seed sequences, some common targets could be detected (Figure S5A). These miRNA pairs could bind different regions in UTR (untranslated regions) of target mRNAs, although the phenomenon was rare (larger amounts of specific targets could be detected). The common targets were more popular between the canonical miRNA sequences and their 5′ isomiRs, despite the fact that "seed shifting" could be detected between them (Figures S5B and S5C). There were about half of target mRNAs of 5′ isomiRs that were shared by the canonical miRNA sequences, although these 5′ isomiRs were involved in novel seed sequences via "seed shifting" events.

4. Discussion

4.1. Evolutionary Divergence between miRNAs from Different Arms. miRNAs have been widely regarded as a class of crucial negative regulatory molecules with important biological roles, especially for their roles in tumorigenesis. Based on the current annotated human miRNAs, similar numbers of 5p-miR and 3p-miR show well-conserved sequences across different species, although they are involved in inconsistent length distributions and nucleotide compositions, including multiple repetitive nucleotides (Figures 1(a)–1(c), Figure 2, and Table 1). This difference may be influenced by larger sample sizes. Simultaneously, mirtrons have been reported as alternative precursors for miRNA biogenesis in vertebrates [43], which may lead to the difference of nucleotide compositions because of nucleotide biases in mirtrons. There are 849 pairs that are identified as miR-#-5p and miR-#-3p, and significant difference in length distributions and nucleotide compositions is detected between the two arms (Figures 1(b)–1(d), Table 1, and Table S3). Evolutionary analysis shows that both dominant and nondominant miRNAs are conserved, although the nondominant miRNA is associated with more nucleotide variation across homologous miRNAs and different species [15]. Phylogenetic relationship shows that these multicopy pre-miRNAs are located in different clusters (Figure 3), which suggests the similar distributions of miRNA genes across different species. The well-conserved sequence contributes to stable miRNA-mRNA regulatory network, and simultaneously, the evolutionary process is also controlled by functional pressures. The two arms of pre-miRNA showed various evolutionary patterns via different levels of nucleotide substitutions and insertions/deletions (Figure S1, Figure 2, and Table S3), which may influence stem-loop structure of pre-miRNA (Table S5). However, both of the two arms are always well-conserved in the functional region, termed the "seed sequences" (Figure 3(a) and Table S3). These results suggest that both products from the two arms are regulatory molecules, although they always have various expression levels.

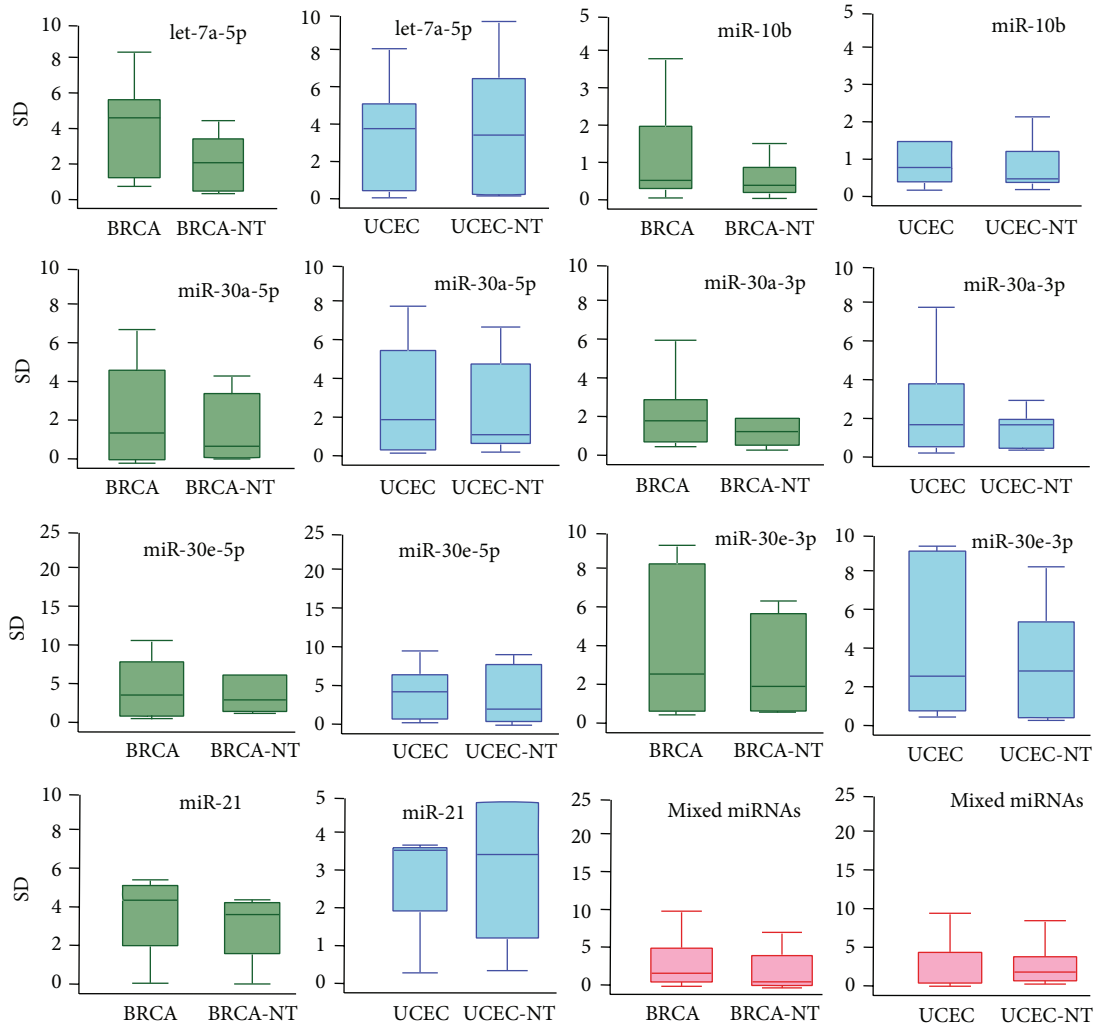


FIGURE 5: Box plots of miRNAs in BRCA and UCEC using standard deviation (SD). The box in green indicates the SD distribution of BRCA samples (BRCA and BRCA-NT), the box in blue indicates the SD distribution of UCEC samples (UCEC and UCEC-NT), and the box in pink indicates mixed miRNAs in different samples.

Homologous and clustered miRNAs are commonly found in miRNAs [44]. No significant relationships between these homologous miRNAs can be detected (Figure 2(b) and Table S4). These findings indicate relatively rapid evolutionary patterns between homologous miRNAs, especially between the less well-conserved nondominant strands (Figure 2(b)). Despite the possibility that these miRNAs have evolved from the common ancient miRNA gene, varied nucleotides in miRNAs, especially in the "seed sequences," will generate novel miRNAs with novel candidate target mRNAs. Simultaneously, coevolution of miRNA and target mRNAs also contributes to the varied miRNAs across different species [45]. Taken together, homologous miRNAs may provide a method to generate novel miRNA genes via duplication events, and multicopy pre-miRNAs are probably transitional products. The driving force should be mainly derived from functional and evolutionary pressures, which largely contributes to the dynamic miRNAome, and enriches the potential relationships between different miRNAs.

4.2. Expression and Function between miRNAs from Different Arms. Similar to our previous studies [21, 46, 47], we found that only several isomiRs (always 1–3) are dominantly expressed, and others have lower expression rate (Figure 4 and Tables S6, S7, and S8). The interesting distributions are consistent in different individuals, including samples from patients with disease and healthy controls. The similar distributions suggest that isomiR expression patterns are always stable across different samples [21, 26]. The characteristics of these dominant isomiRs provide the possibility of imprecise cleavage of Drosha and Dicer through pre-miRNA processing and miRNA maturation processes. Indeed, due to the smaller size of miRNA sequence (~22 nt), degradation of hairpins may also be one factor that contributes to rare isomiRs [48]. Although the distribution of isomiR expression is similar across different samples, no significant correlations can be found between isomiR expression profiles of miR-#-5p and miR-#-3p (Figure 4). Simultaneously, various standard values of deviation can be found (Figure 5 and Figures S2, S3, and

S4). Compared to control samples, samples from patients with disease may be involved in larger expression divergence across different samples (Figure 5). This suggests that a more flexible expression of isomiRs can be detected across different samples from patients with disease compared to control samples. Functional analysis showed that some common target mRNAs between miR-#-5p and miR-#-3p can be detected, although they have no different sequences and most target mRNAs are specific (Figure S5A). Simultaneously, more shared target mRNAs are obtained between the canonical miRNA and 5′ isomiRs despite being with "seed shifting" events (Figures S5B and S5C). The interesting results imply that multiple isomiRs may coordinately contribute to the specific biological processes by binding different regions in UTR. Moreover, 3′ addition events (isomiRs with additional nontemplate nucleotides in 3′ ends) are quite common in isomiRome, while no further analysis is performed in the present study based on the previous TCGA datasets. The phenomenon of 3′ additions may have versatile biological roles, including affecting target selection or miRNA stability [22, 24, 26, 49]. Collectively, analyzing multiple isomiRs and their expression patterns is the first step towards a systematic understanding of the miRNA world, including the genesis and regulatory roles of miRNAs.

miRNAs are likely to be members of miRNA gene families/clusters sharing high sequence similarity or close location distribution. These homologous/clustered miRNAs may have evolved from ancestor genes via part or tandem historic duplication events [15, 50–52]. Previous study reported that homologous miRNAs are likely to show similar isomiR expression patterns [47], and our results are consistent with this observation (Figure 4 and Table S7). The similarity in the expression patterns implies that the pre-miRNA processing and miRNA maturation processes should be derived from the ancestral gene, which may contribute to the potential interactions in the regulatory network [47]. Moreover, we found that deregulated miRNAs are likely to have different types of isomiRs (miR-30a, miR-30e, and miR-10b, Figure 4 and Tables S6, S7, and S8). These deregulated miRNAs have been reported in breast cancer [53, 54], and the moderate expression patterns can be detected. No enough evidence indicates that miRNA with moderate isomiR expression is likely to be abnormally expressed and contributes to abnormal biological roles. More studies, especially for experimental validation, are needed to further study the small noncoding RNAs at the isomiR level.

4.3. Selection of 5p and 3p or Switching between the Two Arms in miRNAome/isomiRome. The phenomenon of arm selection shows that miRNAs may be derived from different arms, and the arm switching phenomenon suggests that the two arms may also show dynamic expression patterns. miRNAs from the two arms (they can form miRNA:miRNA duplex) always show different evolutionary patterns and also have various expression levels and isomiR expression patterns. Most of pre-miRNAs only produce one dominant and one rare miRNAs in specific samples, although the expression rate of the two miRNAs may be changed in other samples (arm switching phenomenon). Indeed, the two arms of many pre-miRNAs are conserved (especially in "seed sequences"), providing the possibility to be regulatory molecules, and the arm switching phenomenon further enriches the dynamic miRNAome by controlling miRNA expression profiles to adapt to functional and/or evolutionary needs. Expression and evolution patterns in miR-#-5p and miR-#-3p are relatively independent, and they are prone to regulate different targets. Based on the phenomena of arm selection or arm switching, the dynamic miRNAome also represents the multiple and dynamic isomiRome at the isomiR level. These isomiRs provide more information towards further understanding of miRNAs, in that isomiR expression patterns may indicate the characteristics of pre-miRNA processing and miRNA maturation processes. Thus it is worth exploring the biological roles of miRNAs at the isomiR level and the origin of miRNAs (5p or 3p) and related miRNAs based on miRNA gene family/cluster. Taken together, the arm selection and/or arm switching may be an important method to regulate miRNAome and isomiRome, and the dynamic miRNA and isomiR expression profiles will adapt to functional and/or evolutionary pressures.

Abbreviations

miRNA:	MicroRNA
ncRNA:	Noncoding RNA
pri-miRNA:	Primary miRNA
pre-miRNA:	Precursor miRNA
miRNA*:	miRNA star
SNP:	Single nucleotide polymorphism
MJ:	Median-joining
TCGA:	The Cancer Genome Atlas
BRCA:	Breast cancer
OV:	Ovarian serous cystadenocarcinoma
UCEC:	Uterine corpus endometrial carcinoma
SD:	Standard deviation
pma:	*Petromyzon marinus*
dre:	*Danio rerio*
xtr:	*Xenopus tropicalis*
aca:	*Anolis carolinensis*
gga:	*Gallus gallus*
eca:	*Equus caballus*
bta:	*Bos taurus*
mdo:	*Monodelphis domestica*
mmu:	*Mus musculus*
hsa:	*Homo sapiens*.

Conflict of Interests

The authors declare no potential conflict of interests with respect to the authorship and/or publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 61301251, 81473070, and 81373102), the Research Fund for the Doctoral Program of Higher Education of China (20133234120009), the National Natural Science Foundation of Jiangsu (no. BK20130885), the Natural Science Foundation of the Jiangsu Higher Education

Institutions (nos. 12KJB310003 and 13KJB330003), Shandong Provincial Key Laboratory of Functional Macromolecular Biophysics, and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

[1] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, 2009.

[2] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[3] Y. Lee, C. Ahn, J. Han et al., "The nuclear RNase III Drosha initiates microRNA processing," *Nature*, vol. 425, no. 6956, pp. 415–419, 2003.

[4] J. Han, Y. Lee, K.-H. Yeom, Y.-K. Kim, H. Jin, and V. N. Kim, "The Drosha-DGCR8 complex in primary microRNA processing," *Genes & Development*, vol. 18, no. 24, pp. 3016–3027, 2004.

[5] J. Han, Y. Lee, K.-H. Yeom et al., "Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex," *Cell*, vol. 125, no. 5, pp. 887–901, 2006.

[6] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, "miRBase: microRNA sequences, targets and gene nomenclature," *Nucleic Acids Research*, vol. 34, pp. D140–D144, 2006.

[7] S. C. Li, W. C. Chan, M. R. Ho et al., "Discovery and characterization of medaka miRNA genes by next generation sequencing platform," *BMC Genomics*, vol. 11, no. 4, article S8, 2010.

[8] S. Griffiths-Jones, J. H. L. Hui, A. Marco, and M. Ronshaugen, "MicroRNA evolution by arm switching," *EMBO Reports*, vol. 12, no. 2, pp. 172–177, 2011.

[9] N. Cloonan, S. Wani, Q. Xu et al., "MicroRNAs and their isomiRs function cooperatively to target common biological pathways," *Genome Biology*, vol. 12, no. 12, article R126, 2011.

[10] A. Marco, J. H. L. Hui, M. Ronshaugen, and S. Griffiths-Jones, "Functional shifts in insect microRNA evolution," *Genome Biology and Evolution*, vol. 2, no. 1, pp. 686–696, 2010.

[11] S.-C. Li, Y.-L. Liao, M.-R. Ho, K.-W. Tsai, C.-H. Lai, and W.-C. Lin, "miRNA arm selection and isomiR distribution in gastric cancer," *BMC Genomics*, vol. 13, supplement 1, article S13, 2012.

[12] S.-C. Li, Y.-L. Liao, W.-C. Chan et al., "Interrogation of rabbit miRNAs and their isomiRs," *Genomics*, vol. 98, no. 6, pp. 453–459, 2011.

[13] W. C. Cheng, I. F. Chung, T. S. Huang et al., "YM500: a small RNA sequencing (smRNA-seq) database for microRNA research," *Nucleic Acids Research*, vol. 41, no. 1, pp. D285–D294, 2013.

[14] S.-C. Li, K.-W. Tsai, H.-W. Pan, Y.-M. Jeng, M.-R. Ho, and W.-H. Li, "MicroRNA 3' end nucleotide modification patterns and arm selection preference in liver tissues," *BMC Systems Biology*, vol. 6, no. 2, article S14, 2012.

[15] L. Guo and Z. Lu, "The fate of miRNA∗ strand through evolutionary analysis: implication for degradation as merely carrier strand or potential regulatory molecule?" *PLoS ONE*, vol. 5, no. 6, Article ID e11387, 2010.

[16] K. Okamura, M. D. Phillips, D. M. Tyler, H. Duan, Y.-T. Chou, and E. C. Lai, "The regulatory activity of microRNA star species has substantial influence on microRNA and 3' UTR evolution," *Nature Structural and Molecular Biology*, vol. 15, no. 4, pp. 354–363, 2008.

[17] K. Okamura, A. Ishizuka, H. Siomi, and M. C. Siomi, "Distinct roles for argonaute proteins in small RNA-directed RNA cleavage pathways," *Genes and Development*, vol. 18, no. 14, pp. 1655–1666, 2004.

[18] G. Jagadeeswaran, Y. Zheng, N. Sumathipala et al., "Deep sequencing of small RNA libraries reveals dynamic regulation of conserved and novel microRNAs and microRNA-stars during silkworm development," *BMC Genomics*, vol. 11, no. 1, article 52, 2010.

[19] P. Landgraf, M. Rusu, R. Sheridan et al., "A mammalian microRNA expression atlas based on small RNA library sequencing," *Cell*, vol. 129, no. 7, pp. 1401–1414, 2007.

[20] R. D. Morin, G. Aksay, E. Dolgosheina et al., "Comparative analysis of the small RNA transcriptomes of *Pinus contorta* and *Oryza sativa*," *Genome Research*, vol. 18, no. 4, pp. 571–584, 2008.

[21] L. Guo, Q. Yang, J. Lu et al., "A comprehensive survey of miRNA repertoire and 3′ addition events in the placentas of patients with pre-eclampsia from high-throughput sequencing," *PLoS ONE*, vol. 6, no. 6, Article ID e21072, 2011.

[22] C. T. Neilsen, G. J. Goodall, and C. P. Bracken, "IsomiRs—the overlooked repertoire in the dynamic microRNAome," *Trends in Genetics*, vol. 28, no. 11, pp. 544–549, 2012.

[23] L. W. Lee, S. Zhang, A. Etheridge et al., "Complexity of the microRNA repertoire revealed by next-generation sequencing," *RNA*, vol. 16, no. 11, pp. 2170–2180, 2010.

[24] H. A. Ebhardt, H. H. Tsang, D. C. Dai, Y. Liu, B. Bostan, and R. P. Fahlman, "Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications," *Nucleic Acids Research*, vol. 37, no. 8, pp. 2461–2470, 2009.

[25] F. Kuchenbauer, R. D. Morin, B. Argiropoulos et al., "In-depth characterization of the microRNA transcriptome in a leukemia progression model," *Genome Research*, vol. 18, no. 11, pp. 1787–1797, 2008.

[26] A. M. Burroughs, Y. Ando, M. J. L. De Hoon et al., "A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness," *Genome Research*, vol. 20, no. 10, pp. 1398–1410, 2010.

[27] A. F. Fernandez, C. Rosales, P. Lopez-Nieva et al., "The dynamic DNA methylomes of double-stranded DNA viruses associated with human cancer," *Genome Research*, vol. 19, no. 3, pp. 438–451, 2009.

[28] S. Lu, Y. H. Sun, and V. L. Chiang, "Adenylation of plant miRNAs," *Nucleic Acids Research*, vol. 37, no. 6, pp. 1878–1885, 2009.

[29] C. Shao, Q. Wu, J. Qiu et al., "Identification of novel microRNA-like-coding sites on the long-stem microRNA precursors in *Arabidopsis*," *Gene*, vol. 527, no. 2, pp. 477–483, 2013.

[30] J. Zhang, S. Zhang, S. Li et al., "A genome-wide survey of microRNA truncation and 3′ nucleotide addition events in larch (*Larix leptolepis*)," *Planta*, vol. 237, no. 4, pp. 1047–1056, 2013.

[31] L. Guo, H. Zhang, Y. Zhao, S. Yang, and F. Chen, "Selected isomiR expression profiles via arm switching?" *Gene*, vol. 533, no. 1, pp. 149–155, 2014.

[32] P. Loher, E. R. Londin, and I. Rigoutsos, "IsomiR expression profiles in human lymphoblastoid cell lines exhibit population and gender dependencies," *Oncotarget*, vol. 30, no. 5, pp. 8790–8802, 2014.

[33] W. P. Gilks, J. K. Abbott, and E. H. Morrow, "Sex differences in disease genetics: evidence, evolution, and detection," *Trends in Genetics*, vol. 30, pp. 453–463, 2014.

[34] A. Kozomara and S. Griffiths-Jones, "MiRBase: integrating microRNA annotation and deep-sequencing data," *Nucleic Acids Research*, vol. 39, no. 1, pp. D152–D157, 2011.

[35] 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.

[36] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.

[37] P. Librado and J. Rozas, "DnaSP v5: a software for comprehensive analysis of DNA polymorphism data," *Bioinformatics*, vol. 25, no. 11, pp. 1451–1452, 2009.

[38] L. Guo, B. Sun, F. Sang, W. Wang, and Z. Lu, "Haplotype distribution and evolutionary pattern of miR-17 and miR-124 families based on population analysis," *PLoS ONE*, vol. 4, no. 11, Article ID e7944, 2009.

[39] D. Bryant and V. Moulton, "Neighbor-Net: an Agglomerative Method for the Construction of Phylogenetic Networks," *Molecular Biology and Evolution*, vol. 21, no. 2, pp. 255–265, 2004.

[40] D. H. Huson, "SplitsTree: analyzing and visualizing evolutionary data," *Bioinformatics*, vol. 14, no. 1, pp. 68–73, 1998.

[41] I. L. Hofacker, "Vienna RNA secondary structure server," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3429–3431, 2003.

[42] R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen et al., "ViennaRNA Package 2.0," *Algorithms for Molecular Biology*, vol. 6, no. 1, article 26, 2011.

[43] E. Berezikov, W.-J. Chung, J. Willis, E. Cuppen, and E. C. Lai, "Mammalian mirtron genes," *Molecular Cell*, vol. 28, no. 2, pp. 328–336, 2007.

[44] L. Guo, Y. Zhao, H. Zhang, S. Yang, and F. Chen, "Integrated evolutionary analysis of human miRNA gene clusters and families implicates evolutionary relationships," *Gene*, vol. 534, no. 1, pp. 24–32, 2014.

[45] S. Lehnert, P. Van Loo, P. J. Thilakarathne, P. Marynen, G. Verbeke, and F. C. Schuit, "Evidence for co-evolution between human MicroRNAs and Alu-repeats," *PLoS ONE*, vol. 4, no. 2, Article ID e4456, 2009.

[46] L. Guo, F. Chen, and Z. Lu, "Multiple isomiRs and diversity of miRNA sequences unveil evolutionary roles and functional relationships across animals," in *MicroRNA and Non-Coding RNA: Technology, Developments and Applications*, pp. 127–144, 2013.

[47] L. Guo, H. Li, T. Liang et al., "Consistent isomiR expression patterns and 3′ addition events in miRNA gene clusters and families implicate functional and evolutionary relationships," *Molecular Biology Reports*, vol. 39, no. 6, pp. 6699–6706, 2012.

[48] M. R. Friedländer, W. Chen, C. Adamidi et al., "Discovering microRNAs from deep sequencing data using miRDeep," *Nature Biotechnology*, vol. 26, no. 4, pp. 407–415, 2008.

[49] S. L. Fernandez-Valverde, R. J. Taft, and J. S. Mattick, "Dynamic isomiR regulation in Drosophila development," *RNA*, vol. 16, no. 10, pp. 1881–1888, 2010.

[50] J. Hertel, M. Lindemeyer, K. Missal et al., "The expansion of the metazoan microRNA repertoire," *BMC Genomics*, vol. 7, article 25, 2006.

[51] L. F. Sempere, C. N. Cole, M. A. Mcpeek, and K. J. Peterson, "The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint," *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, vol. 306, no. 6, pp. 575–588, 2006.

[52] A. Grimson, M. Srivastava, B. Fahey et al., "Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals," *Nature*, vol. 455, no. 7217, pp. 1193–1197, 2008.

[53] F. Yu, H. Deng, H. Yao, Q. Liu, F. Su, and E. Song, "Mir-30 reduction maintains self-renewal and inhibits apoptosis in breast tumor-initiating cells," *Oncogene*, vol. 29, no. 29, pp. 4194–4204, 2010.

[54] L. Ma, J. Teruya-Feldstein, and R. A. Weinberg, "Tumour invasion and metastasis initiated by microRNA-10b in breast cancer," *Nature*, vol. 449, no. 7163, pp. 682–688, 2007.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 619438, 9 pages
http://dx.doi.org/10.1155/2015/619438

Research Article

Multi-Instance Multilabel Learning with Weak-Label for Predicting Protein Function in Electricigens

Jian-Sheng Wu,[1] Hai-Feng Hu,[2] Shan-Cheng Yan,[1] and Li-Hua Tang[1]

[1]*School of Geographic and Biological Information, Nanjing University of Posts and Telecommunications, Nanjing 210046, China*
[2]*School of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210046, China*

Correspondence should be addressed to Jian-Sheng Wu; jansen@njupt.edu.cn

Received 1 October 2014; Accepted 16 December 2014

Academic Editor: Yuedong Yang

Nature often brings several domains together to form multidomain and multifunctional proteins with a vast number of possibilities. In our previous study, we disclosed that the protein function prediction problem is naturally and inherently Multi-Instance Multilabel (MIML) learning tasks. Automated protein function prediction is typically implemented under the assumption that the functions of labeled proteins are complete; that is, there are no missing labels. In contrast, in practice just a subset of the functions of a protein are known, and whether this protein has other functions is unknown. It is evident that protein function prediction tasks suffer from *weak-label* problem; thus protein function prediction with incomplete annotation matches well with the MIML with weak-label learning framework. In this paper, we have applied the state-of-the-art MIML with weak-label learning algorithm MIMLwel for predicting protein functions in two typical real-world electricigens organisms which have been widely used in microbial fuel cells (MFCs) researches. Our experimental results validate the effectiveness of MIMLwel algorithm in predicting protein functions with incomplete annotation.

1. Introduction

Automated annotation of protein functions is challenging in the postgenomic era. With the rapid growth of the number of sequenced genomes, the overwhelming majority of protein products can only be annotated by computational approaches [1]. Nature usually brings multiple domains together to construct multidomain and multifunctional proteins with a vast number of possibilities [2]. The large part of genomic proteins, two-thirds in unicellular organisms and more than 80% in Metazoa, belongs to multidomain proteins [3]. In a multidomain protein, each domain can fulfill its own function independently, or in a coordinated manner with its neighbors [4]. Zhou and Zhang [5] proposed the Multi-Instance Multilabel learning (MIML) framework, where one object is represented by a bag of instances and the object is valid to have several labels simultaneously. Labels of training examples are known; however, labels of instances are unknown. We can regard each domain as an input instance and represent each biological function with an output label. In our previous study, it is disclosed that the protein function prediction problem is naturally and inherently MIML learning tasks [6]. Previously, prediction of protein functions was typically operated with the assumption that the functions of labeled proteins are complete; that is, there are no missing labels [7, 8]. Instead of things, in practice we just know a part of the functions of a protein, and whether this protein has other functions is unknown. Namely, these proteins have an incomplete annotation of their functions [9]. This kind of protein functions prediction problem with incomplete annotation can be referred to as the Multilabel Multi-Instance with weak-label learning task.

During the past several years, many Multilabel Multi-Instance learning algorithms have been developed [5, 10–12]. In our previous study, we proposed an ensemble MIML learning framework EnMIMLNN and design three algorithms for protein function prediction tasks by combining

the advantage of three kinds of Hausdorff distance metrics [6]. On the other hand, in the past few years, there are multiple algorithms which have been proposed for the weak-label learning problem. Sun et al. studied the weak-label learning problem in multilabel learning and proposed a method called weak-label learning (WELL) [13]. WELL deems the fact that classification boundary for each label should go across the low density regions, and any given label will not be correlative to the majority of instances [13]. Bucak et al. [14] studied the incomplete class assignment task for annotating images and proposed an approach called MLR-GR. MLR-GR optimizes the ranking errors and group Lasso loss by a convex optimization approach. Qi et al. [15] applied the Hierarchical Dirichlet Process to append missing labels for a set of images. In addition, Wang et al. [16] designed an approach for annotating weakly labeled facial images.

Although the underlying nature of predicting protein functions with incomplete annotation matches well with the Multi-Instance Multilabel with weak-label learning framework, till now there is no attempt that has been made under this learning framework. Jiang had proposed a multi-label semisupervised learning algorithm, PfunBG, to predict protein functions, employing a birelational graph (BG) of proteins and function annotations [17]. Yu et al. [7, 8] had proposed a protein function prediction method with multi-label weak-label learning (ProWL) and a variant of ProWL (ProWLIF) in order to complete the partial annotation of proteins. Both ProWL and ProWL-IF replenish the functions of proteins under the assumption that proteins are partially annotated [7, 8]. However, multilabel learning framework is evidently degenerated versions of MIML learning framework [5, 12]. Such degenerated strategies may lose useful information in the instance spaces, and this further hurts prediction performance [5, 12]. Recently, Yang et al. [18] proposed the MIMLwel (MIML with weak-label) approach which works by assuming that highly relevant labels share some common instances, and the underlying class means of bags for each label are with a large margin. MIMLwel makes use of the label relationship, and experiments had validated the effectiveness of MIMLwel in handling the Multilabel Multi-Instance with weak-label learning problem [18].

Microbial fuel cells (MFCs) are devices that can use bacterial metabolism to produce an electrical current from a wide range of organic substrates [19]. Due to the promise of sustainable energy production from organic wastes, research has intensified in the MFCs field in the last few years [19]. In this paper, we have applied the MIMLwel algorithm for annotating protein functions in two typical real-world electricigens genomes (i.e., *Geobacter sulfurreducens, Shewanella loihica PV-4*) which have been widely used in the MFCs researches. Our experimental results validate the effectiveness of MIMLwel algorithm in predicting functions of proteins in the electricigens genomes with incomplete annotation. In addition, it is worth mentioning that our approach is a general method for predicting protein functions with incomplete annotation.

2. The Formulation of the Protein Function Prediction Task with Incomplete Annotation

Nature often assembles multiple domains together to form multidomain and multifunctional proteins with high possibility, and each domain may implement its own function independently or in a cooperated manner with its neighbors. We can regard each domain as an input instance and take each biological function as an output label. Labels of the training examples are known; however, labels of instances are unknown. In our previous work, we disclose that the protein function prediction problem is naturally and inherently Multi-Instance Multilabel (MIML) learning tasks [6]. Previous studies typically predict the functions of proteins under the assumption that the functions of labeled proteins are complete; that is, there are no missing labels. In contrast, in most real cases we just know a subset of the functions of a protein, and whether this protein has other functions is unknown. Namely, these proteins have an incomplete annotation for molecular functions [9]. This type of protein function prediction problem with incomplete annotation can be inferred to as the Multilabel Multi-Instance with weak-label learning task.

We study the Multi-Instance Multilabel weak-label learning framework for protein function prediction with incomplete annotation for two tasks as illustrated in Table 1. In the tables, each row indicates the function annotation for a protein, and each column denotes a function label. Table 1(a) presents the complete annotated proteins, with 1 and 0 showing function annotations (F1–F5) on the six proteins P1–P6. In Table 1(b), 1 denotes the known relevant functions, "?" represents the missing functions and will be set to 0 s, and all the 0 s indicate the candidates for being predicted as relevant. In Task 2 as shown by Table 1(c), the definitions of 1 and 0 are the same as in Table 1(b). However, the aim of the weak-label learning is to make use of the incomplete annotated proteins (P1–P4) to predict the functions of proteins P5 and P6, which are completely unlabeled.

Formally, we represent by $\{X_i, Y_i\ (i = 1, 2, \ldots, m)\}$ the training dataset with m examples. X_i is the ith protein in the training dataset, and X_i is a bag with n_i instances $\{x_{i,1}, x_{i,2}, \ldots, x_{i,n_i}\}$. Y_i denotes the Gene Ontology terms which are assigned to X_i, and $Y_i = [y_{i,1}, \ldots, y_{i,L}] \in \{0, 1\}^L$ is a label vector with L labels, where $y_{i,l} = +1$ if the lth label is positive for X_i, and 0 otherwise. Note that the labels of instances $x_{i,j}$'s $(i = 1, \ldots, m; j = 1, \ldots, n_i)$ are untagged. In the MIML weak-label setting, Y is unknown and instead we are just given a partial label matrix $\widehat{Y} \in \{0, 1\}^{m \times L}$. Specifically, for X_i, a label vector $\widehat{Y} = [\widehat{y}_{i,1}, \ldots, \widehat{y}_{i,L}]$ is given, where $\widehat{y}_{i,l} = +1$ if the lth label is assigned for X_i, and 0 otherwise. Different from the full label matrix, $\widehat{y}_{i,l} = 0$ tells us nothing. The goal is to predict all the positive labels for unseen bags [18].

3. Datasets and Methods

3.1. Data and Feature Extraction. Microbial fuel cells (MFCs) are devices that can make use of bacterial metabolism to

Table 1: Task overview for the "*weak-label*" problem in protein function prediction tasks. "1" represents relevant function, "?" denotes missing function and will be transformed to a "0", and P5 and P6 in Table 1(c) are completely unannotated (sources from [8]).

(a) Original

	F1	F2	F3	F4	F5
P1	0	1	0	1	0
P2	0	0	1	0	1
P3	1	1	0	0	1
P4	0	1	1	0	0
P5	1	0	0	1	0
P6	0	1	0	0	0

(b) Task 1

	F1	F2	F3	F4	F5
P1	0	?	0	1	0
P2	0	0	?	?	1
P3	1	?	0	?	1
P4	?	1	1	0	0
P5	1	0	?	?	0
P6	0	1	?	0	0

(c) Task 2

	F1	F2	F3	F4	F5
P1	0	?	0	1	0
P2	0	0	?	?	1
P3	1	?	0	?	1
P4	?	1	1	0	0
P5	?	?	?	?	?
P6	?	?	?	?	?

obtain an electrical current from a wide range of organic substrates [19]. Due to the promise of sustainable energy production from organic wastes, research has booming in this field during the last few years [19]. Recently, the increased interest in MFCs technology was highlighted by the discovery of *Geobacter sulfurreducens*, a bacterial strain capable of high current production [19]. In addition, the genome-wide sequences of multiple *Shewanella* strains have been completed and annotated, opening the door to explore the diversity of their extracellular electron transfer mechanisms [20]. In this paper, two typical real-world electricigens organisms which have been widely used in microbial fuel cells (MFCs) researches (i.e., *Geobacter sulfurreducens, Shewanella loihica PV-4*) are considered for predicting their protein functions. For each organism, complete proteome with manually annotated function has been downloaded from the Universal Protein Resource (UniProt) databank [21] (released by April, 2014) by querying the terms of {"organism name" AND "reviewed: yes" AND "keyword: Complete proteome"}.

Redundancy among protein sequences of each organism is removed by clustering operation using the *blastclust* executable program in the BLAST package [22] from NCBI with a threshold of 90% as sequence identity, and a nonredundant dataset is obtained by keeping only the longest sequence in each cluster for each organism [23]. Then, each nonredundant dataset is uploaded as a *txt* file into the Batch CD-Search servers [24] of NCBI for getting the conserved domains of each protein. For each domain, a frequency vector with 216 dimensions is employed for its representation where each element indicates the frequency of a triad type [25]. Protein function can be annotated in several ways, and the most well-known and widely used one is given by Gene Ontology Consortium [26] which offers ontology in three aspects: molecular function, biological process, and cellular location. In this study, we concentrate on the molecular function aspect. We achieve the GO molecular function terms with manual annotation for a protein from the downloaded UniProt format text file. Then, the same scheme as [27] is assigned for produce label vectors for a protein based on a hierarchal directed acyclic graph (DAG) of GO molecular function, and the latest version (December 2006) of GO function ontology is adopted as the bases of the functional terms and their relations in this work.

Table 2: Characteristics of the data sets.

Organism	Examples	Classes	Instances per bag (mean ± std.)	Labels per example (mean ± std.)
Geobacter sulfurreducens	379	320	3.20 ± 1.21	3.14 ± 3.33
Shewanella loihica PV-4	373	344	3.14 ± 1.19	3.55 ± 5.00

Under the MIML learning framework, each protein is described as a bag of instances where each instance represents a domain and is tagged with a set of GO molecular function terms (multiple labels). Detailed descriptions of the datasets, that is, complete proteome on the two above organisms, are shown in Table 2. For example, there are 373 proteins (examples) with a sum of 344 gene ontology terms (label classes) on molecular function in the *Shewanella loihica PV-4* dataset (Table 2). The average number of instances (domains) per bag (protein) is 3.14 ± 1.19, and the average number of labels (GO terms) per example (protein) is 3.55 ± 5.00 (Table 2).

3.2. The MIMLwel Approach. In this paper, the MIMLwel (MIML with weak-label) approach is adopted for the weak-label setting [18]. MIMLwel assumes that highly relevant labels usually share common instances, and the underlying class means of bags for each label are separated with a large margin [18].

Formally, the training dataset with m examples can be represented by $\{X_i, Y_i \ (i = 1, 2, \ldots, m)\}$. X_i corresponds to the ith example in the training dataset, and X_i is a bag with n_i instances $\{x_{i,1}, x_{i,2}, \ldots, x_{i,n_i}\}$. Y_i denotes the labels which are assigned to X_i, and $Y_i = [y_{i,1}, \ldots, y_{i,L}] \in \{0, 1\}^L$ is a label vector with L labels, where $y_{i,l} = +1$ if the lth label is positive for X_i, and 0 otherwise. Notice that the labels of instances $x_{i,j}$'s ($i = 1, \ldots, m; j = 1, \ldots, n_i$) are unknown. In the MIML weak-label setting, however, only a subset of labels are

Table 3: Performance of the MIMLwel methods with different weak-label ratios on two datasets.

Datasets	W.L.R.	HL↓	maF1↑	miF1↑
Geobacter sulfurreducens	20%	0.010 ± 0.002	0.003 ± 0.004	0.032 ± 0.035
	40%	0.010 ± 0.002	0.009 ± 0.005	0.116 ± 0.038
	60%	0.010 ± 0.002	0.016 ± 0.006	0.201 ± 0.034
	80%	0.011 ± 0.001	**0.019 ± 0.007**	**0.245 ± 0.050**
Shewanella loihica PV-4	20%	0.013 ± 0.002	0.009 ± 0.008	0.145 ± 0.111
	40%	0.010 ± 0.002	0.005 ± 0.003	0.092 ± 0.039
	60%	0.011 ± 0.003	0.010 ± 0.006	0.167 ± 0.072
	80%	0.011 ± 0.003	**0.011 ± 0.005**	**0.186 ± 0.043**

tagged. Specifically, for X_i, a label vector $\widehat{Y} = [\widehat{y}_{i,1}, \ldots, \widehat{y}_{i,L}] \in \{0,1\}^{m\times L}$ is given, where $\widehat{y}_{i,l} = +1$ if the lth label is assigned for X_i, and 0 otherwise. The goal is to predict all the positive labels for unseen bags [18].

For simplicity, L linear models were employed, and each one is for a label; that is, $f_l(X) = w_l^T \Phi^C(X)$ where each w_l denotes a d-dimensional linear predictor $[w_{l,1}, w_{l,2}, \ldots, w_{l,d}]^T$ and w_l^T is the transpose of w_l. To make use of label relationship, a label relation matrix $R \in [0,1]^{L\times L}$ is considered, where $R_{l,\widetilde{l}} = 1$ if the two labels are related, and 0 otherwise. Let $\mathbf{W}_{\mathbf{l},\widetilde{\mathbf{l}}}$ indicate $[\mathbf{w}_\mathbf{l}, \mathbf{w}_{\widetilde{\mathbf{l}}}]$ for the pair of related labels $(l, \widetilde{l})$. MIMLwel assumes that highly related labels usually share common instances, indicating that many rows of $\mathbf{w}_{\mathbf{l},\widetilde{\mathbf{l}}}$ values should be equal to zero; this can be characterized by a convexly relaxed term $\|\mathbf{w}(\mathbf{l},\widetilde{\mathbf{l}})\|_{(\mathbf{2},\mathbf{1})}$, which is a convex relaxation of $\|\mathbf{w}(\mathbf{l},\widetilde{\mathbf{l}})\|_{(\mathbf{2},\mathbf{0})}$. Thus, the goal of MIMLwel is to obtain $W = [w_1, \ldots, w_L]$ and an output matrix $\widehat{Y}$ to meet that

$$\begin{aligned} \min_{\mathbf{w},\overline{\mathbf{Y}}} \quad & -\eta \sum_{l=1}^{L} V\left(\left\{\overline{y}_{i,l}, X_i\right\}_{i=1}^{m}, \mathbf{w}_\mathbf{l}\right) + \sum_{\mathbf{1}<\mathbf{l},\widetilde{\mathbf{l}}\le\mathbf{L}} \mathbf{R}_{\mathbf{l},\widetilde{\mathbf{l}}} \left\|\mathbf{W}_{\mathbf{l},\widetilde{\mathbf{l}}}\right\|_{\mathbf{2},\mathbf{1}}^{\mathbf{2}} \\ \text{s.t.} \quad & \frac{\left|\overline{Y}_l - \widehat{Y}_l\right|_1}{\left|\widehat{Y}_l\right|_1} \le \epsilon; \\ & \overline{y}_{i,l} = \widehat{y}_{i,l} \quad \text{if } \widehat{y}_{i,l} = 1, \ \forall l = 1, \ldots, L, \end{aligned} \tag{1}$$

where V is a loss function for each label, $|\cdot|_1$ represents the l_1-norm, ϵ controls the sparsity of $|\overline{Y}_l - \widehat{Y}_l|_1$, and η trades off the empirical risk and model complexity.

3.3. Experimental Configuration. In this paper, we adopt three popular multilabel learning evaluation criteria, that is, *Hamming loss (HL), macro-F1 (maF1), and micro-F1 (miF1)* [28–30]. *Hamming loss* assesses how many times on average a bag label pair is wrongly predicted. The smaller the value of hamming loss, the better the performance. *Macro-F1* computes *F1* measure on each class label at first and then averages over all class labels. *Macro-F1* is more influenced by the performance of the classes owning fewer examples. The larger the value of *macro-F1*, the better the performance. *Micro-F1* globally calculates the *F1* measure on the predictors over all bags and all class labels. *Micro-F1* is more affected by the performance of the classes involving more examples. The larger the value of *micro-F1*, the better the performance. The definition of these criteria can be found in [30]. We repeat 10-fold cross validation for each dataset ten times and the mean ± std. performances are presented for the proposed and compared methods.

4. Results and Discussion

4.1. Performance of the MIMLwel Method. In our experiments we consider four weak-label ratios (W.L.R.) [18], defined as $|\widetilde{\mathbf{Y}}_{\cdot,l}|_1/|\mathbf{Y}_{\cdot,\mathbf{1}}|_\mathbf{1}$, from 20% to 80% with 20% as the interval. Table 3 illustrates the performances of MIMLwel based on each kind of W.L.R. on the *Geobacter sulfurreducens* and *Shewanella loihica PV-4* datasets. For each evaluation criterion, ↑(↓) indicates the larger (smaller), the better the performance; the best results on each evaluation criterion are highlighted in boldface. As indicated in Table 3, the results show that, with the rising of W.L.R., the model performance of MIMLwel has been greatly improved.

The MIMLwel approach [18] involves two different parameters, that is, the scaling factor μ and the fraction parameter α. Figure 1 shows how the MIMLwel algorithm is implemented on the two datasets with 80% weak-label ratios (W.L.R.) under different parameter configurations, where the performance is measured in terms of *HL, maF1,* and *miF1.* Here, μ varies from 0.2 to 1.0 with an interval of 0.2 when α is fixed to 0.1, and α increases from 0.02 to 0.1 with an interval of 0.02 with the fixed μ equal to 1.0. It is indicated that the performance of the MIMLwel algorithms achieves the perk in most cases by setting the scaling factor μ to 1.0 and the fraction parameter α to 0.1. In this paper, the MIMLwel algorithm is implemented by setting the scaling factor μ to 1.0 and the fraction parameter α to 0.1.

4.2. Performance Comparison. In this paper, we compare the MIMLwel algorithm with four state-of-the-art MIML algorithms, that is, MIMLkNN [31], MIMLNN [12], MIMLRBF [32], and MIMLSVM [5], under different configuration of weak-label ratios (W.L.R.) on the *Geobacter sulfurreducens* dataset (Table 4) and *Shewanella loihica PV-4* dataset (Table 5). The codes of compared MIML algorithms are shared by their authors, and these algorithms are implemented using the best parameters reported in the papers. Specifically, for MIMLkNN, the number of nearest neighbors and the number of citers are set to 10 and 20, respectively [31]; for MIMLNN, the number of clusters is set to 40% of

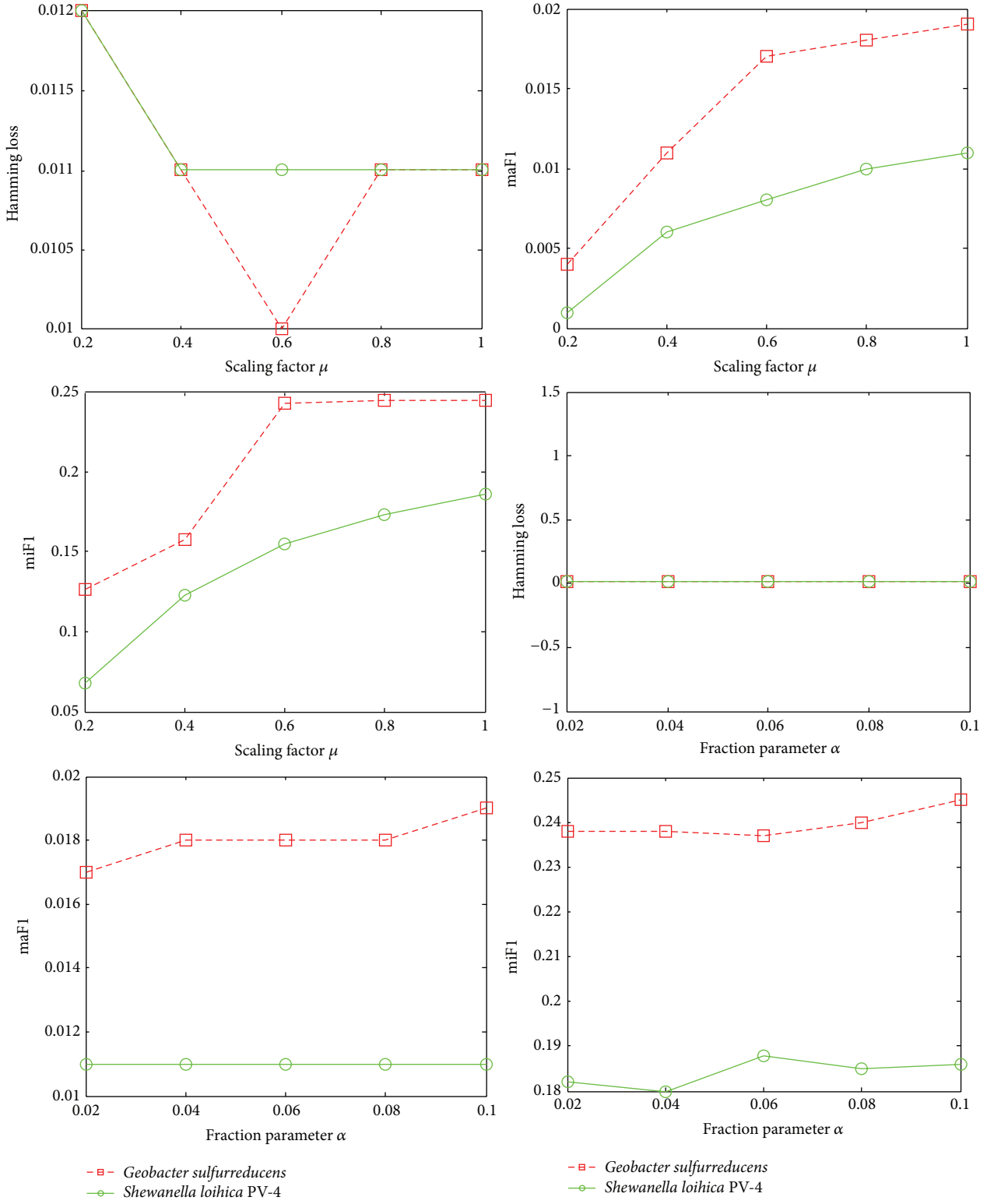


Figure 1: The performance of MIMLwel on all two datasets with 80% weak-label ratios (W.L.R.) under different values of scaling factor μ when the fraction parameter α is fixed to 0.1 and different values of the fraction parameter α when the scaling factor μ is fixed to 1.0. The performance of MIMLwel reaches the perk in most cases by setting the scaling factor μ to 1.0 and the fraction parameter α to 0.1.

Table 4: Comparison results (mean ± std.) of MIMLwel models with four state-of-the-art MIML methods with different weak-label ratios on the *Geobacter sulfurreducens* dataset.

W.L.R.	Methods	HL↓	maF1↑	miF1↑
20%	MIMLwel	**0.010 ± 0.002**	0.003 ± 0.004	**0.032 ± 0.035**
	MIMLNN	**0.010 ± 0.002**	0.000 ± 0.000	0.000 ± 0.000 ●
	MIMLRBF	**0.010 ± 0.002**	0.002 ± 0.003	0.002 ± 0.003 ●
	MIMLSVM	0.012 ± 0.002	**0.005 ± 0.003**	0.005 ± 0.003 ●
	EnMIMLNN {metric}	**0.010 ± 0.002**	0.002 ± 0.002	0.001 ± 0.002 ●
40%	MIMLwel	**0.010 ± 0.002**	**0.009 ± 0.005**	**0.116 ± 0.038**
	MIMLNN	**0.010 ± 0.002**	0.000 ± 0.000	0.000 ± 0.000 ●
	MIMLRBF	**0.010 ± 0.002**	0.004 ± 0.004	0.003 ± 0.003 ●
	MIMLSVM	0.012 ± 0.001	0.006 ± 0.003	0.006 ± 0.003 ●
	EnMIMLNN {metric}	**0.010 ± 0.002**	0.003 ± 0.004	0.003 ± 0.003 ●
60%	MIMLwel	0.010 ± 0.002	**0.016 ± 0.006**	**0.201 ± 0.034**
	MIMLNN	0.010 ± 0.001	0.001 ± 0.001	0.001 ± 0.001 ●
	MIMLRBF	**0.009 ± 0.001**	0.009 ± 0.007	0.008 ± 0.007 ●
	MIMLSVM	0.011 ± 0.001	0.008 ± 0.003	0.008 ± 0.003 ●
	EnMIMLNN {metric}	0.010 ± 0.001	0.009 ± 0.004	0.008 ± 0.004 ●
80%	MIMLwel	0.011 ± 0.001	**0.019 ± 0.007**	**0.245 ± 0.050**
	MIMLNN	0.010 ± 0.001	0.002 ± 0.001 ●	0.002 ± 0.001 ●
	MIMLRBF	**0.009 ± 0.000**	0.009 ± 0.004 ●	0.008 ± 0.004 ●
	MIMLSVM	0.011 ± 0.001	0.008 ± 0.002 ●	0.008 ± 0.002 ●
	EnMIMLNN {metric}	**0.009 ± 0.001**	0.013 ± 0.004	0.012 ± 0.004 ●

the training bags, and the regularization parameter used to compute matrix inverse is set to 1 [12]; for MIMLRBF, the scaling factor and the fraction parameter are set to 0.6 and 0.1, respectively [32]; for MIMLSVM, the number of clusters is set to 20% of the training bags and the Gaussian kernel width is set to 0.2 [5]. Tables 4 and 5 summarize the experimental results of each compared algorithm on the *Geobacter sulfurreducens* dataset and *Shewanella loihica PV-4* dataset, respectively. For each evaluation criterion, "↓" indicates "the smaller the better," while "↑" indicates "the bigger the better." Furthermore, the best results on each evaluation criterion are highlighted in boldface. It is indicated that the MIMLwel algorithm performs quite well in terms of most criteria in two datasets (Tables 5 and 6). Specifically, paired t-tests at 95% significance level indicate that the MIMLwel algorithm achieves significantly better performance than compared methods in most cases, as shown by the overwhelming ●'s in Tables 4 and 5.

4.3. Case Study. Table 6 presents two example results. The first protein with the UniProt ID "Q74BW7" from the *Geobacter sulfurreducens* organism has seven ground-truth labels: {GO:0008270, GO:0046872, GO:0000287, GO:0051539, GO:0030145, GO:0005506, GO:0004160}. After training examples with 80% weak-label ratios by different MIML methods, the trained model is then used to predict the GO molecular function labels of this protein. The correctly predicted GO molecular function labels by each method are highlighted in boldface. It is shown in Table 6 that MIMLwel successfully predicts most of the ground-truth labels (6/7); however, it predicts one more label, that is, GO:0005524, which is not in the ground-truth list. Nevertheless, the label GO:0005524 that denotes "ATP binding" may be not a conflict with the true molecular function in UniProt. MIMLRBF and EnMIMLNN{metric} predict two ground-truth labels but still miss a lot (5/7). MIMLNN reports no prediction result, and MIMLSVM only reports a wrong GO molecular function label. Similar situation also happen in the second example with the UniProt ID "A3QFX5" from the *Shewanella loihica PV-4* organism as indicated in Table 6.

5. Conclusion

In our previous study, we disclosed that the protein function prediction problem is naturally and inherently Multi-Instance Multilabel (MIML) learning tasks. Automated protein function prediction was typically implemented under the assumption that the functions of labeled proteins are complete; that is, there are no missing labels. In contrast, in practice just a subset of the functions of a protein are known, and whether this protein has additional functions is unknown. It is evident that the protein function prediction tasks suffer from weak-label problems, and we disclose that prediction of protein functions with incomplete annotation matches well with the MIML with weak-label learning framework in this paper. In this paper, we have applied the state-of-the-art MIML with weak-label learning algorithm MIMLwel for predicting protein function in two typical real-world electricigens organisms which have been widely used in microbial fuel cells (MFCs) researches. Our experimental results show that MIMLwel is superior to most state-of-the-art MIML algorithms, which validates the effectiveness of

Table 5: Comparison results (mean ± std.) of MIMLwel models with four state-of-the-art MIML methods with different weak-label ratios on the *Shewanella loihica PV-4* dataset.

W.L.R.	Methods	HL↓	maF1↑	miF1↑
20%	MIMLwel	0.013 ± 0.002	**0.009 ± 0.008**	**0.145 ± 0.111**
	MIMLNN	**0.010 ± 0.002**	0.000 ± 0.000	0.000 ± 0.000 ●
	MIMLRBF	0.011 ± 0.003	0.001 ± 0.001	0.001 ± 0.001 ●
	MIMLSVM	0.012 ± 0.002	0.005 ± 0.002	0.004 ± 0.002 ●
	EnMIMLNN {metric}	**0.010 ± 0.003**	0.001 ± 0.001	0.001 ± 0.001 ●
40%	MIMLwel	**0.010 ± 0.002**	**0.005 ± 0.003**	**0.092 ± 0.039**
	MIMLNN	**0.010 ± 0.002**	0.000 ± 0.000	0.000 ± 0.000 ●
	MIMLRBF	**0.010 ± 0.002**	0.001 ± 0.002	0.001 ± 0.002 ●
	MIMLSVM	0.012 ± 0.002	0.004 ± 0.002	0.004 ± 0.002 ●
	EnMIMLNN {metric}	**0.010 ± 0.002**	0.001 ± 0.003	0.001 ± 0.003 ●
60%	MIMLwel	0.011 ± 0.003	**0.010 ± 0.006**	**0.167 ± 0.072**
	MIMLNN	**0.010 ± 0.003**	0.001 ± 0.001	0.001 ± 0.001 ●
	MIMLRBF	**0.010 ± 0.004**	0.004 ± 0.004	0.003 ± 0.003 ●
	MIMLSVM	0.012 ± 0.003	0.005 ± 0.001	0.005 ± 0.002 ●
	EnMIMLNN {metric}	**0.010 ± 0.003**	0.005 ± 0.003	0.004 ± 0.003 ●
80%	MIMLwel	0.011 ± 0.003	**0.011 ± 0.005**	**0.186 ± 0.043**
	MIMLNN	0.010 ± 0.003	0.002 ± 0.001	0.001 ± 0.001 ●
	MIMLRBF	**0.009 ± 0.003**	0.008 ± 0.005	0.007 ± 0.005 ●
	MIMLSVM	0.012 ± 0.003	0.005 ± 0.002	0.005 ± 0.001 ●
	EnMIMLNN {metric}	0.010 ± 0.003	0.006 ± 0.004	0.005 ± 0.003 ●

Table 6: Comparison results on two examples.

Organism/UniProt ID	Molecular function in UniProt	Methods	GO molecular function list		
Geobacter sulfurreducens/ Q74BW7	(1) 4 iron, 4 sulfur cluster binding (2) Dihydroxy-acid dehydratase activity (3) Metal ion binding	Ground truth	GO:0008270	GO:0046872	GO:0000287
			GO:0051539	GO:0030145	GO:0005506
			GO:0004160		
		MIMLwel	GO:0005524	**GO:0008270**	**GO:0046872**
			GO:0000287	**GO:0030145**	**GO:0005506**
		MIMLNN		Null	
		MIMLRBF	**GO:0000287**	**GO:0005506**	
		MIMLSVM	GO:0050567		
		EnMIMLNN {metric}	**GO:0000287**	**GO:0005506**	
Shewanella loihica PV-4/A3QFX5	(1) ATP binding (2) Nucleoside-triphosphatase activity (3) Zinc ion binding	Ground truth	GO:0003924	GO:0005524	GO:0004386
			GO:0008270	GO:0016887	GO:0046961
			GO:0005215	GO:0017111	GO:0004004
			GO:0008094	GO:0008565	
		MIMLwel	**GO:0005524**	**GO:0004386**	**GO:0016887**
			GO:0046961	**GO:0004004**	**GO:0008094**
			GO:0043565		
		MIMLNN		Null	
		MIMLRBF	**GO:0005524**	**GO:0004386**	**GO:0016887**
			GO:0046961	**GO:0004004**	**GO:0008094**
		MIMLSVM	**GO:0008270**		
		EnMIMLNN {metric}	**GO:0005524**	**GO:0016887**	**GO:0004004**

MIMLwel algorithm in predicting protein functions with incomplete annotation.

Conflict of Interests

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

Acknowledgments

This research was supported by the National Science Foundation of China (61203289, 61071092, and 61205057), China Postdoctoral Science Foundation (20110490129, 2013T60523), and Natural Science Foundation of the Higher Education Institutions of Jiangsu Province, China (12KJB520010).

References

[1] P. Radivojac, W. T. Clark, T. R. Oron et al., "A large-scale evaluation of computational protein function prediction," *Nature Methods*, vol. 10, no. 3, pp. 221–227, 2013.

[2] C. Chothia, "One thousand families for the molecular biologist," *Nature*, vol. 357, no. 6379, pp. 543–544, 1992.

[3] G. Apic, J. Gough, and S. A. Teichmann, "Domain combinations in archaeal, eubacterial and eukaryotic proteomes," *Journal of Molecular Biology*, vol. 310, no. 2, pp. 311–325, 2001.

[4] C. J. Tsai and R. Nussinov, "Hydrophobic folding units derived from dissimilar monomer structures and their interactions," *Protein Science*, vol. 6, no. 1, pp. 24–42, 1997.

[5] Z.-H. Zhou and M.-L. Zhang, "Multi-instance multi-label learning with application to scene classification," in *Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS '06)*, pp. 1609–1616, December 2006.

[6] J.-S. Wu, S.-J. Huang, and Z.-H. Zhou, "Genome-wide protein function prediction through multi-instance multi-label learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 5, pp. 831–902, 2014.

[7] G. Yu, H. Rangwala, C. Domeniconi, G. Zhang, and Z. Yu, "Protein function prediction with incomplete annotations," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 3, pp. 579–591, 2013.

[8] G. Yu, G. Zhang, H. Rangwala, C. Domeniconi, and Z. Yu, "Protein function prediction using weak-label learning," in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pp. 202–209, October 2012.

[9] L. Peña-Castillo, M. Tasan, C. L. Myers et al., "A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence," *Genome Biology*, vol. 9, supplement 1, article S2, 2008.

[10] F. Briggs, X. Z. Fern, and R. Raich, "Rank-loss support instance machines for MIML instance annotation," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*, pp. 534–542, Beijing, China, August 2012.

[11] S.-H. Yang, H. Zha, and B.-G. Hu, "Dirichlet-bernoulli alignment: a generative model for multi-class multi-label multi-instance corpora," in *Proceedings of the 23th Annual Conference on Neural Information Processing Systems*, pp. 2143–2150, MIT Press, Vancouver, Canada, 2009.

[12] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artificial Intelligence*, vol. 176, no. 1, pp. 2291–2320, 2012.

[13] Y.-Y. Sun, Y. Zhang, and Z.-H. Zhou, "Multi-label learning with weak label," in *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI '10)*, pp. 593–598, 2010.

[14] S. S. Bucak, R. Jin, and A. K. Jain, "Multi-label learning with incomplete class assignments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 2801–2808, Providence, RI, USA, June 2011.

[15] Z. Qi, M. Yang, Z. M. Zhang, and Z. Zhang, "Mining partially annotated images," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*, pp. 1199–1207, ACM, August 2011.

[16] D. Wang, S. C. H. Hoi, Y. He, and J. Zhu, "Mining weakly labeled web facial images for search-based face annotation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 166–179, 2014.

[17] J. Q. Jiang, "Learning protein functions from bi-relational graph of proteins and function annotations," in *Algorithms in Bioinformatics*, pp. 128–138, Springer, New York, NY, USA, 2011.

[18] S.-J. Yang, Y. Jiang, and Z.-H. Zhou, "Multi-instance multi-label learning with weak label," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI '13)*, pp. 1862–1868, AAAI Press, August 2013.

[19] A. E. Franks and K. P. Nevin, "Microbial fuel cells, a current review," *Energies*, vol. 3, no. 5, pp. 899–919, 2010.

[20] G. J. Newton, S. Mori, R. Nakamura, K. Hashimoto, and K. Watanabe, "Analyses of current-generating mechanisms of *Shewanella loihica* PV-4 and *Shewanella oneidensis* MR-1 in microbial fuel cells," *Applied and Environmental Microbiology*, vol. 75, no. 24, pp. 7674–7681, 2009.

[21] R. Apweiler, A. Bairoch, C. H. Wu et al., "UniProt: the universal protein knowledgebase," *Nucleic Acids Research*, vol. 32, pp. D115–D119, 2004.

[22] 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.

[23] J. Wu, H. Liu, X. Duan et al., "Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature," *Bioinformatics*, vol. 25, no. 1, pp. 30–35, 2009.

[24] 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, supplement 1, pp. D225–D229, 2011.

[25] J. Wu, D. Hu, X. Xu, Y. Ding, S. Yan, and X. Sun, "A novel method for quantitatively predicting non-covalent interactions from protein and nucleic acid sequence," *Journal of Molecular Graphics and Modelling*, vol. 31, pp. 28–34, 2011.

[26] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology. The gene ontology consortium," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[27] Ö. S. Saraç, V. Atalay, and R. Cetin-Atalay, "GOPred: GO molecular function prediction by combined classifiers," *PLoS ONE*, vol. 5, no. 8, Article ID e12382, 2010.

[28] N. Ghamrawi and A. McCallum, "Collective multi-label classification," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05)*, pp. 195–200, ACM, November 2005.

[29] M. Rogati and Y. Yang, "High-performing feature selection for text classification," in *Proceedings of the 11th International Conference on Information and Knowledge Management*, pp. 659–661, ACM, 2002.

[30] S.-J. Yang, Y. Jiang, and Z.-H. Zhou, "Multi-Instance multi-label learning with weak label," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI '13)*, pp. 1862–1868, August 2013.

[31] M.-L. Zhang, "A k-nearest neighbor based multi-instance multi-label learning algorithm," in *Proceedings of the 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI '10)*, vol. 2, pp. 207–212, Arras, France, October 2010.

[32] M.-L. Zhang and Z.-J. Wang, "MIMLRBF: RBF neural networks for multi-instance multi-label learning," *Neurocomputing*, vol. 72, no. 16–18, pp. 3951–3956, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 821596, 7 pages
http://dx.doi.org/10.1155/2015/821596

Research Article

Nucleosome Organization around Pseudogenes in the Human Genome

Guoqing Liu,[1,2] Fen Feng,[2] Xiujuan Zhao,[1,2] and Lu Cai[1,2]

[1] *The Institute of Bioengineering and Technology, Inner Mongolia University of Science and Technology, Baotou 014010, China*
[2] *School of Mathematics, Physics, and Biological Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China*

Correspondence should be addressed to Guoqing Liu; gqliu1010@163.com and Lu Cai; nmcailu@163.com

Received 9 September 2014; Accepted 17 December 2014

Academic Editor: Elena Papaleo

Pseudogene, disabled copy of functional gene, plays a subtle role in gene expression and genome evolution. The first step in deciphering RNA-level regulation of pseudogenes is to understand their transcriptional activity. So far, there has been no report on possible roles of nucleosome organization in pseudogene transcription. In this paper, we investigated the effect of nucleosome positioning on pseudogene transcription. For transcribed pseudogenes, the experimental nucleosome occupancy shows a prominent depletion at the regions both upstream of pseudogene start positions and downstream of pseudogene end positions. Intriguingly, the same depletion is also observed for nontranscribed pseudogenes, which is unexpected since nucleosome depletion in those regions is thought to be unnecessary in light of the nontranscriptional property of those pseudogenes. The sequence-dependent prediction of nucleosome occupancy shows a consistent pattern with the experimental data-based analysis. Our results indicate that nucleosome positioning may play important roles in both the transcription initiation and termination of pseudogenes.

1. Introduction

Pseudogenes are produced from protein-coding genes during evolution. Though highly homologous with their parent genes, pseudogenes are unable to synthesize functional protein due to the defects in their sequences. There are two major types of pseudogenes: duplicated pseudogenes and processed pseudogenes (or retropseudogenes). The former type is created by genomic duplication and the latter by retrotransposition [1, 2]. For each type, the abnormalities occurred in either the protein-coding regions or the regulatory regions of parent genes leading to the loss of protein-coding ability of pseudogenes. Duplicated pseudogenes are often distributed in the flanking of the parent genes and may still maintain the upstream regulatory sequences of their parents due to their duplicative origin. Processed pseudogenes are usually characterized by absence of intron-like segments, decayed poly A tail, frame shifts, and premature stop codons. Processed pseudogenes are thought to be nonautonomous retrotransposons which are probably mobilized by long interspersed elements (LINEs), a kind of autonomous retrotransposons in the genome [3, 4]. Processed pseudogenes occur in a great number of eukaryotes, especially in mammalian genomes [5, 6].

Many unexpected discoveries of biological functions for pseudogenes challenge the popular belief that pseudogenes are nonfunctional and simply molecular fossils. A nitric oxide synthase (NOS) pseudogene functions as a regulator of the paralogous protein-coding neuronal nitric oxide synthase (nNOS) gene by producing antisense RNA that forms a duplex with some of the gene's mRNA [7, 8]. The Makorin1-p1 pseudogene in mouse regulates the stability of the mRNA of its homologous Makorin1 gene probably by producing RNA which competes for the freely available repressor molecules that inhibit the homologous gene expression [9]. Some pseudogenes can also compete with their parent genes for microRNA binding, thereby modulating the repression of the functional gene by its cognate miRNA [10]. The transcription of MYLKP1 pseudogene, which is upregulated in cancer cells,

creates a noncoding RNA (ncRNA) that inhibits the mRNA expression of its parent MYLK gene [11]. Moreover, recent studies have documented that a subset of pseudogenes generates endogenous small interfering RNAs (endo-siRNAs) and suppresses gene expression by means of the RNA interference pathway in mouse oocytes [12, 13], subsequently in rice [14], most lately in African Trypanosoma brucei [15], a unicellular eukaryote. These observations suggested that pseudogenes might be an alternative source of natural antisense transcripts that regulate the activity of sense transcripts of their parent genes. Besides, pseudogenes may have a whole set of functions related to intracellular immunobiology [2, 16, 17].

The variety of known or suspected pseudogene functions discovered to date suggests that pseudogenes as a whole have a wide range of previously unsuspected functions. Of the functions, RNA-level functions are of great importance and are most frequently discussed. The prerequisite of understanding the RNA-level functions of pseudogenes is to explore their transcriptional activity. It has been shown that the nucleosome, a fundamental composing unit of the chromatin structure in eukaryotes, affects gene transcription in that it modulates the accessibility of underlying genomic sequence to proteins [18]. How does nucleosome positioning affect pseudogene transcription? Seeking to answer the question, we analyze the nucleosome organization around the pseudogenes in human. Nucleosome occupancy is measured by both a sequence-dependent computational model and experimental data [19]. The computational model emphasizes the sequence-dependency of nucleosome positioning, while the nucleosome occupancy inferred from in vivo experimental data reflects the joint effect of DNA sequence and other external factors, such as chromatin remodeler, DNA methylation, histone modification, and polymerase II binding, on nucleosome positioning [19–21]. The two methods may have different implications for the dependency of pseudogene transcription on chromatin structure.

2. Materials and Methods

2.1. Materials

2.1.1. Transcribed and Nontranscribed Pseudogenes. A total of 201 consensus pseudogenes, including 124 processed pseudogenes and 77 duplicated pseudogenes, were identified in ENCODE regions [22]. Of the ENCODE pseudogenes, 38 pseudogenes have evidence of transcription, and others are considered to be nontranscribed. The sequences and annotation information (genomic position, strand, and positions of start positions and end positions) of the pseudogenes mapping to the human genome (hg18) were retrieved from UCSC (http://www.genome.ucsc.edu/). The type and transcriptional information of the pseudogenes were downloaded from the pseudogene database (http://www.pseudogene.org/). The number of transcribed pseudogenes in ENCODE regions is too small, so we refer to the genome-wide transcribed processed pseudogenes that were identified by Harrison et al. [23]. The annotation of the 192 transcribed processed pseudogenes that corresponds to the human genome (version hg18) was taken from the pseudogene database (http://www.pseudogene.org/). The transcribed processed pseudogenes were identified by mapping three sources of expressed sequences (Refseq mRNAs, Unigene consensuses, and ESTs from dbEST) onto the processed pseudogenes. Oligonucleotide microarray data was used to further verify the expression of the selected transcribed pseudogenes [23]. The sequences surrounding the start sites and end sites of the transcribed pseudogenes were retrieved from the human complete genome (hg18) by using the positional information of the pseudogenes. The statistics of the pseudogenes are listed in Table 1.

Table 1: The statistics of pseudogenes.

	Transcribed	Nontranscribed
Processed	192	106
Duplicated	0	57
Total	192	163

2.1.2. Human Nucleosome Occupancy. Experimental data-based nucleosome occupancy profile mapping to the human genome (hg18) was taken from Schones et al. [19]. It was based on maps of nucleosome positions in both resting and activated human CD4+ T cells generated by direct sequencing of nucleosome ends using the Solexa high-throughput sequencing technique. The two nucleosome profiles (resting and activated) have a resolution of 10 bp. We applied cubic spline fitting to each of the profiles to obtain nucleosome occupancy at each genomic site. We also estimated nucleosome occupancy by a sequence-dependent computational model described in detail in the Methods section.

2.2. Methods

2.2.1. Conformational Energy Calculation. Conformational energy is to be calculated on the basis of the geometrical description of DNA double helix structure. According to Cambridge Convention [24], each base pair of DNA is viewed as a rigid board, and its position relevant to its neighbor is specified by roll, tilt, twist, slide, shift, and rise. Nucleosomal DNA bending appeared to be due to periodic variations in both roll and tilt in the crystal structure 1kx5 [18]. The periodic changes reflected the helix twisting that altered the rotational position of each base-pair step (or dinucleotide step) relative to the dyad. In addition to the general trend of periodic changes, variations in the roll and tilt at each base-pair step were also dependent on the property of individual dinucleotide.

Nucleosomal DNA deformation is viewed as forced bending. It is assumed that torque F_b is uniformly distributed along the DNA. We consider DNA bending to be analogous to the bending of a rod of multiple segments with variable stiffness. For a bending force exerted by the histone octamer on a segment of the DNA, the conformational energy at each step along the sequence depends on both the corresponding dinucleotide flexibility and the phasing of the dinucleotide with respect to the dyad. According to simple elastic model,

deformations of roll and tilt from their equilibrium values at dinucleotide step i are described as

$$\rho(i) - \rho_0(i) = \frac{F_b \cos\Omega_i}{k_\rho(i)}, \qquad \tau(i) - \tau_0(i) = \frac{F_b \sin\Omega_i}{k_\tau(i)}. \tag{1}$$

The bending energy is then calculated by

$$E_b(i) = \frac{1}{2}k_\rho(i)\left[\rho(i) - \rho_0(i)\right]^2 + \frac{1}{2}k_\tau(i)\left[\tau(i) - \tau_0(i)\right]^2 = \frac{F_b^2}{2k_\rho(i)}\cos^2\Omega_i + \frac{F_b^2}{2k_\tau(i)}\sin^2\Omega_i, \tag{2}$$

where $\rho(i)$ and $\tau(i)$ are, respectively, the actual roll and tilt angle at dinucleotide step i, $\rho_0(i)$ and $\tau_0(i)$, which are dependent on the dinucleotide at step i, are, respectively, the roll and tilt without torque, $k_\rho(i)$ and $k_\tau(i)$ are the dinucleotide-dependent force constants, and Ω_i is the accumulated twist (ω) at the center of step i, counted from the dyad position. For 147 bp nucleosomal core DNA, its structure is symmetrical with respect to the dyad that is located at the centeral nucleotide, and the dinucleotide steps from the dyad are labeled as $i = \pm1, \pm2, \pm3, \ldots, \pm73$ towards downstream and upstream directions. The step ±1 is half step away from the dyad; thus the accumulated twist is calculated as follows:

$$\Omega_i = \begin{cases} 0.5\omega_1 + \sum_{2}^{i}\omega_i, & \text{if } i > 0, \\ -\left(0.5\omega_{-1} + \sum_{i}^{-2}\omega_i\right), & \text{if } i < 0. \end{cases} \tag{3}$$

The bending energy for the central L-bp segment of a nucleosomal DNA is the sum of corresponding dinucleotide steps:

$$E_b = \sum_{-(L-1)/2}^{(L-1)/2} E_b(i) = \sum_{-(L-1)/2}^{(L-1)/2}\left[\frac{F_b^2}{2k_\rho(i)}\cos^2\Omega_i + \frac{F_b^2}{2k_\tau(i)}\sin^2\Omega_i\right], \tag{4}$$

where L is a positive odd number and less than or equal to 147.

In (4), F_b is determined by utilizing its relationship with the total bending angle of the core DNA. In the crystal structure of core particles, about 10 bp at each end has no contact with the histone octamers, and therefore the sequence dependency of nucleosome positioning is reflected merely in the central 129 bp part of the nucleosomal DNA. The central 129 bp part of the nucleosomal core DNA bends around histone octamer about 579° (α) under the stress of F_b, and the α is due to contribution of ρ and τ at every step:

$$\alpha = \sum_{i=-64}^{64}\left[\rho(i)\cos\Omega_i + \tau(i)\sin\Omega_i\right]. \tag{5}$$

Table 2: The dinucleotide-dependent force constants and parameters ρ_0 and τ_0.

Step	k_ρ	k_τ	ρ_0	τ_0
AA/TT	0.2	0.406	0.76	−1.84
AT	0.124	0.641	−1.39	0
AG/CT	0.077	0.28	3.15	−1.48
AC/GT	0.085	0.302	0.91	−0.64
TA	0.064	0.365	5.25	0
TG/CA	0.059	0.393	5.95	−0.05
TC/GA	0.097	0.408	3.87	−1.52
GG/CC	0.075	0.218	3.86	0.4
GC	0.057	0.256	0.67	0
CG	0.04	0.255	4.25	0

Combining (1) and (5) leads to

$$F_b = \frac{\alpha - \sum_i \rho_0(i)\cos\Omega_i - \sum_i \tau_0(i)\sin\Omega_i}{\sum_i\left(\cos^2\Omega_i/k_\rho(i)\right) + \sum_i\left(\sin^2\Omega_i/k_\tau(i)\right)}. \tag{6}$$

The empirical parameters of our model for conformational energy calculation consist of force constants (k_ρ and k_τ) and roll and tilt angles (ρ_0 and τ_0) for 10 dinucleotides at the equilibrium state (Table 2). The dinucleotide-dependent parameters ρ_0 and τ_0 averaged over a large pool of DNA-protein complexes and force constants k_ρ and k_τ are taken from the paper of Morozov et al. [25]. A constant $\omega = 34.8°$, average twist for the 1kx5 X-ray crystal structure of nucleosome-bound DNA, was used for all dinucleotide steps.

2.2.2. Nucleosome Occupancy Estimation. According to Boltzmann distribution, the potential of forming a nucleosome which centers at position j in a DNA segment of N bp is defined as

$$S_j = e^{-\beta E_j}, \tag{7}$$

where $\beta = 1/k_BT$, k_B is Boltzmann constant, T is the room temperature, $M = 147$ (nucleosome size), and E_j is the deformation energy of the underlying DNA of the nucleosome which occupies positions $j-(M-1)/2$ through $j+(M-1)/2$. For simplicity, we assume $\beta = 1$ in calculation. Nucleosome occupancy at the base-pair position j is measured by the average of the nucleosome formation potentials over l-bp window:

$$O_j = \frac{\sum_{i=j-(l-1)/2}^{j+(l-1)/2} S_i}{l}. \tag{8}$$

In this study, $l = 51$, of which performance was validated in our other study (unpublished).

Normalized nucleosome occupancy at every base-pair is calculated by the log-ratio between the corresponding absolute nucleosome occupancy O_i and the average nucleosome occupancy $\langle O_i\rangle$ per base-pair across the genome as

$$O_j^{\text{nor}} = \log\frac{O_j}{\langle O_j\rangle}. \tag{9}$$

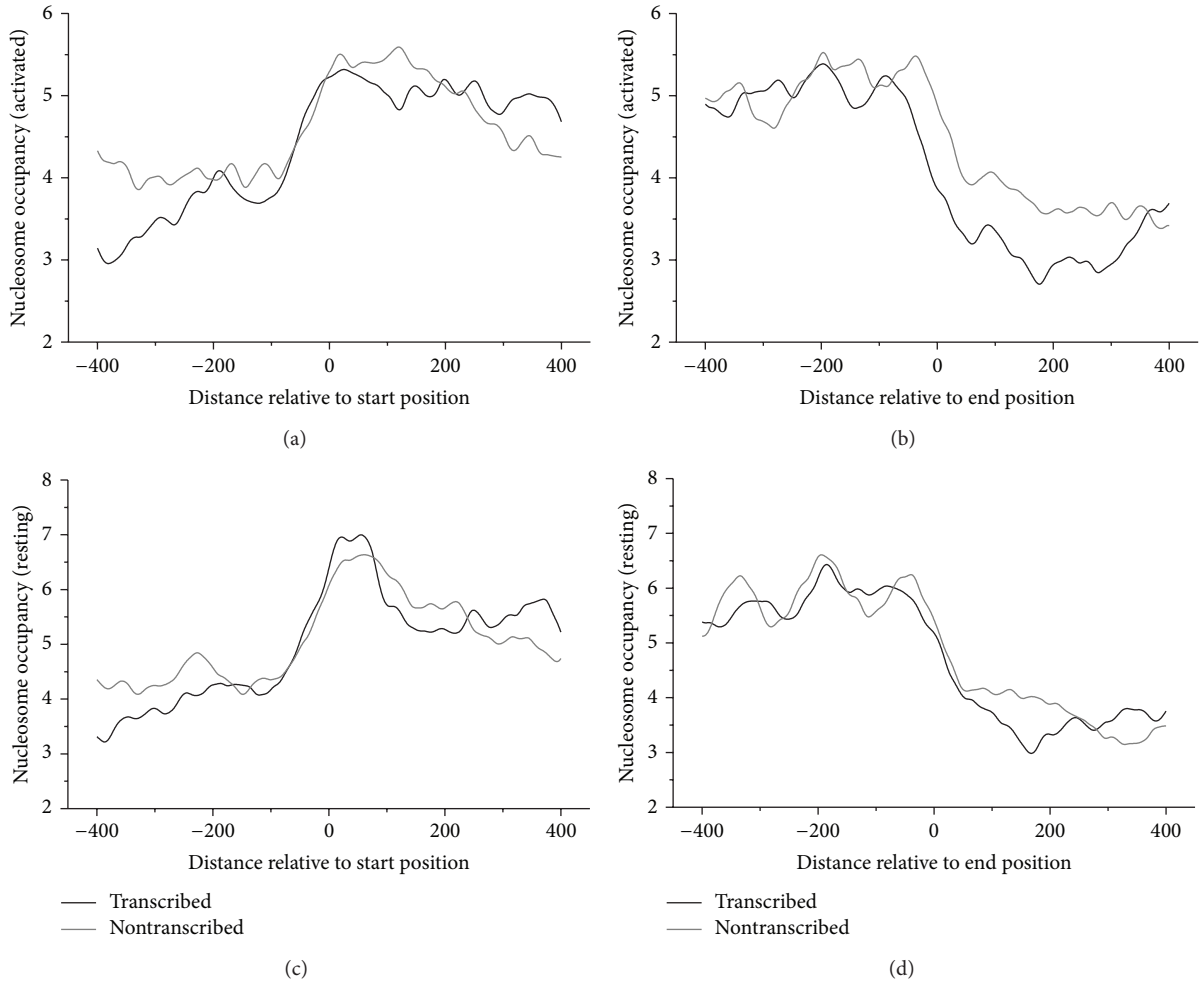


Figure 1: Experimental nucleosome occupancy around start positions and end positions of pseudogenes.

3. Results and Discussion

3.1. Experimental Nucleosome Occupancy around Pseudogenes. As shown in Figure 1, nucleosome occupancy exhibits clear distribution pattern around the start positions and end positions of pseudogenes: (1) nucleosomes are depleted upstream and enriched downstream of the start positions; (2) nucleosomes are enriched upstream and depleted downstream of the end positions; (3) the nucleosome depletion pattern is similar between transcribed pseudogenes and nontranscribed pseudogenes; (4) nucleosome occupancy profile shows similar pattern between resting and activated human CD4+ T cells.

An obvious nucleosome depletion detected upstream of the start positions of transcribed pseudogenes, suggesting that the nucleosome depletion at the region may promote the pseudogene transcription by exposing the underlying sequence in a linker region, which is accessible for transcription factor binding. A similar depletion at the region downstream of the end positions of transcribed pseudogenes might imply the role of nucleosome positioning in transcription termination by facilitating the sequence to form hairpin structure to terminate transcription. Note that the nucleosome depleted regions detected upstream of the start positions and downstream of the end positions of transcribed pseudogenes match well with the transcription start region and transcription end region of the pseudogenes, respectively.

As compared with transcribed pseudogenes, nucleosome depletion both upstream and downstream of the nontranscribed pseudogenes is unexpected since nucleosome depletion in those regions is thought to be unnecessary in light of the nontranscriptional property of those pseudogenes.

3.2. Sequence-Dependent Prediction for Nucleosome Occupancy around Pseudogenes. The overall distribution trend of experimentally determined nucleosome occupancy around both start positions and end positions of pseudogenes is reproduced successfully by our computational model (Figure 2). It has been demonstrated in the previous study that predicted occupancy has a better correlation with in vitro nucleosome occupancy than in vivo occupancy [26], as our prediction depends solely on the physical properties of DNA and reflects the sequence-dependent nucleosome-forming ability. In the present paper, the depletion of nucleosomes

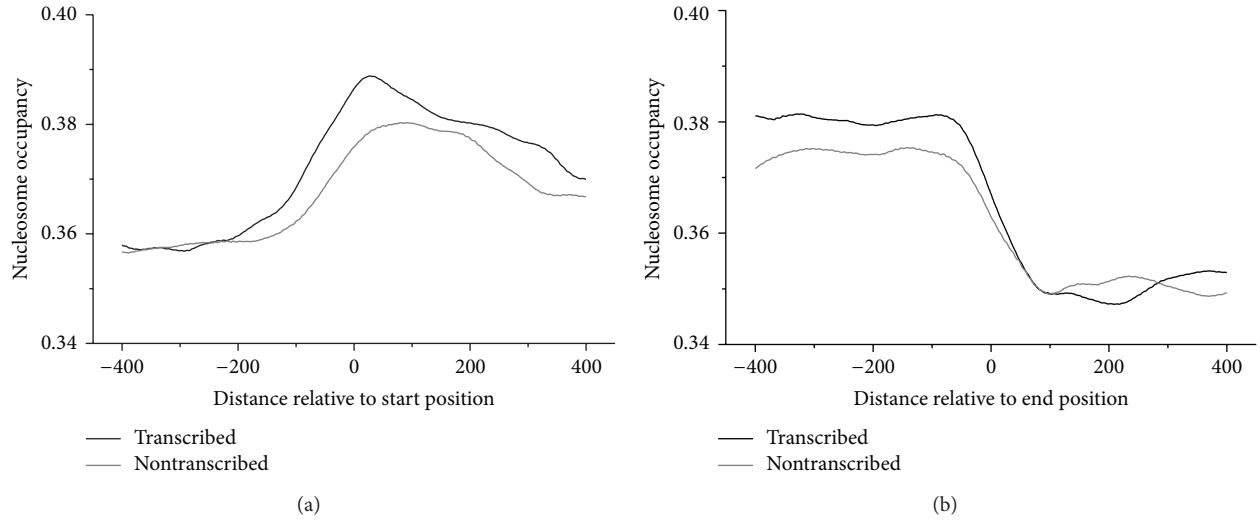


Figure 2: Calculated nucleosome occupancy around start positions and end positions of pseudogenes. Analysis of variance (ANOVA) shows significant differences of average nucleosome occupancy between transcribed and nontranscribed pseudogenes ($P < 0.001$).

both upstream of start positions and downstream of end positions and enrichment of nucleosomes both downstream of start positions and upstream of the end positions merely reflect the sequence properties to form nucleosome. The consistence of the overall distribution trend of nucleosome occupancy around pseudogenes between sequence-dependent prediction (Figure 2) and in vivo case (Figure 1) suggests that the DNA sequence is an important determinant of nucleosome positioning in human as in yeast. Our sequence-based model predicted nucleosome depletions both upstream and downstream of the nontranscribed pseudogenes. This suggests that the in vivo nucleosome depletions surrounding the nontranscribed pseudogenes are dominated by DNA sequence.

3.3. The Effect of Sequence Degeneration of Pseudogenes on Nucleosome Formation. Pseudogenes provide a natural resource of relics for researchers to explore the chromatin response to sequence mutations that are enriched in pseudogenes. Specifically, a number of structurally similar but not identical pseudogenes can be produced from a single functional gene during evolution. In particular, each of the high-transcriptional ribosomal protein genes tends to have many, in some cases over 100, pseudogenes. A simple way to test the possible change of nucleosome distribution over pseudogenes is to correlate the nucleosome occupancy over the pseudogenes with their evolutionary distances. To do this, we first downloaded the annotation (hg16-based) for 2536 ribosomal protein (RP) pseudogenes [27] from the pseudogene database (http://pseudogene.org/) and then remapped them onto the hg18 human genome using Lift program accessed at http://www.genome.ucsc.edu/. 2401 RP pseudogenes were successfully mapped. From them, duplicated pseudogenes and pseudogenic fragments that account only a small percentage of pseudogenes were removed. Finally, we retained 1931 processed pseudogenes whose sequences and annotations (GC content, DNA identity to their ancestral genes, etc.) are available at http://pseudogene.org/. We computed the average nucleosome occupancy over each pseudogene from the hg18-based experimental nucleosome reads data (the same to the procedure described in Section 2.1.2). Sequence-dependent predictive model was also applied to the pseudogenes to get average nucleosome occupancy over each one. The correlations among the variables for each RP pseudogene family were computed (Table 3).

Our data clearly illustrate that predicted nucleosome occupancy over pseudogenes tends to positively correlate with their DNA identity, suggesting that the ability of the pseudogenes to form nucleosome(s) tends to decline in the process of their evolution. However, we did not detect a positive correlation between experimental nucleosome occupancy and DNA identity. There are three possible reasons for this. Firstly, the effects of some nonsequence factors which are likely to play a larger role in nucleosome positioning in human than in simple eukaryotes, such as yeast, exceed the sequence-induced effect on nucleosome positioning [19]. Secondly, it is also possible that the mutations occurring in some pseudogenes are so little and trivial that they cannot bring about a significant effect on the nucleosome-forming ability of pseudogenes. Thirdly, the high substitution rates in nucleosome-enriched regions [28] are likely to result in the weak negative correlation between nucleosome occupancy and pseudogene identity.

We also found a significant correlation between pseudogenes' divergence and their predicted nucleosome occupancy, indicating again the decreasing trend of nucleosome-forming ability of pseudogenes during their degradation process. Furthermore, there is a strong positive correlation of predicted nucleosome occupancy of pseudogenes with their GC content, consistent with the previous finding that GC content dominates intrinsic nucleosome occupancy [29]. The GC-dependency of nucleosome occupancy and the decrease

Table 3: The proportion of significant Spearman correlations between nucleosome occupancy and pseudogene characteristics with regard to 79 RP pseudogene families.

	pgene GC	Identity[a]	Divergence[a]
Predicted	68/77[b] ($R = 0.817$, 68+)[c]	41/77 ($R = 0.589$, 39+)	41/77 ($R = -0.622$, 2+)
Experimental	3/77 ($R = -0.02$, 1+)	3/77 ($R = -0.106$, 1+)	5/77 ($R = 0.026$, 2+)

[a]The "Identity" and "Divergence" of pseudogenes from the coding sequences of their functional RP genes were taken from Zhang et al. 2002 [27]. The "Divergence" was computed with the program MEGA2, using the Kimura two-parameter model and pairwise deletion.

[b]Among 79 RP pseudogene families, there are two RP pseudogene families whose lengths are not up to 129 bp, a minimum required size for nucleosome occupancy prediction.

[c]The average of the significant Spearman correlation coefficients and the number of positive significant correlations were indicated in the parenthesis.

of GC content of pseudogenes with time [6] could explain the reduced intrinsic preference of pseudogenes for nucleosome-forming during evolution.

4. Conclusion

In this report, we analyzed the organization of nucleosomes around pseudogenes and compared between transcribed and nontranscribed pseudogenes. Experimental data-based analysis shows nucleosome depletion both upstream of the start positions and downstream of the end positions of transcribed pseudogenes, suggesting that nucleosome positioning plays an important role in both transcription initiation and transcription termination of pseudogenes. A similar depletion of nucleosomes is detected for nontranscribed pseudogenes, which is likely to be caused by sequence-dependent nucleosome-inhibitory effect. We also applied a sequence-dependent model for calculating nucleosome occupancy to pseudogenes and obtained consistent pattern with experimental nucleosome organization.

Conflict of Interests

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

Acknowledgments

This work was supported by Grants from the National Natural Science Foundation (61102162, 61271448, and 61361014) and the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT-14-B10).

References

[1] J. Maestre, T. Tchénio, O. Dhellin, and T. Heidmann, "mRNA retroposition in human cells: processed pseudogene formation," *The EMBO Journal*, vol. 14, no. 24, pp. 6333–6338, 1995.

[2] E. S. Balakirev and F. J. Ayala, "Pseudogenes: are they "junk" or functional DNA?" *Annual Review of Genetics*, vol. 37, pp. 123–151, 2003.

[3] C. Esnault, J. Maestre, and T. Heidmann, "Human LINE retrotransposons generate processed pseudogenes," *Nature Genetics*, vol. 24, no. 4, pp. 363–367, 2000.

[4] H. H. Kazazian Jr., "Mobile elements: drivers of genome evolution," *Science*, vol. 303, no. 5664, pp. 1626–1632, 2004.

[5] P. M. Harrison, N. Echols, and M. B. Gerstein, "Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome," *Nucleic Acids Research*, vol. 29, no. 3, pp. 818–830, 2001.

[6] Z. Zhang, P. M. Harrison, Y. Liu, and M. Gerstein, "Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome," *Genome Research*, vol. 13, no. 12, pp. 2541–2558, 2003.

[7] S. A. Korneev, J.-H. Park, and M. O'Shea, "Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene," *The Journal of Neuroscience*, vol. 19, no. 18, pp. 7711–7720, 1999.

[8] S. A. Korneev, V. Straub, I. Kemenes et al., "Timed and targeted differential regulation of nitric oxide synthase (NOS) and anti-NOS genes by reward conditioning leading to long-term memory formation," *The Journal of Neuroscience*, vol. 25, no. 5, pp. 1188–1192, 2005.

[9] S. Hirotsune, N. Yoshida, A. Chen et al., "An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene," *Nature*, vol. 423, no. 6935, pp. 91–96, 2003.

[10] L. Poliseno, L. Salmena, J. Zhang, B. Carver, W. J. Haveman, and P. P. Pandolfi, "A coding-independent function of gene and pseudogene mRNAs regulates tumour biology," *Nature*, vol. 465, no. 7301, pp. 1033–1038, 2010.

[11] Y. J. Han, S. F. Ma, G. Yourek, Y.-D. Park, and J. G. N. Garcia, "A transcribed pseudogene of *MYLK* promotes cell proliferation," *The FASEB Journal*, vol. 25, no. 7, pp. 2305–2312, 2011.

[12] O. H. Tam, A. A. Aravin, P. Stein et al., "Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes," *Nature*, vol. 453, pp. 534–538, 2008.

[13] T. Watanabe, Y. Totoki, A. Toyoda et al., "Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes," *Nature*, vol. 453, no. 7194, pp. 539–543, 2008.

[14] X. Guo, Z. Zhang, M. B. Gerstein, and D. Zheng, "Small RNAs originated from pseudogenes: cis- or trans-acting?" *PLoS Computational Biology*, vol. 5, no. 7, Article ID e1000449, 2009.

[15] Y.-Z. Wen, L.-L. Zheng, J.-Y. Liao et al., "Pseudogene-derived small interference RNAs regulate gene expression in African *Trypanosoma brucei*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 20, pp. 8345–8350, 2011.

[16] T. Hayakawa, T. Angata, A. L. Lewis, T. S. Mikkelsen, N. M. Varki, and A. Varki, "A human-specific gene in microglia," *Science*, vol. 309, no. 5741, p. 1693, 2005.

[17] T. Benatar and M. J. H. Ratcliffe, "Polymorphism of the functional immunoglobulin variable region genes in the chicken by exchange of sequence with donor pseudogenes," *European Journal of Immunology*, vol. 23, no. 10, pp. 2448–2453, 1993.

[18] T. J. Richmond and C. A. Davey, "The structure of DNA in the nucleosome core," *Nature*, vol. 423, no. 6936, pp. 145–150, 2003.

[19] D. E. Schones, K. Cui, S. Cuddapah et al., "Dynamic regulation of nucleosome positioning in the human genome," *Cell*, vol. 132, no. 5, pp. 887–898, 2008.

[20] T. N. Mavrich, I. P. Ioshikhes, B. J. Venters et al., "A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome," *Genome Research*, vol. 18, no. 7, pp. 1073–1083, 2008.

[21] Y. Zhang, Z. Moqtaderi, B. P. Rattner et al., "Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo," *Nature Structural and Molecular Biology*, vol. 16, no. 8, pp. 847–852, 2009.

[22] D. Zheng, A. Frankish, R. Baertsch et al., "Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution," *Genome Research*, vol. 17, no. 6, pp. 839–851, 2007.

[23] P. M. Harrison, D. Zheng, Z. Zhang, N. Carriero, and M. Gerstein, "Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability," *Nucleic Acids Research*, vol. 33, no. 8, pp. 2374–2383, 2005.

[24] R. E. Dickerson, "Definitions and nomenclature of nucleic acid structure parameters," *Journal of Biomolecular Structure & Dynamics*, vol. 6, no. 4, pp. 627–634, 1989.

[25] A. V. Morozov, K. Fortney, D. A. Gaykalova, V. M. Studitsky, J. Widom, and E. D. Siggia, "Using DNA mechanics to predict in vitro nucleosome positions and formation energies," *Nucleic Acids Research*, vol. 37, no. 14, pp. 4707–4722, 2009.

[26] J.-Y. Wang and G. Liu, "Calculation of nucleosomal DNA deformation energy: its implication for nucleosome positioning," *Chromosome Research*, vol. 20, no. 7, pp. 889–902, 2012.

[27] Z. Zhang, P. Harrison, and M. Gerstein, "Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome," *Genome Research*, vol. 12, no. 10, pp. 1466–1482, 2002.

[28] S. Washietl, R. Machné, and N. Goldman, "Evolutionary footprints of nucleosome positions in yeast," *Trends in Genetics*, vol. 24, no. 12, pp. 583–587, 2008.

[29] D. Tillo and T. R. Hughes, "G+C content dominates intrinsic nucleosome occupancy," *BMC Bioinformatics*, vol. 10, article 442, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 817134, 12 pages
http://dx.doi.org/10.1155/2015/817134

Research Article

Predicting Homogeneous Pilus Structure from Monomeric Data and Sparse Constraints

Ke Xiao, Chuanjun Shu, Qin Yan, and Xiao Sun

State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China

Correspondence should be addressed to Xiao Sun; xsun@seu.edu.cn

Received 24 October 2014; Accepted 5 January 2015

Academic Editor: Themis Lazaridis

Type IV pili (T4P) and T2SS (Type II Secretion System) pseudopili are filaments extending beyond microbial surfaces, comprising homologous subunits called "pilins." In this paper, we presented a new approach to predict pseudo atomic models of pili combining ambiguous symmetric constraints with sparse distance information obtained from experiments and based neither on electronic microscope (EM) maps nor on accurate *a priori* symmetric details. The approach was validated by the reconstruction of the gonococcal (GC) pilus from *Neisseria gonorrhoeae*, the type IVb toxin-coregulated pilus (TCP) from *Vibrio cholerae*, and pseudopilus of the pullulanase T2SS (the PulG pilus) from *Klebsiella oxytoca*. In addition, analyses of computational errors showed that subunits should be treated cautiously, as they are slightly flexible and not strictly rigid bodies. A global sampling in a wider range was also implemented and implied that a pilus might have more than one but fewer than many possible intact conformations.

1. Introduction

Type IV pili (T4P) and T2SS (Type II Secretion System) pseudopili are thin flexible filaments extending beyond microbial surfaces [1, 2] and are descended from a common ancestor [3]. Pili from different species might have similar quaternary structures, for they are assembled by highly conserved biogenesis machinery, which comprises more than a dozen proteins. These pili are composed of small, initially inner membrane-localized proteins called "pilins," the conformations of which consist of a highly conserved N-terminal α-helix and a relatively less conserved C-terminal globular domain [4]. For their importance in mobility or protein secretion, T4P and T2SS play significant roles in microbial pathogenicity and are of considerable interest as potential targets of drugs or vaccine. Moreover, some special pili contribute to the process of extracellular electron transfer (EET), known as "microbial nanowires," which inspire research efforts to understand the physicochemical basis for their conductivity [5]. All these researches will benefit if molecular structures of these pili, which might imply the mechanisms of their assembly and functions, are provided. However, difficulties caused by insolubility of subunits, heterogeneous assembly, flexibility, and presence of other surface appendages, such as cytochromes, obstruct researches of T4P and T2SS at atomic resolution.

A traditional way to study pilus structures at atomic resolution would be a combination of high-resolution structures of subunits, examined by X-ray crystallography or Nuclear Magnetic Resonance (NMR) experiments, and low-resolution envelops of pili filaments provided by cryoelectronic microscope (cryo-EM) data, which are obtained from specimens at cryogenic temperatures. To date more than a dozen pilin subunits, or at least pilin fragments, have been determined and archived in the Protein Data Bank (PDB) [6–10], but only one assembled structure of type IV pilus, the *Neisseria gonorrhoeae* (gonococcal or GC) T4P, has been obtained [11]. Meanwhile, several other pilus pseudo atomic models have been proposed, such as the type IVb toxin-coregulated pilus (TCP) from *Vibrio cholerae* [12] and pseudopili of the pullulanase T2SS (the PulG pili) from *Klebsiella oxytoca* that consist essentially of the major pseudopilin subunit PulG [13], which might also imply the difficulty of acquisition of intact pilus filaments structures.

Craig and colleagues [11] obtained the GC pilus structure by combining X-ray crystallography and cryo-EM data. A 2.3 Å resolution pilin crystal structure has been docked into 12.5 Å resolution cryo-EM maps quantitatively, by utilizing iterative helical real space reconstruction [14]. A TCP pseudo atomic structure from *Vibrio cholera* was also proposed by using the same method. The model was modified for several times, according to newly generated DXMS and cryo-EM and negative stain reconstruction [4, 12, 15]. Besides, Campos et al. described a strategy based on helical symmetry from lower resolution EM studies, on conformation restraints validated experimentally and on molecular modeling, and apply it to the PulG pseudopilus [13, 16]. The three pilus models above are all based on EM data of pili filaments and show common helical symmetry.

In this paper, we proposed an alternative approach which is based neither on higher EM maps nor on accurate *a priori* symmetric information. The new strategy enforces symmetry on the conformations among equivalent subunits in the pili assembly and then "guesses" the symmetric details, combining information of single pilin structures and sparse constraints. It is based on two assumptions: (1) T4P and T2SS pili are helically symmetric; (2) there are few differences in structure between pilins packed in crystals and in pili. As all known pilus structures show a common symmetry: a right-handed helix with ~4 subunits per turn, and the pilin subunits have similar non-globular structure which consists of a globular head and a long N-terminal hydrophobic helix, both assumptions seem strong.

This approach, combining distance constraints obtained from a variety of experiments with helical symmetric information, penalizes conformations which include constrained atom pairs that are out of range, reduces the sampling space, and then biases the process of docking efficiently. It involves two steps: a low-resolution step and a high-resolution one. The former narrows down the range of possible symmetric details, while the latter builds and refines full atomic models. A GC pilus structure was reconstructed by using this method, so were the TCP and PulG pilus. Results of the reconstruction verified that the proposed method could recover the structural details of pilus models. This study is a special case of integrating external constraint data with specific prediction methods and could be an efficient way to predict T4P or T2SS pilus structures, by combining with proper restraints.

2. Materials and Methods

2.1. The Overall Workflow. The overall workflow (shown in Figure 1) involves two separated steps: a low-resolution step and a high-resolution one. The first phase aims to find out potential pilus conformations from a wide range of sampling space. It enforces helical symmetry on pilus conformations and generates low-resolution models, or decoys as we call it, by using structures of single subunits, during which the side chains are represented by pseudo atoms. Output models of the global sampling at low resolution could be further analyzed and filtered by their energy scores and clustering results, for the following local refinement in the second step.

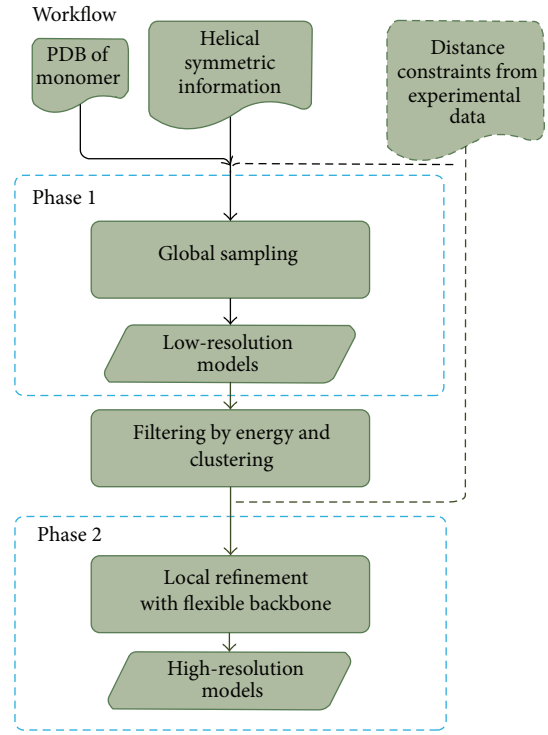


Figure 1: Computational workflow for pilus structure modeling. Distance constraints indicated by dashed lines are optional.

The second phase carries out local docking perturbations and full atomic refinements around the initial structure and generates high-resolution models. Distance constraints from experiments could be applied to both phases. Moreover, each step could be utilized independently for specific purposes. All the modeling processes are implemented by using the Symmetric Docking protocol in Rosetta software suite [17].

2.2. Preparing for Pilin Subunits. Since the GC pilus structure is known as the only structure of intact pilus, the GC pilin subunit was extracted from its pilus structure [11] (PDB ID: 2HIL), with a complete N-terminal α-helix and a C-terminal globular domain.

The crystal structure of the TCP pilin TcpA (PDB ID: 1OQV) lacks the 28 amino acid residues in the N-terminal. Because of the fact that TcpA and the *Pseudomonas aeruginosa* PAK pilin (PDB ID: 1OQW) are 75% similar in α1N (32% identity) [12], A full length TcpA structure was modeled by employing the coordinates of α1N in PAK pilin [4].

The PulG pilin structure derived from crystallography [10] (PDB ID: 1T92) lacks both the C-terminal and N-terminal segments. As proposed by Campos et al. [13] the C-terminal residues were modeled by utilizing the β2-β3 loop of closely homologous GspG (PDB ID: 3G20) from enterohemorrhagic *Escherichia coli*, and the N-terminal helix, considering its high conservation among T4P and T2SS major pilins, was also built by using the coordinates of α1N from PAK pilin.

For the homology modeling, we used MODELLER to reconstruct the C-terminal of the PulG pilin, SWISS-MODEL, and Pymol to build and superimpose the N-terminal α-helix. All the pilin models were relaxed by Rosetta Relax protocol to eliminate steric clashes, before the calculations.

2.3. Use of Conformational Constraints

2.3.1. Helical Symmetry. Up-to-date data indicate that all known T4P and T2SS pili have similar intact structures with a right-handed helical symmetry along their assembly directions. The GC pilus shows symmetry with a 10.5 Å rise along the symmetric axis and a 100.8° rotation around the axis. Meanwhile, the rise and rotation angle of the TCP are 8.4 Å and 96.7° and 10.4 Å and 84.7° for the PulG pseudopilus.

Taking into account the phenomenon mentioned above, helical symmetry was enforced during all the calculations, which was implemented by defining a symmetrical conformational space through six degrees of freedom (DOF) of rigid-body [17, 18]: the translation along the axis; the rotation around the axis; the distance between the axis and the center of mass (COM) of subunits; and three dimensions of orientation of subunits, that is, x, y, and z. Only one master subunit was taken into real calculation, and all other pilins were just translated from the master through these DOFs.

Considering the fact that the GC pilus structure is the only known full-atom conformation of pilus, an initial helical symmetric definition was extracted from it (shown in Supplementary Material available online at http://dx.doi.org/10.1155/2015/817134) and would be applied to all the following calculations. In addition, initial ranges of the six DOFs could be set to ensure the sampling is taken under some specific situations, for example, specific starting positions and searching ranges.

2.3.2. Distance Constraints. Information from a variety of experimental data could be introduced as distance constraints for our approach and applied to both low- and high-resolution steps. For the three pili discussed in this paper, the distance information was obtained as constrained pairs from either the full atomic structure already known or related experiments, such as cysteine crosslinking, salt bridge charge reversal experiments, and hydrogen/deuterium exchange mass spectrometry (DXMS).

First of all, all the F1/E5 (phenylalanine in position 1 and glutamic acid in position 5) pairs were used as constraints (M1/E5 for TCP), for the simple reason that the proximity of N-terminal nitrogen and Glu5 might be a conserved feature for most T4P and T2SS pili [19] and probably contribute to the stabilization.

Secondly, pairs of residues in adjacent N-terminal α-helices were used as constraints to keep the subunits oriented along the axis: for the GC pilus, the pair V9/L16 was determined from the cryo-EM derived model; for the PulG pilus, the pair I10/L16 was from cysteine crosslinking experiments; and for the TCP, a pair of V9/V16 was assumed, as such a distance constraint would help to keep the N-terminal α-helices of subunits packed closely in the core of pili.

Besides, other constraints which could be derived from various experiments were also added. For the GC pilus, residue pairs were picked out if they are in close proximity to each other in the intact structure, for instance, several charged residue pairs with the distance between the Cα atoms less than 10 Å (R30/E49, K76/D153, and K74/E113). Also special atoms pairs, N99/R112, were used as constraints because they are so close in the GC structure and probably form a hydrogen bond to stabilize the whole conformation. For the PulG pilus, restraints derived from cysteine crosslinking and salt bridge charge reversal experiments, or inferred from the proposed structure [13] (R26/L76, R26/E83, K68/I179, V9/V16, and M1/E5), were taken into account. For the TCP, constraints derived from charge reversal experiments and DXMS [15] (R26/L76, R26/E83, and K68/I179) were added.

In addition to the experimental data mentioned above, from which distance constraints came in this paper, other "lightweight," high-throughput experimental approaches could also potentially yield such constraints [20].

To apply this distance information to the procedures of modeling, distance constraints between pairs of atoms were characterized as energy penalty functions for Rosetta and then energy score penalties would be attached when sampling outside the constraints. All the close distance constraints for the method are set by a flat harmonic function,

$$f(x) = \begin{cases} 0, & \left|\text{distance} - x_0\right| \le \text{tolerance}, \\ \left(\dfrac{\text{distance} - x_0 - \text{tolerance}}{\text{sd}}\right)^2, & \text{distance} > x_0 + \text{tolerance}, \\ \left(\dfrac{\text{distance} - x_0 + \text{tolerance}}{\text{sd}}\right)^2, & \text{distance} < x_0 - \text{tolerance}, \end{cases} \quad (1)$$

where x_0 represents the center of constraints, which is an estimation of distance between a pair of atoms, tolerance gives the acceptable bound of constraints, and sd stands for standard deviations.

The flat harmonic function guaranteed that models were penalized if the Euclidean distance between two atoms is either too small or too large. For global sampling at low resolution, the three parameters were heuristically chosen as $x_0 = 10$, tolerance $= 5$, and sd $= 0.5$ and the constraints were enforced on Cα atoms of each residue. For full-atom modeling, the parameters were set as 4, 2, and 0.5, respectively, and the restraints were added on N-O atom pairs in salt bridges [21].

Besides, constraints that define distant relationships from the TCP DXMS were set by bounded constraint function, which describes a linear relationship between the penalty and the distance if it is out of range.

2.3.3. Ambiguous Constraints. Since the arrangement of subunits in the pili remains unknown until their assemblies are

determined, it is improper to assign an interaction specifically to two subunits. Ambiguous constraints were therefore used during these calculations. Ambiguous contact between two residues described above could be depicted as an enumeration of all combinations of the residue pairs, respectively, from two different subunits, $C1, C2, \ldots, Cn$, and then the ambiguous constraint is described by $\min(C1, C2, \ldots, Cn)$, which picks the minimum from all the scores of possible pairs. Since the total number of subunits was 15 for our calculation, an ambiguous contact should be a combination of 14 possible residue pairs ($2 * 7$, only the master subunit in the middle and the upper 7 subunits are taken into account because of the symmetry). The constraints were implemented by employing Rosetta Constraint Files [22].

2.4. Global Sampling. The global sampling phase was used for searching potential pilus conformations from a wide range of sampling space at low resolution, during which the subunits were treated as rigid-body backbones with side chains in centroid mode. A helical symmetry was enforced on the process of sampling as described in last section, and distance constraints were also applied.

Subunits were aligned along the pilus axis to some extent, with the α-helix approximately parallel to the axis, in order to optimize the initial position and then accelerate the searching process.

2.5. Local Refinement. The local refinement phase, started from a specific initial position, aimed to generate full atomic models with conformational details. A new symmetric definition was generated from the starting point and applied to the calculation with distance constraints being used too. During the local refinement procedure, a small initial perturbation was added to the subunits in low-resolution models first, then side chains were added, by a Monte Carlo Minimization which optimized both the backbones and side chains, and finally, a fast simulated annealing step was employed to relax the full atomic models with flexible backbones.

Command lines for execution of the two steps are shown in Supplemental Material.

2.6. Validating and Analyzing the Models. Although a common criterion for judging the results from such calculations is that the best model is with the lowest energy, exceptions are not uncommon during structure modeling. Also, some deviations would be got because of the insufficiency of score functions, the artifacts during data processing, or even the errors from the native structures themselves. To avoid these, we used a combination of clustering and energy score to evaluate our models, as native structure might be situated within a broad basin of low-energy conformations, to keep the efficiency and robustness of structure [23]. Thus, we chose low-energy models from the largest clusters. For low-resolution models, total energy of the master subunits was employed, while for high-resolution models, we also took the interface energy into account as it is an approximation to binding energy [24, 25] and depicts the stability of protein docking.

The pilus structures were clustered based on Root Mean Square Differences (RMSD) of the Cα positions. A similar strategy of RMSD calculation to the one taken by Campos et al. [16] was used, in which models were rotated around and shifted along their symmetric axes so that the lowest RMSDs could be determined. RMSDs over three consecutive subunits were calculated. Considering that our methods include variables from six degrees of freedom due to rigid-body translations and rotations in addition to differences in the subunit structures, such RMSD could be used to evaluate the accuracy and sufficient to depict structural details of differences among models.

3. Results and Discussion

A set of sampling processes have been completed, combining pilin structures with a variety of constraint conditions and searching ranges, shown in Table 1.

For global sampling of the GC pilus at low resolution, 4 different combinations of distance information were applied (lines 1–4, column 3 in Table 1), with 0, 2, 4, and 6 constrained pairs, respectively. The percentage of models in the largest cluster is shown, with symmetric details of the lowest-energy structure from the cluster, as a demonstration. The cutoff of clustering was set to 1.75 Å or 2.5 Å, depending on the constraints and convergent speed. Similarly, low-resolution calculations of the TCP and PulG pilus, with or without constraints, were employed and are shown in Table 1 (lines 5–8), respectively. For each calculation, at least 1000 decoys were sampled and clustered into different groups; only the largest group was taken into account for further processing. To balance the accuracy and computational efficiency, we chose proper low-resolution models from GC3, TC2, and PG2 as initial models for the following high-resolution sampling, that is, GC5, TCP3, and PG3, of which the details are shown in lines 9–11, using a criterion combining energy score and clustering. During each local refinement procedure, at least 1000 models were finally generated.

Specific constraints used in each calculation are described both in Materials and Methods and in Table 1.

3.1. Global Searching "Guesses" the Symmetry. As all the known T4P and T2SS pili show right-handed helical symmetry, it is a strong assumption that all T4P and T2SS pili would have similar symmetric modes. In order to narrow the searching space and save computational time, an initial searching range with six DOFs (illustrated in Figure S1) has been set, with rotation angle per unit between 80° and 100°, rise along axis between 5 Å and 15 Å, and COM (center of mass) radius for each subunit between 15 Å and 30 Å. In addition, the orientation of subunits was also perturbed in three dimensions.

It can be inferred that proper distance constraints would narrow down the sampling space and then have a strong influence on the convergence at the largest cluster, as depicted in both Table 1 and Figure 2. Take the GC pilus as example; when with the same cutoff, the number of decoys in the largest cluster increased from 18.10% to 36.00% (Table 1, lines

Table 1: Overview of calculations.

Index[1]	Pilus type	Distance constraints[2]	1st cluster[3]	Detailed symmetric info.[4]				Cutoff[5] (Å)
				Rise (Å)	Rotation angle (°)	Radius (Å)	Units per turn	
GC1	T4Pa	None	18.10%	11.35	99.32	18.96	3.62	2.50
GC2	T4Pa	F1/E5, N99/R112	36.00%	10.61	100.80	20.28	3.57	2.50
GC3*	T4Pa	F1/E5, N99/R112, R30/E49, V9/L16	28.60%	11.58	99.36	18.63	3.62	1.75
GC4	T4Pa	F1/E5, N99/R112, R30/E49, V9/L16, K76/D153, K74/E113	50.00%	11.87	98.62	18.69	3.65	1.75
TCP1	T4Pb	None	10.00%	7.69	100.08	27.88	3.60	2.50
TCP2*	T4Pb	R26/L76 R26/E83 K68/I179§ V9/V16 M1/E5	21.80%	8.16	98.92	26.08	3.64	2.50
PG1	T2SS	None	7.20%	14.62	76.85	16.08	4.68	2.50
PG2*	T2SS	D48/R87, E29/K51, R78/D124, R78/D117, I10/L16, F1/E5	55.00%	10.42	83.00	20.25	4.34	1.75
GC5¶	T4Pa	F1/E5, N99/R112, R30/E49, V9/L16,	72.58%	10.97	100.42	19.16	3.59	1.75
TCP3¶	T4Pb	R26/L76, R26/E83, K68/I179, V9/V16, M1/E5	60.00%	7.44	98.72	25.90	3.65	2.50
PG3¶	T2SS	D48/R87, E29/K51, R78/D124, R78/D117, I10/L16, F1/E5	55.00%	10.75	86.32	20.03	4.17	1.75

[1]Indices of calculations, GC: the GC pilus, TCP: the TCP, PG: the T2SS pseudopilus (PulG pilus).
[2]Distance constraints applied to the calculations, in the form of atom pairs.
[3]Percentage of decoys in the largest cluster.
[4]Detailed symmetric information of picked decoys (with the lowest energy score) from the largest cluster, described by the rise along the axis, the rotation angle, the radius of COM (center of mass) of each subunit, and also the number of subunits per turn.
[5]Cutoffs applied in clustering processes.
*Calculation selected for high-resolution sampling in the second step.
¶Calculation in high-resolution mode.
§Constraint defining a distance no less than 10 Å.

1 and 2) and from 28.60% to 50.00% (Table 1, lines 3 and 4) after new distance constraints being added. The grey spots in Figures 2(a), 2(b), 2(c), and 2(d) also show the trend that the more the constraints there are, the more convergent the decoys will be. Similarly, data of the TCP and PulG pilus show the same tendency, as the largest cluster of the TCP doubles (Table 1, lines 5 and 6) and the one of the PulG pilus increases from 7.20% to 55.00% (Table 1, lines 7 and 8). Obviously, proper constraints would reduce the sampling space and keep the sampling models "closer" from each other.

The distribution of clusters however shows a quite different tendency against the whole samples. For the GC pilus, no matter how the restraint conditions and the overall trend of distribution change, the cluster seems "stable," near the native conformation with an average RMSD of approximately 2.5 Å versus the native structure. Even when there were no constraints, decoys tend to converge at clusters near the native conformation, which might imply that information of a single pilin monomer can decide a pilus structure independently, to some extent. This phenomenon can also be found in the TCP and PulG pilus modeling, as shown in Figures 2(e) and 2(f).

As mentioned in Materials and Methods above, we combined energy scores with the results of clustering to pick out proper structures for the following calculations. Also, since the largest cluster is always near the native conformation, the decoys with the lowest-energy score from the largest cluster of each calculation were selected. Their symmetric information has been extracted and shown in the first eight lines of Table 1. The errors are less than 1.5 Å of translation and 2° of rotation angle.

3.2. Local Refinement Reveals the Structural Details. After global sampling within narrow ranges, the lowest-energy decoys among the largest clusters were picked out as the starting conformations for local refinements in high resolution. As described in Materials and Methods, a small initial perturbation was added to the subunits in centroid mode first, then a Monte Carlo Minimization optimized both the backbones and the side chains, and finally a fast relaxation was applied and the backbones of subunits were flexible during the last step. All the three pili were reconstructed, shown in Table 1 (lines 9–11) and Figures 3 and 4. Only for the GC pilus can we compare the results of the procedure with the intact pilus structure obtained from crystallography and EM data. More than 70% of the full atomic models

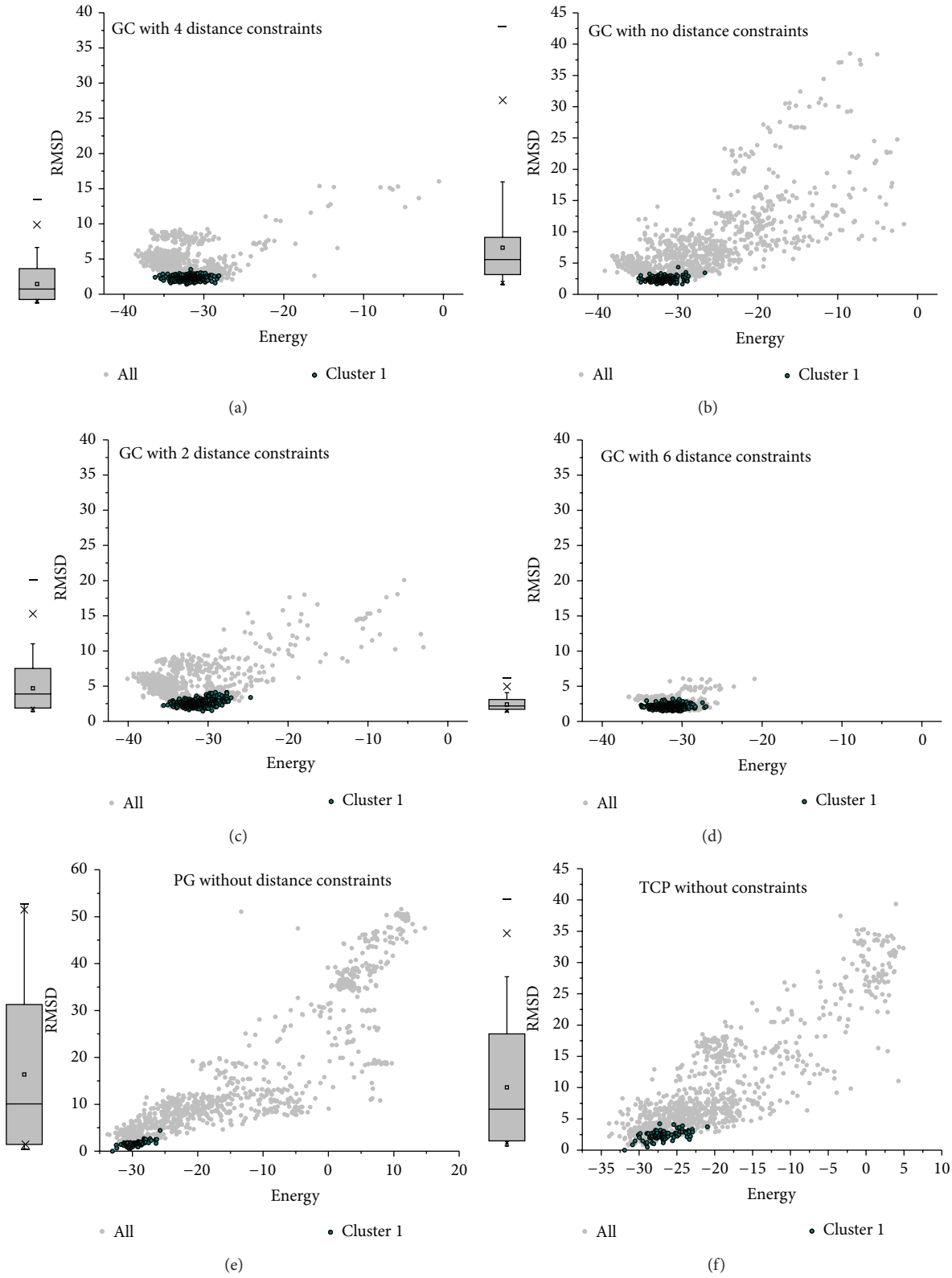


Figure 2: RMSD landscapes from the native structure (for the GC pilus) or the structure with the lowest energy in the largest cluster (for the TCP and PulG pilus) versus energy scores. The grey plots and the boxes show the distributions of RMSDs for all the models from each calculation, and the dark cyan plots show the distribution of cluster 1 (the largest cluster). (a) The GC pilus with 4 distance constraints, (b) the GC pilus with no distance constraints, (c) the GC pilus with 2 distance constraints, (d) the GC pilus with 6 distance constraints, (e) the PulG pseudopilus with no distance constraints, and (f) the TCP with no distance constraints.

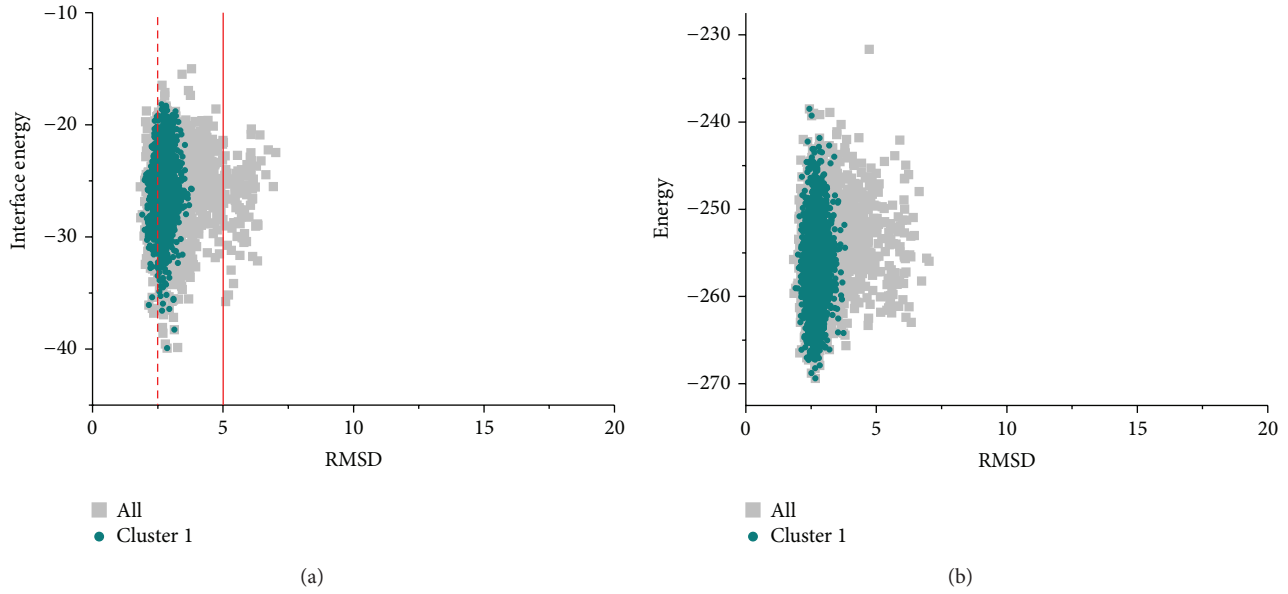


Figure 3: The interface energy and total energy landscapes of full atomic models of the GC pilus versus RMSD from the native conformation. Both show convergence near the native structure.

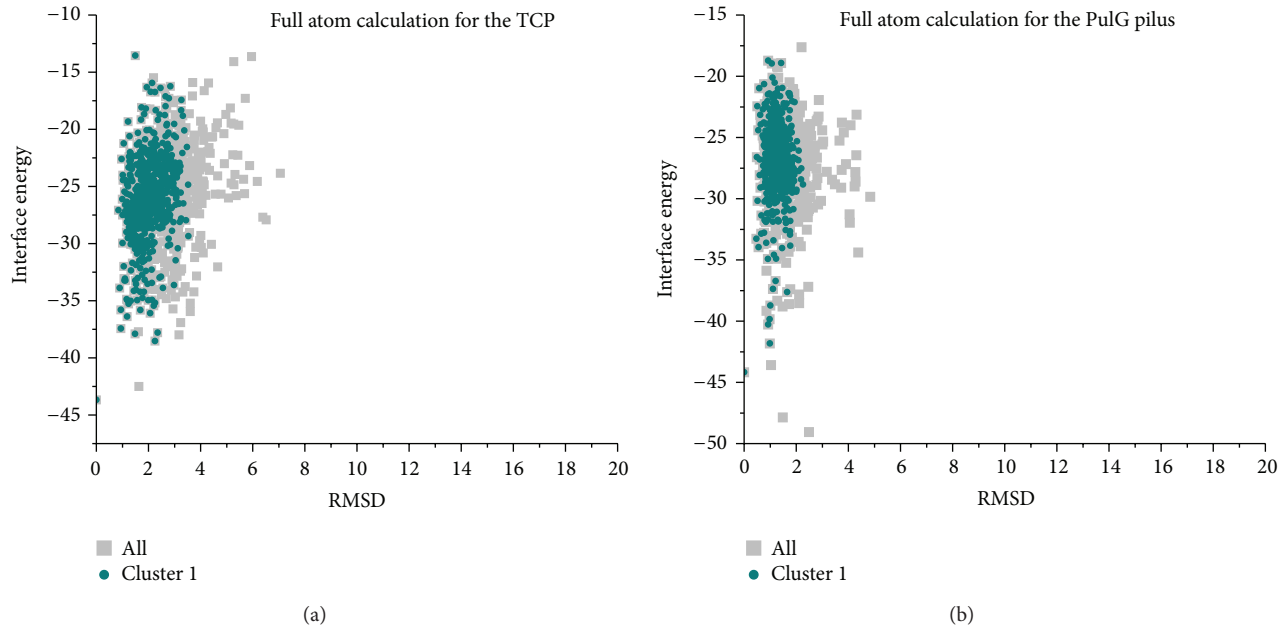
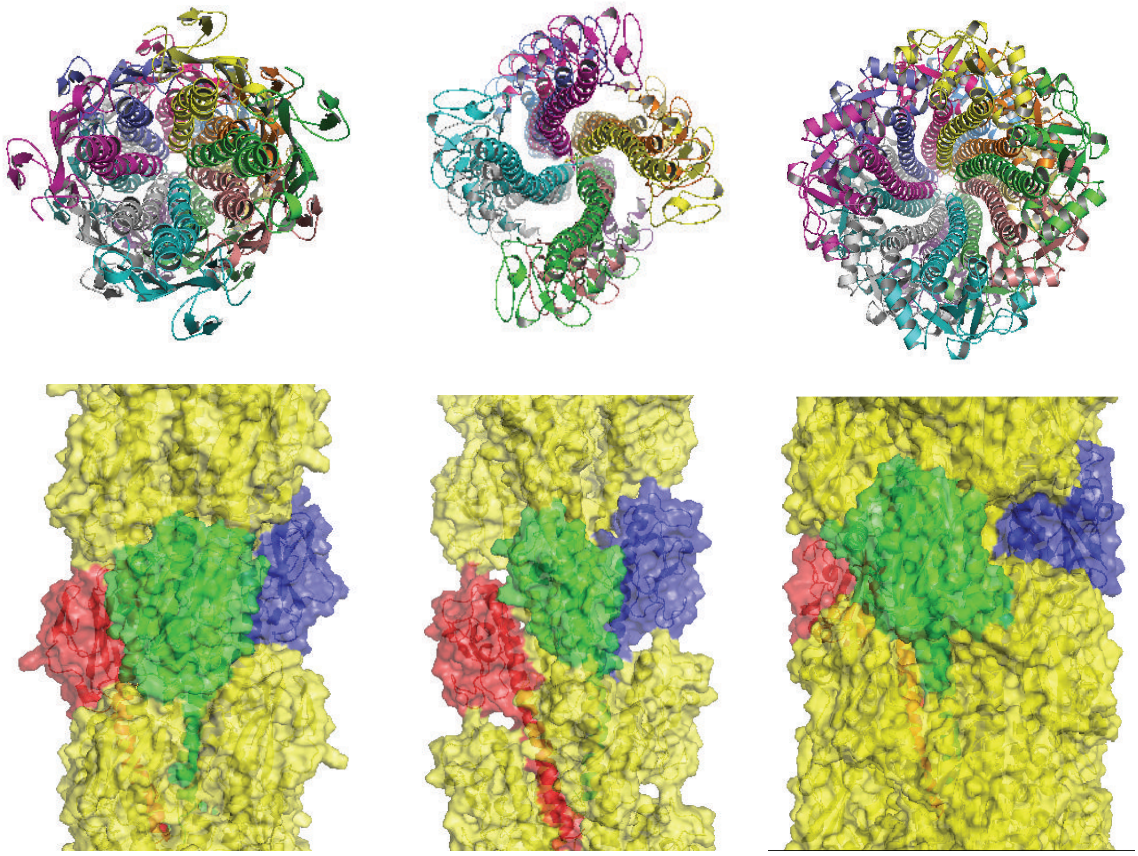


Figure 4: The interface energy landscapes of full atomic models versus RMSD from the lowest-energy conformation in the largest cluster. Left, the TCP. Right, the PulG pilus.

are clustered into the first group, as shown in Figure 3. To distinguish between correct models and incorrect ones, we clustered the results and used both total energy and interface energy score as reference. Figure 3 shows high correlation between the interface energy scores and the RMSDs from the native structure, as the RMSDs of models tend to converge at the point with the lowest score. Besides, most models in the largest cluster are near the lowest-energy model, therefore the native model. Actually, RMSD of the model with the lowest energy is about 2.8 Å from the native one, and its symmetric information is shown in Table 1 (10.97 Å for rise along axis, 100.42° for rotation angle, and 19.16 Å for COM radius). Moreover, most models are clustered between 1 and 5 Å away from the native conformation, with an average of around 2.5 Å, which is exactly similar to the estimated error in atomic position of the native GC pilus model. This accuracy is at the same level as the former method proposed by Campos et al. [16] based on molecular modeling, while the way to



(a) (b) (c)

Figure 5: Reconstruction of pili. For structures with the lowest interface energy score from the largest cluster of each pilus, three consecutive monomers are shown in red, green, and blue. Left, the GC pilus. Middle, the PulG pseudopilus. Right, the TCP.

take account of RMSD is a little different. Despite the fact that deviations of their models were calculated over all the subunits, our approach takes more degrees of freedom into account, especially the rotation around and the rise along the symmetric axis, as described above. Because of this, our method does not require the detailed symmetric information directly and thus depends less on *a priori* knowledge. In addition, such RMSDs seem sufficient to depict structural details of differences of models.

For the TCP and PulG pilus, due to the absence of published experimentally validated full atomic structures, only the landscape of interface energy versus the lowest-energy model in the largest cluster is depicted in Figure 4. Similar to the GC pilus, most models in the full-atom mode tend to cluster into a large group. The models with the lowest interface energy were taken into account. As shown in Table 1, the TCP model gets a rise of 7.44 Å along the helical axis, 98.72° of the rotation angle around the axis, and a 25.90 Å radius of COM; meanwhile, the symmetric information of the PulG pilus model is 10.75 Å, 86.32°, and 20.03 Å, respectively. Moreover, we compared our TCP model to a pseudo atomic model determined by Craig et al., and the RMSD is about 1.4 Å, which might also validate the reasonableness of our model.

All the three models show close packing of the pilin subunits, with N-terminal α-helices inside the core of pili (Figure 5), which are coincident with former models. Analyses on these structures reveal that such accuracy could recover details which are consistent with experimental phenomena. For example, although there are some deviations between the reconstructed model and the native structure, local details of the structures, such as the aromatic residue stacking of the GC pilus [12], can also be recovered (Figure 6).

3.3. Is the Rigid-Body Assumption Strong Enough? In order to figure out where the deviations of models come from, several other calculations were employed. A local refinement of the native GC pilus structure was applied, and the results also show a deviation from the native structure (Figure 7), which is correspondent with the results in the last section. This phenomenon poses a question on whether these RMSDs are derived from artifacts during calculations or from the error of the native structure itself. To eliminate the influence caused by Rosetta energy function artifacts, we used the crystallographic rigid-body transforms to build a repeating lattice [26] out of the model and carried out all-atom refinement in both the lattice and the native symmetric pilus structure.

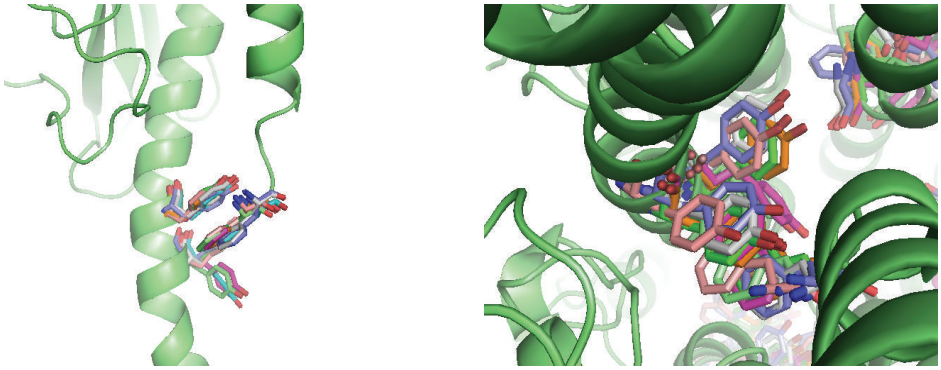


Figure 6: The N-terminal of the GC pilus has three aromatic residues whose side chains are positioned to stack, with F1 from one subunit being inserted between Y24 and Y27 from an adjacent subunit. Aromatic residues from the 10 lowest-energy models are depicted in different colors. Only one backbone of these models is shown (in ribbon).

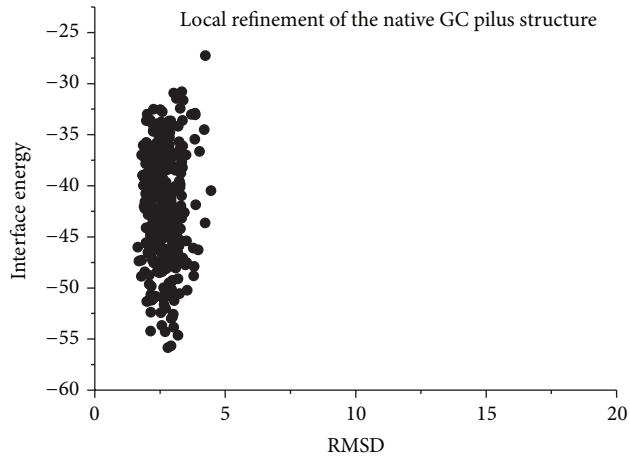


Figure 7: Energy landscapes of local refinement in the native GC pilus structure and the results also show an average deviation around 2.5 Å from the native structure.

As shown in Figure 8, the models generated from the two procedures exhibit evident differences. RMSDs of the lattice sampling from the native subunit (PDB ID: 2HI2) tend to converge at the point less than 1 Å; by contrast, the pilus sampling exhibits a convergence of RMSD at around 2 Å from the native structure.

To address the difference, we superimposed subunits models into the native structure and found out that evident conformational diversification is shown on the N-terminal α-helices while the C-terminal regions remain stable. The analysis confirms that the change of subunit conformation is based on variation of outer environment rather than artifacts derived from the software calculation and also implies that the models we got might have a more reasonable conformation, as our model also has a better MolProbity [27] score than the native one which includes more steric clashes (shown in Table S1).

As mentioned in Introduction section, the current modeling methods of pili, not only the first step of our approach, but also the one which Craig et al. used for the GC pilus [11, 14], are based on the assumption that there are few structural differences between pilins packed in crystals and in pili. Taking into account all the discussion above, this assumption might still be applicable but need to be carefully handled. Considering the rigid-body process is an efficient way to reduce computational complexity, introducing some flexibility during the modeling procedure, or at least parts of the procedure, could be a better choice.

3.4. Global Searching in Larger Range Implies More Features of Pili. To get a more complete view of pilus confirmations, global searching with larger ranges has been applied. The initial searching range was set with rotation angle between 0° and 180°, the rise along axis between 5 Å and 15 Å, and the COM (center of mass) radius of each subunit between 15 Å and 30 Å.

As mentioned in the global searching segment, the structures of the GC pilus tend to converge into a smaller structural space, even with little distance constraints. The results of the large range global searching also support this point (Figure 9). The energy scores have fallen into several troughs at specific rotation angles. Meanwhile, the distributions of diameter and rise are also related with rotation angles, shown in Figure S2, which imply that the initial rotation angle is a

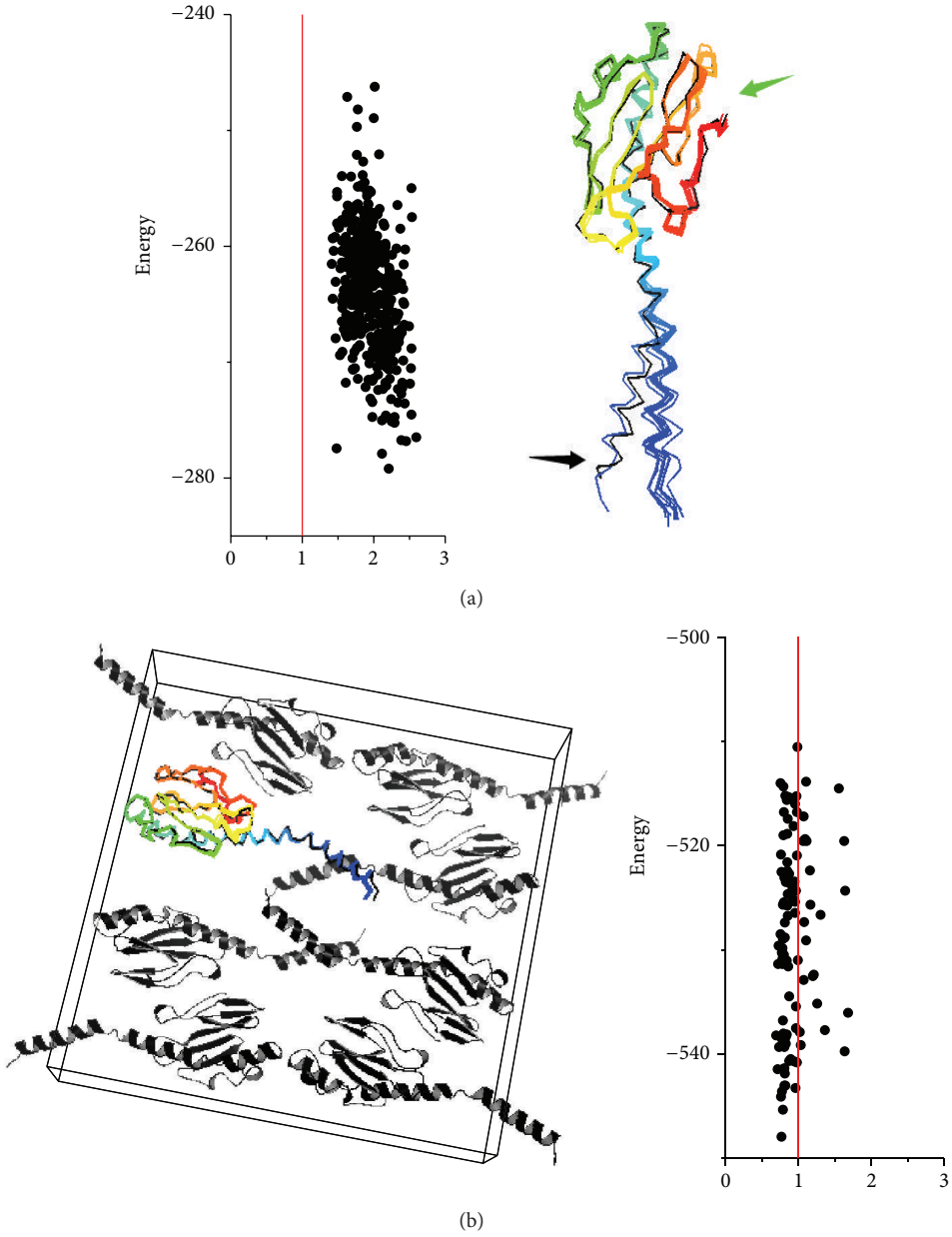


Figure 8: Energy landscapes of the GC pilin in pili versus in unit cells. (a) shows a landscape plot and lowest-energy ensemble for the GC pilin packed in a pilus environment. Evident deviations on the N-terminal helix are shown (indicated by black arrow) between the models (colored) and the native structure (black). (b) shows the same subunits simulated in the crystal environment, including its oligomeric binding partners.

key element of the modeling. A similar trend of convergence has also been observed in the sampling of both the TCP and PulG pilus, shown in Figure S3.

All these results indicate that the conformation of pili tends to converge at some specific region, and at least in the common rotation range, which has ~4 subunits per turn, the assembled conformation should be unique for each pilus.

Taking into account the assumption made by Cisneros et al. [28] that the assembly mode and details of major pilin are influenced by some other factors such as minor pilins, it can be implied that, from the perspective of docking energy, a pilus might have more than one possible assembly mode, and if it is true, these assembly modes are limited and influenced by the pilus assembly machinery.

4. Conclusions

In this paper, we describe an approach to predict full atomic models of pilus, with sparse constraint data. This method is independent of detailed symmetric data and can "guess" the symmetry from pilin structures and sparse distance constraints which could be obtained from various experiments

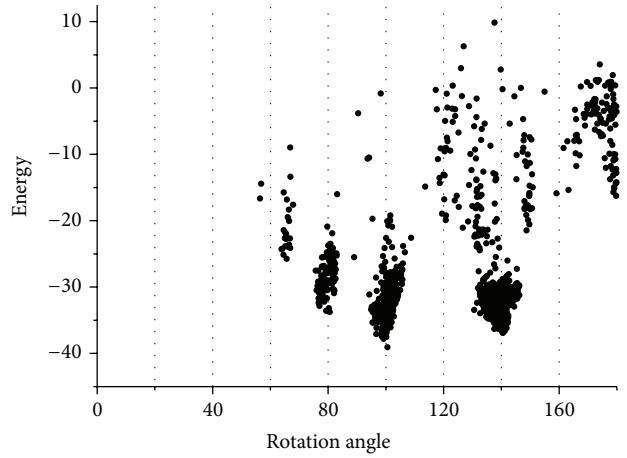


Figure 9: Landscape of energy score versus rotation angle in large range global searching for the GC pilus.

such as cysteine crosslinking, salt bridge charge reversal experiments, and DXMS. Models of the GC, TCP, and PulG pilus assembly conformations were reconstructed by this method and validated by known structural details of these three pili.

The method combines a low-resolution step with a full atomic one. During the first step, a global searching is performed within a range which contains common helical symmetric information of T4P and T2S pilus. After that, a local refinement is applied in the largest cluster of the first step. The low-resolution models from the first step tend to cluster near the native conformation, and the high-resolution phase, to some extent, can recover the structural details of the native structures.

To assess the quality of models, we use a combination of clustering and energy score to judge output models, as native structure might be situated within a broad basin of low-energy conformations, to fold efficiently and retain robustness to changes in amino acid sequence. The results of the reconstruction of several pili also prove that the criterion is reasonable.

Analyses of the errors for these results show that there are variations of subunits between their conformations packed in crystal and in intact pilus and suggest that we should take a slight flexibility into consideration during the modeling processes, instead of taking pilins as rigid bodies totally.

The global searching in a larger initial range shows that the pilus structures tend to assemble into specific basins, which implies that a pilus may have limited but probably more than one assembly mode, and be influenced by other factors in the pilus assembly machinery, such as minor pilins.

It also can be inferred from this paper that Rosetta could predict the structure of complex macromolecules such as pilus polymers, when with proper constraint information. This method could be a supplement for experimental methods and build pilus models rapidly when without sufficient EM data.

Conflict of Interests

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

Acknowledgments

The authors thank Lisa Craig for providing the new model of the TCP and R. Moretti for technical assistance on the Rosetta forum. This work was supported by the National Natural Science Foundation of China (no. 61472078) and the Key Research Fund of State Key Laboratory of Bioelectronics from Southeast University.

References

[1] L. Craig and J. Li, "Type IV pili: paradoxes in form and function," *Current Opinion in Structural Biology*, vol. 18, no. 2, pp. 267–277, 2008.

[2] C. R. Peabody, Y. J. Chung, M.-R. Yen, D. Vidal-Ingigliardi, A. P. Pugsley, and M. H. Saier Jr., "Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella," *Microbiology*, vol. 149, part 11, pp. 3051–3072, 2003.

[3] M. Ayers, P. L. Howell, and L. L. Burrows, "Architecture of the type II secretion and type IV pilus machineries," *Future Microbiology*, vol. 5, no. 8, pp. 1203–1218, 2010.

[4] L. Craig, R. K. Taylor, M. E. Pique et al., "Type IV pilin structure and assembly: X-ray and EM analyses of Vibrio cholerae toxin-coregulated pilus and *Pseudomonas aeruginosa* PAK pilin," *Molecular Cell*, vol. 11, no. 5, pp. 1139–1150, 2003.

[5] D. R. Lovley, "Electromicrobiology," *Annual Review of Microbiology*, vol. 66, pp. 391–409, 2012.

[6] P. N. Reardon and K. T. Mueller, "Structure of the type IVa major pilin from the electrically conductive bacterial nanowires of geobacter sulfurreducens," *The Journal of Biological Chemistry*, vol. 288, no. 41, pp. 29260–29266, 2013.

[7] S. Ramboarina, P. J. Fernandes, S. Daniell et al., "Structure of the bundle-forming pilus from enteropathogenic *Escherichia coli*," *Journal of Biological Chemistry*, vol. 280, no. 48, pp. 40252–40260, 2005.

[8] M. S. Lim, D. Ng, Z. Zong et al., "Vibrio cholerae El Tor TcpA crystal structure and mechanism for pilus-mediated microcolony formation," *Molecular Microbiology*, vol. 77, no. 3, pp. 755–770, 2010.

[9] B. Hazes, P. A. Sastry, K. Hayakawa, R. J. Read, and R. T. Irvin, "Crystal structure of *Pseudomonas aeruginosa* PAK pilin suggests a main-chain-dominated mode of receptor binding," *Journal of Molecular Biology*, vol. 299, no. 4, pp. 1005–1017, 2000.

[10] R. Köhler, K. Schäfer, S. Müller et al., "Structure and assembly of the pseudopilin PulG," *Molecular Microbiology*, vol. 54, no. 3, pp. 647–664, 2004.

[11] L. Craig, N. Volkmann, A. S. Arvai et al., "Type IV pilus structure by cryo -electron microscopy and crystallography: implications for pilus assembly and functions," *Molecular Cell*, vol. 23, no. 5, pp. 651–662, 2006.

[12] J. Li, E. H. Egelman, and L. Craig, "Structure of the *Vibrio* cholerae Type IVb pilus and stability comparison with the *Neisseria gonorrhoeae* Type IVa pilus," *Journal of Molecular Biology*, vol. 418, no. 1-2, pp. 47–64, 2012.

[13] M. Campos, M. Nilges, D. A. Cisneros, and O. Francetic, "Detailed structural and assembly model of the type II secretion

pilus from sparse data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 29, pp. 13081–13086, 2010.

[14] E. H. Egelman, "A robust algorithm for the reconstruction of helical filaments using single-particle methods," *Ultramicroscopy*, vol. 85, no. 4, pp. 225–234, 2000.

[15] J. Li, M. S. Lim, S. Li et al., "Vibrio cholerae toxin-coregulated pilus structure analyzed by hydrogen/deuterium exchange mass spectrometry," *Structure*, vol. 16, no. 1, pp. 137–148, 2008.

[16] M. Campos, O. Francetic, and M. Nilges, "Modeling pilus structures from sparse data," *Journal of Structural Biology*, vol. 173, no. 3, pp. 436–444, 2011.

[17] I. Andre, P. Bradley, C. Wang, and D. Baker, "Prediction of the structure of symmetrical protein assemblies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 45, pp. 17656–17661, 2007.

[18] F. DiMaio, A. Leaver-Fay, P. Bradley, D. Baker, and I. André, "Modeling symmetric macromolecular structures in Rosetta3," *PLoS ONE*, vol. 6, no. 6, Article ID e20450, 2011.

[19] C. L. Giltner, Y. Nguyen, and L. L. Burrows, "Type IV pilin proteins: versatile molecular modules," *Microbiology and Molecular Biology Reviews*, vol. 76, no. 4, pp. 740–772, 2012.

[20] R. Das and D. Baker, "Macromolecular modeling with Rosetta," *Annual Review of Biochemistry*, vol. 77, pp. 363–382, 2008.

[21] J. E. Donald, D. W. Kulp, and W. F. DeGrado, "Salt bridges: geometrically specific, designable interactions," *Proteins: Structure, Function and Bioinformatics*, vol. 79, no. 3, pp. 898–915, 2011.

[22] Constraint File, https://www.rosettacommons.org/docs/latest/constraint-file.html.

[23] D. Shortle, K. T. Simons, and D. Baker, "Clustering of low-energy conformations near the native structures of small proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 19, pp. 11158–11162, 1998.

[24] C. Wang, P. Bradley, and D. Baker, "Protein-protein docking with backbone flexibility," *Journal of Molecular Biology*, vol. 373, no. 2, pp. 503–519, 2007.

[25] S. Chaudhury, M. Berrondo, B. D. Weitzner, P. Muthu, H. Bergman, and J. J. Gray, "Benchmarking and analysis of protein docking performance in Rosetta v3.2," *PLoS ONE*, vol. 6, no. 8, Article ID e22477, 2011.

[26] M. D. Tyka, D. A. Keedy, I. André et al., "Alternate states of proteins revealed by detailed energy landscape mapping," *Journal of Molecular Biology*, vol. 405, no. 2, pp. 607–618, 2011.

[27] V. B. Chen, W. B. Arendall III, J. J. Headd et al., "*MolProbity*: all-atom structure validation for macromolecular crystallography," *Acta Crystallographica Section D: Biological Crystallography*, vol. 66, part 1, pp. 12–21, 2010.

[28] D. A. Cisneros, P. J. Bond, A. P. Pugsley, M. Campos, and O. Francetic, "Minor pseudopilin self-assembly primes type II secretion pseudopilus elongation," *The EMBO Journal*, vol. 31, no. 4, pp. 1041–1053, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 746980, 6 pages
http://dx.doi.org/10.1155/2015/746980

Research Article

Strong Ligand-Protein Interactions Derived from Diffuse Ligand Interactions with Loose Binding Sites

Lorraine Marsh

Department of Biology, Long Island University, 1 University Plaza, Brooklyn, NY 11201, USA

Correspondence should be addressed to Lorraine Marsh; lmarsh@liu.edu

Received 22 October 2014; Revised 22 December 2014; Accepted 4 January 2015

Academic Editor: Jia-Feng Yu

Many systems in biology rely on binding of ligands to target proteins in a single high-affinity conformation with a favorable ΔG. Alternatively, interactions of ligands with protein regions that allow diffuse binding, distributed over multiple sites and conformations, can exhibit favorable ΔG because of their higher entropy. Diffuse binding may be biologically important for multidrug transporters and carrier proteins. A fine-grained computational method for numerical integration of total binding ΔG arising from diffuse regional interaction of a ligand in multiple conformations using a Markov Chain Monte Carlo (MCMC) approach is presented. This method yields a metric that quantifies the influence on overall ligand affinity of ligand binding to multiple, distinct sites within a protein binding region. This metric is essentially a measure of dispersion in equilibrium ligand binding and depends on both the number of potential sites of interaction and the distribution of their individual predicted affinities. Analysis of test cases indicates that, for some ligand/protein pairs involving transporters and carrier proteins, diffuse binding contributes greatly to total affinity, whereas in other cases the influence is modest. This approach may be useful for studying situations where "nonspecific" interactions contribute to biological function.

1. Introduction

Ligand interactions with proteins may be specific or nonspecific. Many ligands bind to proteins via tight, cooperative interactions, that is, "lock and key" mechanisms. However, other, looser interactions also occur and may have physiological significance (e.g., in multidrug resistance). These interactions can be modeled by molecular dynamics (MD) [1], but the timescales involved in modeling multiple on and off diffusion, especially in and out of solvent, may strain the current limits of the technology. Pure Monte Carlo methods can estimate binding to loose protein cavities but can be inefficient, given the rugged energy profiles of binding and the large binding space occupied by clashing between ligand and receptor.

The MCMC approach has been widely used in physics and statistics to determine the probability distribution of multidimensional processes. For example, the distribution of molecular interactions with DNA has been modeled using MCMC [2]. Modeling uncertainty and low probability states in protein structure prediction also benefits from MCMC and related approaches [3]. A common theme in many structural analyses is the tradeoff between entropic and enthalpic contributions to free energy [4].

In the MCMC process, the probability of a state being occupied (number of steps occupying the state/number of total steps) is proportional to the stationary probability distribution of the process. One major advantage of MCMC is that high probability regions of the distribution are sampled more than low probability regions (importance sampling), increasing efficiency for study of distributions that have extensive regions of low probability. Another advantage is that extensive theoretical and practical applications of MCMC methods show that they are extremely robust and flexible approaches to model probability distributions [5].

For a MCMC method to be applicable, a distribution must exhibit certain traits. In particular, the distribution must satisfy regularity criteria. The process must be ergodic, that is, be capable of returning to any given state. For ligand-receptor interaction, this requires that solvent regions be

finite. The process must be irreducible; that is, all states must be reachable by a random walk. In practice, this means that either ligands must access a solvent region that permits all conformations or the process must be allowed to jump to all allowed states. For instance, a ligand unable to rotate in a given site must be given a statistically valid path to rotate, either in solvent or through jump diffusion. In some cases, diffusion of small molecules will not require extra techniques to achieve irreducibility in large-volume sites.

Ligand binding pockets in proteins are diverse in shape, depth, and size [6]. Though many binding sites exhibit specificity required for their biological function, other molecules require less specificity to fulfill their purpose. The bacterial drug efflux pumps, including RND transporters such as AcrB, serve to export multiple toxic substrates from the cell [7]. Some of these substrates make fairly well-defined contacts with the binding cavity of the pump [8] while others may not. Computational studies have contributed to our understanding of the basis of drug pumping [1, 9]. AcrB preferentially pumps hydrophobic molecules and the pump chamber is lined with phenylalanine residues. Another example of molecules binding diverse substrates is the several families of sterol binding proteins, which also bind other lipids. The human serum albumin protein binds an extremely diverse set of ligands through two pockets and some crevice regions. Pocket 2, which is shallow and solvent accessible, binds the sedative diazepam and the anesthetic halothane amongst many substrates. Drugs may compete with each other for binding to albumin, which suggests some specificity of binding, but could be due to nonspecific occlusion of a hydrophobic patch available for contact [10].

Here, we treat a ligand molecule as a MCMC process diffusing within pose space in a receptor site, with its probability density distribution determined by the K_d of interaction with the receptor given its positional and rotational state for relatively rigid molecules. We show, using the MCMC method, that, for large sites, such as those of the AcrB bacterial drug transporter, multiple states or binding poses contribute to binding efficiency. This approach may have application to modeling other macromolecular interactions such as DNA-protein and protein-domain interactions. This method is also relevant to a number of pharmacological analyses.

2. Methods

2.1. Model Systems. Systems for ligand/protein binding analysis were selected based on the potential for loose, nonspecific interactions. Each of these proteins had large binding regions which offered ample room for the ligand to bind in multiple conformations and at multiple sites within the binding region of the protein. The AcrB multidrug transporter of *E. coli* undergoes cyclic changes that first open a large ligand-binding cavity and then expel the contents outside the cell. The cavity is large enough that all known ligands can, in principle, adopt many conformations within the cavity. The pump is very nonselective, suggesting that efflux does not require interaction with a specific evolutionarily selected binding site. Human serum albumin is an abundant protein in blood that plays many roles, including the transport of fatty acids. It also plays important roles in pharmacology by absorbing diverse drugs and reducing their free concentration in the bloodstream. Human serum albumin has at least two large pocket domains on its surface. Steroid transporters, also in the bloodstream, can bind many hydrophobic compounds including diverse steroids. The steroid transporters have pockets larger than what would be required to bind a simple steroid. These model systems were studied using the known sites of drug interaction.

The open binding chamber of AcrB (PDB ID: 3AOD, chain A) was studied with four ligands. Toluene is a solvent pumped by the exporter [11]; skatole is a toxic hydrophobic molecule ubiquitous in the natural environment of *E. coli*; acridine orange is the dye first used to characterize the exporter; and minocycline is an exported antibiotic crystalized with AcrB in the PDB ID: 3AOD structure. Two sterol binding proteins were studied as well (PDB IDs: 1ZHY and 2A1B). The human serum albumin (HSA) protein has two canonical ligand binding pockets. Pocket 2, reported to bind diazepam and halothane, was studied (PDB ID: 1E7B, chain A) [10, 12].

2.2. The MCMC Process. To study ligand binding in multiple conformations, it was necessary to estimate the proportion of ligand binding in each position. A MCMC process was designed such that the stationary probability in pose space would equal the predicted distribution of ligands in a binding site. To ensure ergodicity the process was constrained to a sphere centered on the initial ligand binding pose. Trial MCMC processes were studied to determine the effect of windowing and various parameters on performance. The spatial scoring window was set, conservatively, at 2 angstroms, since that is a commonly accepted RMSD for significantly similar poses for ligands and synchronized with the poxel definition used (see below). Depending on the number of steps and alpha, the step distance was analyzed. In general, the time to equilibration for an MCMC process is best determined by experimentation. A step size, α, for translation of 0.9 angstroms and an α for rotation of 0.6 radians were used. An important consideration for importance sampling of a rugged distribution is the tradeoff between the density of sampling and number of steps. For most systems studied, less than 200,000 steps were required to populate the majority of poxels. Stochastic jump diffusion to a probable poxel every 1000 steps was used to ensure sampling of all poses.

2.3. Scoring Affinity. A soft Lennard-Jones scoring parameter was employed as is common in successful empirical scoring functions [13–15]. The ΔG of atomic interaction was set at 1 kcal/mol at the VDW distance for the atoms involved [16, 17]. The K_d for interaction was calculated as $K_d = e^{\Delta G/RT}$ with the MCMC Metropolis transition determined by $1/K_d$. ΔG for molecules in solvent was set at 0. For some experiments AutoDock Vina scoring [18] was used as an alternative method with similar results. This MCMC method

is compatible with any scoring function for ligand/protein interaction.

2.4. Definition of Pose Space. Rigid ligands can be positioned in 6-dimensional translational/rotational conformational pose space. Pose space was divided into "poxels" by analogy with 3-dimensional voxels. Poxels were placed 3 angstroms apart in x, y, and z dimensions and 51.4° apart in rotational dimensions. No torsional dimensions were included for the rigid molecules studied here. These poxels were large enough that a single poxel could accommodate most of the motions of known tight-binding ligand/receptor complexes such as those found in PDB ID: 2rnh and 4gid. Note that molecules in adjacent poxels occlude the space of each other but are conformationally distinct and make different contacts with the receptor.

2.5. Analysis of Diffuse Binding. The MCMC method for DBF calculation was programmed in a Perl script. The MCMC process was typically run for 100,000 steps which typically populated the average poxel with more than 20 process visits. The most visited pose (modal) was recorded and entropy enhancement (EE) was calculated as total MCMC steps/modal MCMC steps. In essence, EE is the proportion of the steps a ligand spent at sites other than the modal pose. The less time spent at the modal pose, the more time spent in other poses. EE can be used to calculate the overall K_d (adjusted for multiple poses) as

$$\text{Overall } K_d = \frac{\text{Modal } K_d}{\text{EE}}. \quad (1)$$

EE can be useful for converting the K_d predicted for binding to the single best binding site into the overall binding, allowing for pose flexibility. For typical high-affinity ligand-receptor pairs, the number of poses in the site was 1 or 2 and the MCMC steps for the modal pose equaled or nearly equaled the total number of steps, producing an EE value of ~1 and an overall $K_d \simeq$ modal K_d.

The effective number of poses for each site (DBF) was also calculated. DBF factors that many poses have a probability >0, nonetheless, are enthalpically unfavorable and hence poorly populated. For equally populated poses, the total poses equal DBF. For a more typical Poisson distribution, where the mean number of visits equals the standard deviation of the number of visits, DBF is 0.5x number of poses. DBF is defined as

$$\text{DBF} = \frac{\text{NP}}{\left(\left(\text{Var}_{\text{MV}}\right)/\text{MV}^2\right) + 1}, \quad (2)$$

where NP is the number of $P > 0$ poses, MV is mean visits to a poxel, and Var_{MV} is the variance of MV. DBF is equal to the number of permissible poses if all sites bind equally. However, DBF is equal to ~1.0 if only a single pose has substantial probability, even if many poses with a probability >0 exist. If binding affinity is distributed amongst sites with a Poisson distribution, DBF $\simeq 0.5 *$ NP. DBF is dependent on both the number of possible poses and the distribution of the probability of those poses. A high DBF indicates that diffuse binding plays an important role in overall affinity of a protein for a ligand.

Table 1: Relative contributions of multiple poses to predicted affinity.

Complex	Poses	DBF	EE
Spatial fit scoring			
AcrB/minocycline	36	5 ± 2	2.5 ± 0.8
AcrB/acridine orange	68	14 ± 4	6.9 ± 1.2
AcrB/skatole	1494	57 ± 31	9.7 ± 5.7
AcrB/toluene	2429	661 ± 139	53.7 ± 10.1
OSBP/cholesterol	10	1.3 ± 0.3	1.1 ± 0.2
Cinnamomin/ergosterol	7	2.2 ± 1.1	1.7 ± 0.4
HSA/diazepam	5050	2298 ± 1703	435 ± 333
HSA/halothane	6897	4319 ± 1626	1630 ± 695
BAR/carazolol	3	1.8 ± 0.9	1.5 ± 0.4
Empirical scoring			
AcrB/minocycline	52	8.5 ± 2.0	4.1 ± 1.7
AcrB/acridine orange	81	5.7 ± 3.1	3.3 ± 2.3
AcrB/toluene	1965	726 ± 67	107 ± 10
AcrB/skatole	1637	526 ± 58	76 ± 9

Abreviations: OSBP, oxysterol binding protein; HSA, human serum albumin; BAR, beta-adrenergic receptor.
"Poses", refers to number of possible ($P > 0$) poses in binding region analyzed.
Analyses performed in triplicate with runs of 100,000 MCMC steps.

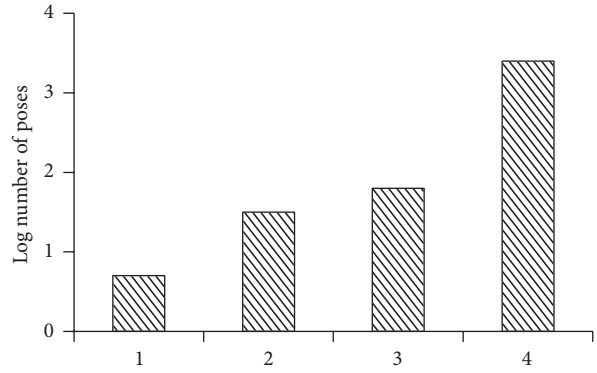


Figure 1: Number of poses in binding site for various ligands. (1) Beta-adrenergic receptor/carazolol (PDB ID: 2rhn_A), a typical tight-binding receptor/ligand complex; (2) AcrB/minocycline (28 heavy atoms); (3) AcrB/acridine orange (20 heavy atoms); and (4) AcrB/toluene (7 heavy atoms).

3. Results and Discussion

3.1. Binding to a Loose Site. The AcrB binding site is approximately 2600 angstroms^3 which is about twice the volume of the typical antibiotics that it pumps and about 12 times the volume of toluene. Analysis of binding of acridine orange (originally used to discover the AcrB gene) and other ligands showed that they could bind in several conformations (Figure 1 and Table 1).

DBF values for various ligands binding to the cavity of the AcrB multidrug pump varied greatly. Not surprisingly, the smaller ligands had higher DBF values than the larger ones (Table 1) indicating that diffuse binding contributes more to

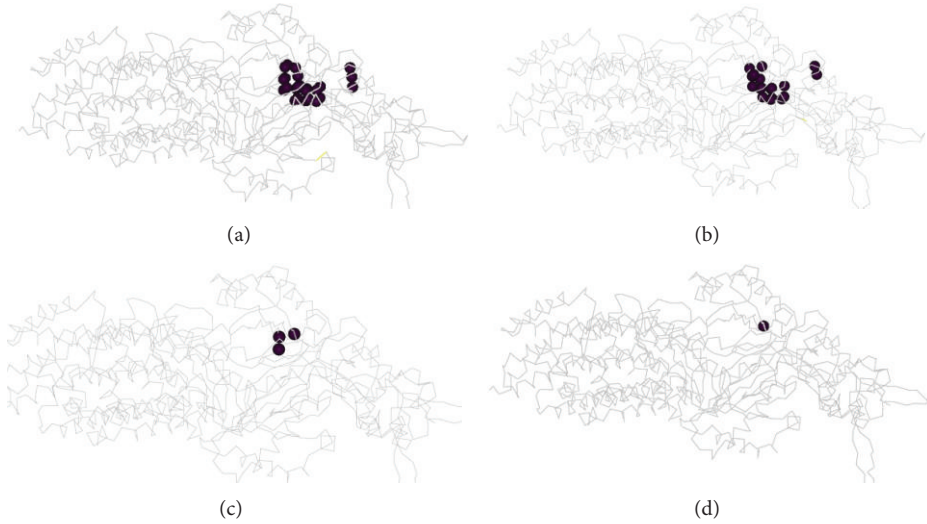


Figure 2: Toluene binding sites in AcrB. The centroids of the sites are shown. Rotationally equivalent poses may occupy a single site. (a) All pose sites are shown. (b) Sites contributing 52% of total binding are shown. (c) Poses contributing 15% of binding are shown. (d) A single pose contributing 10% of binding is shown. Though a relatively high-affinity site exists, low-affinity sites make up the bulk of binding probability.

their overall affinity. Minocycline, which has been crystalized with AcrB in a single pose, had effectively 1-2 poses in this analysis as well, though a number of unfavorable poses had a probability >0. The absolute DBF value may be slightly inflated since the poxel dimensions were conservatively defined and even a relatively tight-binding molecule often has some freedom to move within its site. DBF comparisons between ligands may be more useful for some purposes. Toluene, a substrate that has been shown to be pumped by this transporter had a very high DBF, approximately 100 times that of minocycline. As discussed below, a mixture of many low-affinity sites and one higher affinity site seems to contribute to AcrB affinity for toluene. Acridine orange, the substrate originally used to characterize AcrB, had a low DBF, but more than one pose seemed to contribute to its binding. Skatole, a molecule toxic to *E. coli* but ubiquitous in its environment, was intermediate between acridine orange and toluene. Using empirical scoring, skatole had a DBF more similar to toluene, suggesting that for some ligands the choice of scoring function might be significant. The difference was due to the presence of a single high-affinity AcrB/skatole poxel in the atomistic analysis that was absent in the empirical analysis. However, both scoring methods predicted that most of skatole affinity involved diffuse binding. The other tested ligands had similar DBFs with both scoring methods. For the small ligands, substantial enthalpy-entropy compensation was possible, suggesting that the potential to bind with many poses contributed to efficiency of pumping these molecules. Similar effects may influence patterns of multidrug resistance in other systems [19]. It should be noted that the increased affinity predicted for small molecules such as toluene and skatole is not purely an entropic effect. The molecules have more freedom to move but are also able to make enthalpically favorable (though weak) contacts with the pump in these alternative poses. These loose configurations may resemble other examples of ligands binding in sites larger than what the molecules require. For example, in virtual screening, decoy molecules may appear to be able to bind sites that do not fit well [19].

3.2. Visualization of Relative Poxel Probability. Populated poxels ($P > 0$) were visualized for toluene (Figure 2). Poxels with some probability of being populated were not clustered, indicating that several sites in the large cavity contribute to affinity. The distribution of MCMC visits to poxels was determined. The distribution of poxels contributing about 50% of total affinity was also dispersed (Figure 2). However, one pose alone contributed about 10% of the total probability and could be considered a relatively high-affinity site. If crystallization of AcrB with toluene was successful one might predict that this site would appear as the site of toluene interaction. However, the analysis here suggests that most of the experimentally determined affinity would be due to diffuse, low-affinity interactions. The poxels of minocycline/acridine orange clustered at the region of the minocycline site in the crystal structure PDB ID 3AOD. Skatole poxels were scattered, in a Poisson-like distribution similar to that of toluene.

Overall, the probability distribution of toluene in AcrB resembles a Poisson distribution of site occupancy (Figure 3) with the mean number of poxels occupied increasing as affinity decreases. Skatole, which chemically resembles toluene, had a nearly identical distribution to toluene, though its DBF was lower because of a single pose with higher affinity than the highest affinity site of toluene. Significantly, despite the existence of higher affinity sites, most of the binding of toluene and skatole is still derived from interaction with multiple low-affinity sites. In contrast, the affinity of both acridine orange and minocycline is dominated by one or a few poses, with only smaller, less significant, contributions from alternative sites (Figure 3 and Table 1). The DBF

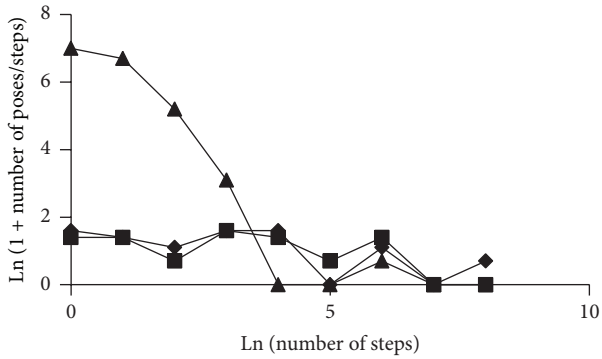


Figure 3: Specific versus nonspecific binding. The number of steps per pose is plotted versus the number of poses with a specific number of steps. A high value on the vertical axis indicates many low-affinity poses. A high value on the horizontal axis indicates a few high-affinity poses. Triangles, AcrB/toluene; squares, AcrB/acridine orange; diamonds, AcrB/minocycline. The graph for skatole was nearly identical to the line for toluene except for a single higher affinity site and is omitted for clarity. See text for details.

values for toluene were roughly 100-fold higher than those for minocycline, indicating a much greater role for diffuse binding in toluene binding than for minocycline binding. Thus small ligands binding to the AcrB site derive an affinity benefit from the ability to bind weakly in many conformations, while larger ligands benefit by binding in only a few conformations, but making more receptor/ligand contacts in each conformation. Acridine orange and minocycline also have fewer poxels that do not clash with the AcrB pump chamber than toluene or skatole.

3.3. Other Nonspecific Complexes. As a comparison with AcrB, two sterol binding proteins were analyzed. These molecules, oxysterol carrier protein and cinnamomin, both bind to a number of sterols with little specificity. They have no sequence or structural similarity. Both bound sterol with only a single pose showing high probability, though other poses did have a probability >0 (Table 1). This result suggests that the flexibility and looseness of their respective binding pockets serve mostly in accommodating ligand diversity and not increasing ligand affinity for the molecules studied.

Human serum albumin (HSA) has been extensively studied for its ability to bind naturally occurring hydrophobic molecules such as fatty acids and also drugs. Competition studies and X-ray crystallographic studies suggest that two large pockets are involved in at least some of the binding [10]. When binding of halothane and diazepam, which bind pocket 2, was studied by the MCMC method considerable pose diversity appeared to be present within the known binding pocket (Table 1). These molecules, both by visual inspection and docking studies with AutoDock Vina, did not have binding sites with the characteristics of known high-specificity sites. They appeared to bind in many poses to the hydrophobic pocket region 2 of HSA and occlude the site, competing in that way with other substrates. As with toluene binding to AcrB, a few higher affinity sites were present, but the overall predicted affinity was dominated by a large number of low-affinity interactions. In contrast, binding of carazolol to the beta-adrenergic receptor (a classic "lock and key" tight interaction) yielded a DBF of only 1.8.

4. Conclusions

The MCMC method provides an approach to conceptualize and study diffuse binding interactions that have heretofore been relegated to the area of "nonspecific" interactions. These types of interactions are likely to play important roles in some ligand binding, DNA-protein interactions and in domain interactions in proteins, especially the interactions of intrinsically disordered domains. In some cases this analysis has shown that nonspecific interactions might also contribute to overall affinity of ligands. In other cases, apparently loose fitting ligands actually bind with a single effective pose. MCMC methods complement molecular dynamic approaches and allow modeling of interactions that occur over relatively long time periods. The MCMC approach is reasonably fast and statistically rigorous and thus provides a link between physics-based methods that attempt to model actual molecular behavior and algorithmic methods that seek to efficiently provide a single best binding conformation.

Conflict of Interests

The author declares no conflict of interests for this paper.

Acknowledgments

The author would like to thank the LIU Biocomputing Facility and Dr. Samuel Watson for helpful comments and suggestions. Software implementing the MISMA MCMC method to calculate DBF values for receptor/ligand pairs is available from the author at http://myweb.brooklyn.liu.edu/lmarsh/misma.html.

References

[1] P. Ruggerone, A. V. Vargiu, F. Collu, N. Fischer, and C. Kandt, "Molecular dynamics computer simulations of multidrug RND efflux pumps," *Computational and Structural Biotechnology Journal*, vol. 5, no. 6, Article ID e201302008, 2013.

[2] Z. Q. Tan, "The limit theorems for maxima of stationary Gaussian processes with random index," *Acta Mathematica Sinica, English Series*, vol. 30, no. 6, pp. 1021–1032, 2014.

[3] E. Freire, "Statistical thermodynamic linkage between conformational and binding equilibria," *Advances in Protein Chemistry*, vol. 51, pp. 255–279, 1998.

[4] J. D. Chodera and D. L. Mobley, "Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design," *Annual Review of Biophysics*, vol. 42, no. 1, pp. 121–142, 2013.

[5] L. Tierney, *Markov Chain Monte Carlo in practice*, Chapman & Hall, New York, NY, USA, 1996.

[6] M. Gao and J. Skolnick, "A comprehensive survey of small-molecule binding pockets in proteins," *PLoS Computational Biology*, vol. 9, no. 10, Article ID e1003302, 2013.

[7] P. Ruggerone, S. Murakami, K. M. Pos, and A. V. Vargiu, "RND efflux pumps: structural information translated into function and inhibition mechanisms," *Current Topics in Medicinal Chemistry*, vol. 13, no. 24, pp. 3079–3100, 2013.

[8] D. Du, Z. Wang, N. R. James et al., "Structure of the AcrAB-TolC multidrug efflux pump," *Nature*, vol. 509, no. 7501, pp. 512–515, 2014.

[9] N. Fischer, M. Raunest, T. H. Schmidt, D. C. Koch, and C. Kandt, "Efflux pump-mediated antibiotics resistance: insights from computational structural biology," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 6, no. 1, pp. 1–12, 2014.

[10] J. Ghuman, P. A. Zunszain, I. Petitpas, A. A. Bhattacharya, M. Otagiri, and S. Curry, "Structural basis of the drug-binding specificity of human serum albumin," *Journal of Molecular Biology*, vol. 353, no. 1, pp. 38–52, 2005.

[11] H. Nikaido and J. M. Pagès, "Broad-specificity efflux pumps and their role in multidrug resistance of Gram-negative bacteria," *FEMS Microbiology Reviews*, vol. 36, no. 2, pp. 340–363, 2012.

[12] A. A. Bhattacharya, S. Curry, and N. P. Franks, "Binding of the general anesthetics propofol and halothane to human serum albumin: high resolution crystal structures," *The Journal of Biological Chemistry*, vol. 275, no. 49, pp. 38731–38738, 2000.

[13] R. A. Friesner, J. L. Banks, R. B. Murphy et al., "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy," *Journal of Medicinal Chemistry*, vol. 47, no. 7, pp. 1739–1749, 2004.

[14] T. Cheng, X. Li, Y. Li, Z. Liu, and R. Wang, "Comparative assessment of scoring functions on a diverse test set," *Journal of Chemical Information and Modeling*, vol. 49, no. 4, pp. 1079–1093, 2009.

[15] L. Marsh, "Prediction of ligand binding using an approach designed to accommodate diversity in protein-ligand interactions," *PLoS ONE*, vol. 6, no. 8, Article ID e23215, 2011.

[16] Z. Simon, M. Vigh-Smeller, Á. Peragovics et al., "Relating the shape of protein binding sites to binding affinity profiles: is there an association?" *BMC Structural Biology*, vol. 10, article 32, 2010.

[17] T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, "DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases," *Journal of Computer-Aided Molecular Design*, vol. 15, no. 5, pp. 411–428, 2001.

[18] O. Trott and A. J. Olson, "Software news and update AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *Journal of Computational Chemistry*, vol. 31, no. 2, pp. 455–461, 2010.

[19] N. M. King, M. Prabu-Jeyabalan, R. M. Bandaranayake et al., "Extreme entropy-enthalpy compensation in a drug-resistant variant of HIV-1 protease," *ACS Chemical Biology*, vol. 7, no. 9, pp. 1536–1546, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 313709, 9 pages
http://dx.doi.org/10.1155/2015/313709

Research Article

A Systematic Analysis of Candidate Genes Associated with Nicotine Addiction

Meng Liu, Xia Li, Rui Fan, Xinhua Liu, and Ju Wang

School of Biomedical Engineering, Tianjin Medical University, 22 Qixiangtai Road, Tianjin 300070, China

Correspondence should be addressed to Ju Wang; wangju@tmu.edu.cn

Received 30 September 2014; Revised 28 December 2014; Accepted 2 January 2015

Academic Editor: Yuedong Yang

Nicotine, as the major psychoactive component of tobacco, has broad physiological effects within the central nervous system, but our understanding of the molecular mechanism underlying its neuronal effects remains incomplete. In this study, we performed a systematic analysis on a set of nicotine addiction-related genes to explore their characteristics at network levels. We found that NAGenes tended to have a more moderate degree and weaker clustering coefficient and to be less central in the network compared to alcohol addiction-related genes or cancer genes. Further, clustering of these genes resulted in six clusters with themes in synaptic transmission, signal transduction, metabolic process, and apoptosis, which provided an intuitional view on the major molecular functions of the genes. Moreover, functional enrichment analysis revealed that neurodevelopment, neurotransmission activity, and metabolism related biological processes were involved in nicotine addiction. In summary, by analyzing the overall characteristics of the nicotine addiction related genes, this study provided valuable information for understanding the molecular mechanisms underlying nicotine addiction.

1. Introduction

Cigarette smoking is the most common form of tobacco use and is one of the main preventable causes of premature death and disability worldwide [1, 2]. Although there are some effective control policies and interventions on tobacco abuse, the negative impact of tobacco dependence on society is still staggering. The World Health Organization estimates that there are currently about 1.3 billion smokers worldwide, resulting in approximately 5 million annual tobacco attributable deaths [3, 4]. If the current trend continues, by 2020, smoking will become the largest single health problem worldwide, causing 10 million deaths annually, mostly in low- and middle-income countries [5]. Despite these grim statistics, cigarette smoking continues to impose substantial health and financial costs on society. According to the Centers for Disease Control and Prevention (CDC), in USA alone, the economic burden caused by smoking to society, including both the direct health care expenditures and the loss of productivity, can be as high as $193 billion a year [6]. In china, the prevalence of smoking remains high with 350 million smokers, and it is estimated that, by 2025, the annual number of deaths attributed to tobacco use will increase from 1.2 million to 2 million [7]. Although many cigarette smokers report a desire to quit smoking [8], few are successful [9, 10]. Thus, developing effective therapeutic approaches that can help smokers achieve and sustain abstinence from smoking, as well as methods that can prevent people from starting smoking, remains a huge challenge in public health.

Nicotine, as the primary psychoactive component of tobacco smoke, produces diverse neurophysiological, motivational, and behavioural effects through interactions with nicotinic acetylcholine receptors (nAChRs) in the central nervous system (CNS). Twins, family and adoption studies have suggested that nicotine addiction is closely related to genetic and environmental factors, and genetic factors play an important role in the risk to the development of addiction [11, 12]. Numerous studies aiming to identify the genetic variants or candidate genes have found a large number of promising genes and chromosomal regions involved in the etiology of nicotine addiction [13]. In addition, various pathways and neurotransmitter systems have been found to be related to the psychoactive and addictive properties of nicotine, such as the mesocorticolimbic dopamine system [14–16], the serotonin

system, the glutamate system, and the GABA system [17–19]. Further, emerging evidence suggests that nicotine can also regulate the expression of genes/proteins involved in various functions such as ERK1/2, CREB, and c-FOS [20–22], as well as the expression state of multiple biochemical pathways, for example, mitogen-activated protein kinase (MAPK), phosphatidylinositol phosphatase signaling, growth factor signaling, and ubiquitin-proteasome pathways [23–25].

During the past decade, the application of high-throughput technologies to nicotine addiction study has greatly enhanced our ability to identify the nicotine addiction-related molecular factors [26–28]. In spite of these progresses, our understanding of the molecular mechanism underlying nicotine addiction is still incomplete. Under such situation, how to integrate the available knowledge and data in heterogeneous datasets to obtain the relevant biological information has become an important task. Among the available approaches to explore the molecular mechanisms underlying various complex diseases, investigating the interactions between proteins encoded by the candidate genes in the human protein-protein interaction (PPI) network has been emerging as a powerful way [29–31]. Furthermore, genes/proteins with similar functions usually interact with each other more closely than those functionally unrelated genes [32], and cluster analysis on the molecular candidates within a PPI network can provide an intuitive view to understand its major biological functions. Taking together, a comprehensive analysis of the candidate genes within a systematic framework may be a powerful approach to analyze the molecular mechanisms underlying complex diseases like nicotine addiction.

In this study, the global network topological properties of nicotine addiction-related genes (NAGenes) were explored in the context of human PPI network and were compared with other gene sets. Then, cluster analysis was utilized to detect the major functional modules related to nicotine addiction in the PPI network. Additionally, the significantly enriched functional clusters were identified for the NAGenes. This study provides useful insights for understanding the molecular mechanisms of nicotine addiction at the systems biological level.

2. Materials and Methods

2.1. Data Sources. Multiple gene sets related to nicotine abuse have been reported [27, 33, 34]. In an earlier study, we obtained 220 NAGenes prioritized via a multisource-based gene approach [35], which represented a relatively comprehensive gene set for nicotine addiction. Briefly, genes identified to be related to nicotine addiction or involved in the physiological response to nicotine exposure or smoking behaviors were collected by integrating four categories of evidence, that is, association studies, linkage analysis, gene expression analysis, and literature search of single gene/protein-based studies. A category-specific score was assigned to each gene and a combined score was computed for all the collected genes based on an optimized weight matrix. Then, the genes were ranked according to the combined scores with a larger score value indicating a potentially higher correlation between the gene and nicotine addiction. Based on the distribution of the combined score of all the genes collected, 220 genes on the top of the list were selected as the prioritized NAGenes.

For the purpose of comparison, we collected two other gene sets, that is, an alcohol addiction-related gene set (alcohol genes) and a cancer-related gene set (cancer genes). Alcohol addiction can evoke the dysfunction of neuronal system and has been suggested to share some biological mechanisms with nicotine addiction. In this study, we selected the gene set with 316 alcohol genes collected by Li et al. [33]. Cancer has been well studied and is expected to have substantially different pathological characteristics from nicotine addiction. We downloaded the cancer genes (522 genes) from the Cancer Gene Census database (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/).

To investigate the network topological characteristics of a gene set, we first need to construct a relatively comprehensive and reliable PPI network. Here, we downloaded the human PPI data from the Protein Interaction Network Analysis (PINA) platform (May 21, 2014) [36], which collected and annotated data from six major protein interaction databases, that is, IntAct, BioGRID, MINT, DIP, HPRD, and MIPS/MPact. Also, we downloaded several related annotation files from NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/) (May 24, 2014), including the Entrez gene information database of human (Homo_sapiens.gene_info.gz), the data set specifying relationship between pairs of NCBI and UniProtKB protein accessions (gene_refseq_uniprotkb_collab.dz), and file containing mappings of Entrez Gene records to Entrez RefSeq Nucleotide sequence records (gene2refseq.gz). For the proteins included in the human PPI database, only those that could be mapped to NCBI Entrez Gene were included in our subsequent analysis. After excluding the redundant and self-interacting pairs, we constructed a human PPI network containing 15,093 nodes and 161,419 edges.

2.2. Global Network Topological Properties. In network analysis, different metrics can be used to describe the network characteristics. We applied four measures to assess the network topological characteristic of NAGenes, that is, degree and degree distribution, clustering coefficient, closeness, and eccentricity. For a network, degree of a node (gene/protein in our case) is the number of direct connections that it has to other nodes in the network, and highly linked nodes are usually thought to make important contribution to the global structure or the behavior organization of a biological network [37, 38]. Degree distribution is the probability distribution of the degrees of all nodes over the whole network. Clustering coefficient quantifies the probability that two nodes linking to the same node connect with each other and describes the overall organization of the relationships within a network [39, 40]. The closeness of a node is the reciprocal of its average distance to each node in the network, while the eccentricity of a node is the distance to its farthest reachable node [41].

2.3. Cluster Analysis within the Global Network. To intuitively observe the biological functions involved in the large nicotine

addiction-related network, we applied the Molecular Complex Detection (MCODE) (Version 1.4) (http://baderlab.org/Software/MCODE) implemented in Cytoscape platform (http://www.cytoscape.net/) to identify the molecule modules or clusters. MCODE is a local clustering algorithm that can effectively detect densely connected regions of a molecular interaction network. In our analysis, the global network that we constructed was uploaded into the Cytoscape [42] and then MCODE was run to detect gene clusters in the network using the haircut option which identified nodes having limited connectivity at the cluster periphery. For the other parameters, the default settings were adopted.

2.4. Functional Annotation Cluster. To assess the candidate genes in the context of function similarity, we performed enrichment analysis on their Gene Ontology (GO) annotations using the Database for Annotation and Integrated Discovery (DAVID) (Version 6.7) [43]. The genes with their gene ID or GenBank Accession Numbers were submitted to DAVID under the functional annotation option specifying Homo sapiens as the species. In the DAVID functional annotation clustering, the significantly overrepresented GO terms, that is, biological process (BP), molecular function (MF), and cellular component (CC), were retrieved by using the options GOTERM_BP_ALL, GOTERM_MF_ALL and GOTERM_CC_ALL. The default parameters and corresponding false discovery rate (FDR) by the Benjamini and Hochberg approach [44] were used to determine the enrichment score.

3. Results and Discussions

3.1. Global Network Topological Properties of NAGenes. PPI network analysis provides an effective approach to investigate the biological themes related to a list of genes at the molecular level. In particular, the topological properties of nodes (genes) and edges (connections between genes) can help to understand the underlying biological themes associated with the network [45]. To depict the network topological properties of NAGenes, we first constructed a human PPI network by integrating information from multiple databases, to which NAGenes were then mapped. Subsequently, the characteristics of the NAGenes in the network were assessed by four network topological measurements, that is, degree, clustering coefficient, closeness, and eccentricity. As a comparison, we also calculated the topological measures of the networks corresponding to alcohol genes and cancer genes.

Of the 220 NAGenes, 208 could be mapped onto the human PPI network and the average degree of these genes was 39.1, which measured the average number of direct connections between each member of NAGenes and other genes included in the PPI network, while, for the alcohol genes, 304 of the 316 genes could be mapped onto the human PPI network, with an average degree of 52.9 and for the cancer genes, 488 of the 519 genes could be mapped onto the human PPI network, with an average degree of 59.8. In order to have a more intuitive understanding of the degree characteristics, we plotted the degree distributions of the three gene sets (Figure 1). As shown, for all the three gene sets, the degrees scattered in a rather large range from 1 to more than 500. But the degree distributions were right-skewed, that is, the majority of the genes had only a few connections with other genes and a small number of genes had a large number of connections. Compared with the NAGenes, the average degree of the alcohol genes appeared to be closer to the cancer genes, but statistical test indicated that significant difference existed between the degrees of all the three gene sets (alcohol genes versus cancer genes, $P = 1.93 \times 10^{-7}$; alcohol genes versus NAGenes, $P = 0.0031$, Wilcoxon rank sum test). The degree distribution of NAGenes was also significantly different from that of both alcohol genes and cancer genes (NAGenes versus alcohol genes, $P = 0.0031$; NAGenes versus cancer genes, $P = 1.93 \times 10^{-13}$, Wilcoxon rank sum test). But compared with the cancer genes, the NAGenes and the alcohol genes tended to have lower or moderate connections, for example, 67% and 54% of the NAGenes, and the alcohol genes fell in the degree interval of 1–20, respectively, while only 37% of the cancer genes were included in this range (Figure 2). A close check of the degree of NAGenes showed that genes with more specific functions, such as those related to synaptic transmission (e.g., neuronal acetylcholine receptor subunit alpha-1 [CHRNA1], CHRNA2, CHRNB1, and CHRNB2), drug metabolism (e.g., N-acetyltransferase 2 [NAT2], tryptophan hydroxylase 2 [TPH2], and cytochrome P450 2A6 [CYP2A6]), and transport (e.g., solute carrier family 9 member 9 [SLC9A9], solute carrier organic anion transporter family member 3A1 [SLCO3A1], and solute carrier family 1 member 2 [SLC1A2]), tended to have smaller degrees, while the genes expressed in a large range of cell types/tissues or involved in broad physiological processes were more likely to larger degrees, for example, nuclear receptor subfamily 3 group C member 1 (NR3C1), beta-2 adrenergic receptor (ADRB2), estrogen receptor alpha (ESR1), and tumor protein p53 (TP53). Thus, although all the members of NAGenes may be nicotine addiction-related, those with smaller degrees are more likely to be involved in biological processes or neuronal activities invoked by nicotine.

Clustering coefficient measures the interconnectivity of neighboring genes in a network. Generally, a gene with larger clustering coefficient has a higher density of network connection. The average clustering coefficients of NAGenes, alcohol genes, and cancer genes were 0.02, 0.03, and 0.06, respectively. To better describe the characteristics of the clustering coefficient, we summarized them using histogram with an interval of 0.1 (Figure 3(a)). Among the three gene sets, the proportion of genes with clustering coefficient of 0 was much higher for NAGenes (67.8%) than the alcohol genes (44.1%) and cancer genes (16.0%). Within the intervals 0-0.1, the proportion of NAGenes included was 96.2%, which is higher than the other two gene sets (alcohol: 95.7%; cancer: 81.6%). Interestingly, when the clustering coefficient was greater than 0.4, the proportion of NAGenes was 0. Thus, NAGenes were likely to be less connected with each other than the alcohol genes or the cancer genes. In addition, we also analyzed the distribution of closeness and eccentricity of the NAGenes in the human PPI network. Usually, a gene with higher closeness is more likely to be a central gene in the network, and a gene with larger eccentricity

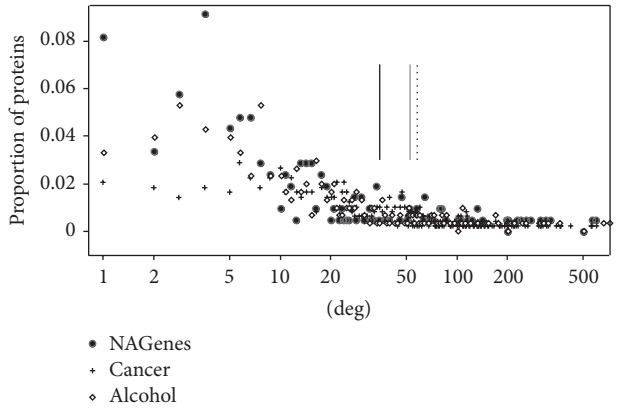


Figure 1: Degree distribution and the average degree of NAGenes, alcohol genes, and cancer genes. y-axis represents the proportion of proteins having a specific degree. Vertical line represents the average value of the degrees. Black line denotes NAGenes, gray line denotes alcohol genes, and dotted line denotes cancer genes.

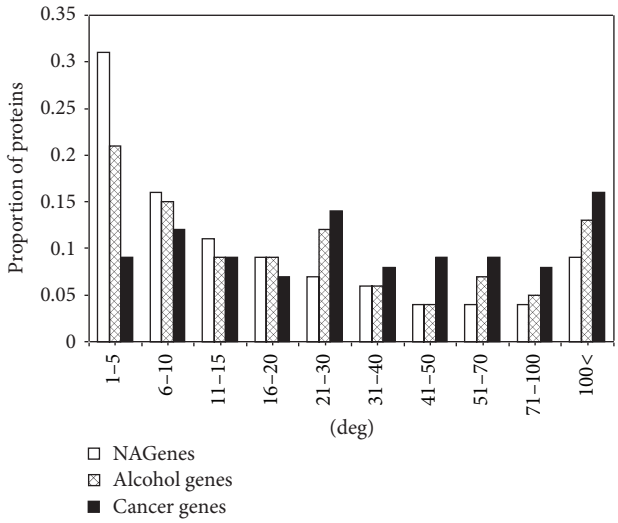


Figure 2: Degree distribution of NAGenes, alcohol genes, and cancer genes. y-axis represents the proportion of proteins having a specific degree.

is closer to the fringe of the network [46, 47]. Figure 3(b) showed that NAGenes had a smaller closeness compared with the alcohol genes or the cancer genes, but the eccentricity distribution of NAGenes showed an opposite trend, following a more right-skewed distribution (Figure 3(c)). These results revealed that the NAGenes may be less central in the PPI network compared with the other two gene sets.

3.2. Cluster Analysis within the Global Network of NAGenes. Besides characterizing the interaction networks with respect to their topological features, the biological network can also be clustered or partitioned into modules, which provides an insight into the overall organization of the relationship within the PPI network [32]. Clustering algorithms have previously been shown to be useful in predicting the molecular modules that participate in similar biological process.

By using the clustering algorithm to the network associated with nicotine addiction, we identified 6 clusters including 81 nodes (genes in our case) and 126 edges. Out of these nodes, 30 (37.04%) were included in the 208 genes mapping into the human PPI network. These clusters were ranked according to their density and the number of proteins (genes) included (Table 1 and Figure 4). As shown, the clusters were involved in multiple biological functional categories. For example, the majority of the genes in cluster I were associated with apoptotic and macromolecular metabolic process. Three genes associated with nicotine addiction, estrogen receptor 1 (ESR1), arrestin beta 1 (ARRB1), and ARRB2, were located close to the center of this cluster (Figure 4). ESR1, as the specific nuclear receptor of sex hormones, widely distributes in the dopaminergic midbrain neurons and is able to modulate the neurotransmitter systems of the brain reward circuitry [48]. Moreover, ESR1 also plays an important role in apoptotic process. ARRB1 and ARRB2 are ubiquitous scaffolding proteins. They can regulate multiple intracellular signaling proteins involved in cell proliferation and differentiation and have important roles in mitogenic and antiapoptotic function of nicotine [49, 50]. The overall functional theme of Clusters II, III, and VI was synaptic transmission. Dopamine receptor D2 (DRD2) and DRD4 are both dopamine receptors that are critical for the reinforcing effects or rewarding behaviors of nicotine [51, 52]. GABA B receptor 1 (GABBR1) and GABBR2, the two receptors of the major inhibitory neurotransmitter GABA, play important roles in the development of nicotine addiction [53].

Each cluster also contained genes not included in NAGenes (Figure 4). A close inspection showed that some of these additional genes were potentially related to nicotine addiction. For example, N-ethylmaleimide-sensitive factor (NSF) [54], ubiquitin b (UBB) [55], small ubiquitin-related modifier 2 (SUMO2) [55], cyclin-dependent kinase 5 (CDK5) [56], and phospholipase C gamma 1 (PLCG1) [57] have been reported to be associated with nicotine addiction or regulated by nicotine exposure. Thus, further exploration on the genes included in these clusters may help us to identify more nicotine addiction-related candidate genes.

3.3. Functional Annotation Analysis. To obtain a more systematic view of the biological function of the genes involved in nicotine addiction, we performed functional enrichment analysis on NAGenes. In earlier study, a preliminary functional annotation analysis showed that genes related to biological processes like neurodevelopment and signal transduction were overrepresented in NAGenes [35]. Here, we provided a more comprehensive exploration on the function features of these genes. For the 220 genes, 73 annotation clusters were identified in the candidate genes (enrichment score > 1.3). Of these annotation clusters, eight clusters with enrichment scores higher than 10 were displayed with the representative GO terms (Figure 5 and Table S1). From a wide view of the annotation clusters, functional annotations associated with neurodevelopment and neurotransmitters were significantly overrepresented in the NAGenes. In the top two annotation clusters (Clusters 1 and 2), eight terms, including transmission of nerve impulse (FDR = 1.85×10^{-28}), synaptic

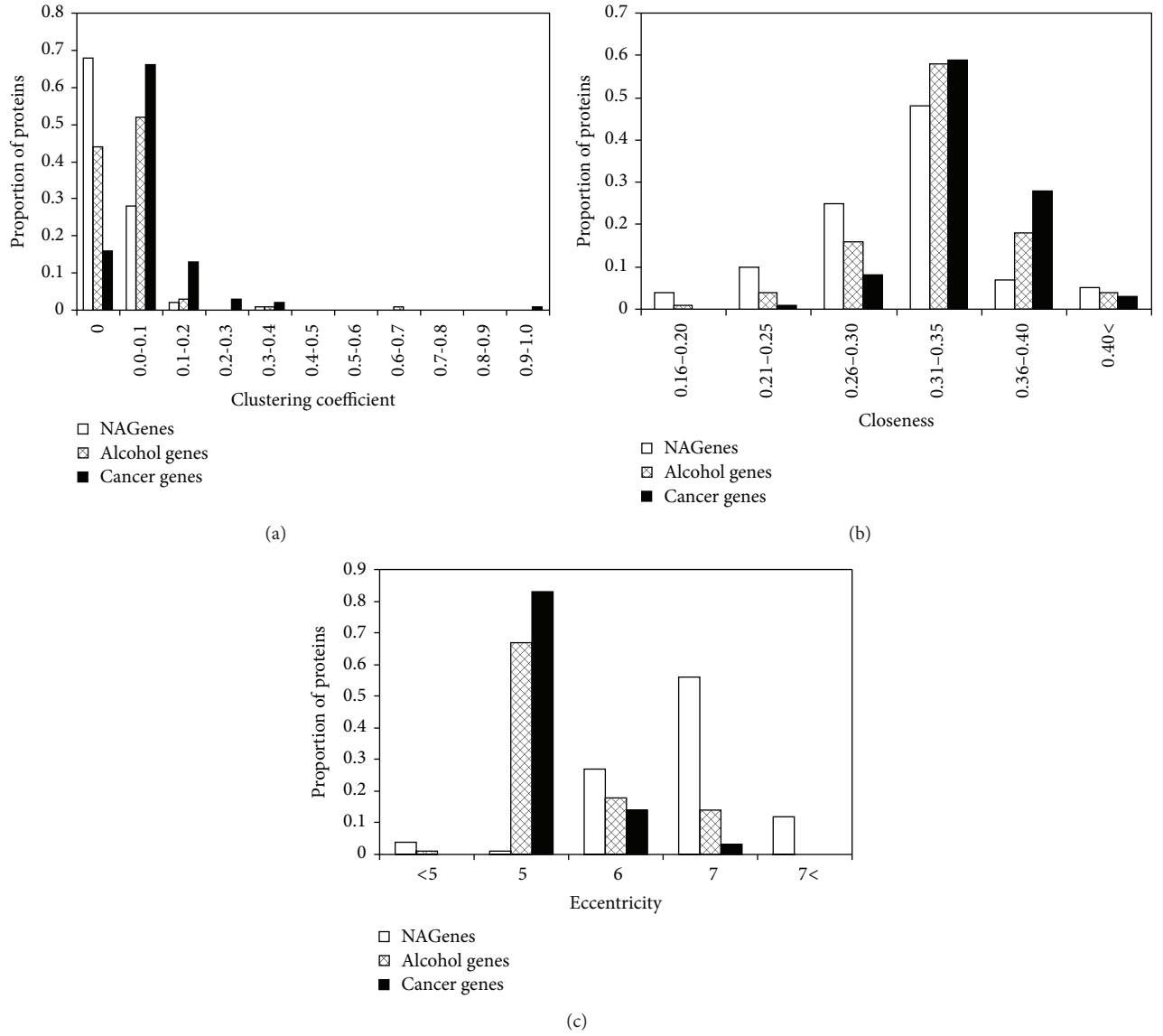


FIGURE 3: Topological measures distribution of NAGenes, alcohol genes, and cancer genes. y-axis represents the proportion of proteins having a specific measurement. (a) Clustering coefficient. (b) Closeness. (c) Eccentricity.

transmission (FDR = 3.32×10^{-28}), system process (FDR = 3.84×10^{-19}), and neurological system process (FDR = 2.76×10^{-18}), were directly related to neurodevelopment, consistent with the previous reports that there is a relationship between the pathology of nicotine addiction and the development of neuron system. Moreover, the majority of terms in Cluster 3 were associated with neurotransmitter receptor or channel activity, for example, extracellular ligand-gated ion channel activity (FDR = 2.28×10^{-19}), neurotransmitter receptor activity (FDR = 5.80×10^{-18}), and acetylcholine receptor activity (FDR = 6.06×10^{-18}) (Table S1). These results indicated the importance of neurotransmitters and related molecules in the development of nicotine addiction. Importantly, we found that calcium ion transport (FDR = 0.02) was also overrepresented in the candidate genes, consistent with the reports that the ligand-gated cation channels play an important role in regulating various neuronal activities by mediating intracellular Ca^{2+} concentration, including neurotransmitter release [58, 59]. In Cluster 7, the overall functional theme was various neurotransmitter or substances metabolic process, such as dopamine metabolic process (FDR = 1.76×10^{-12}), catecholamine metabolic process (FDR = 6.58×10^{-11}), diol metabolic process (FDR = 6.58×10^{-11}), and cellular amino acid derivative metabolic process (FDR = 5.21×10^{-6}). These metabolic processes had important roles not only in the development of nicotine addiction, but also in the harm to human health. In addition, Cluster 8 was concentrated on learning or memory, which reflected a kind of pathological forms of nicotine addiction. In summary, the molecular mechanisms underlying nicotine addiction are

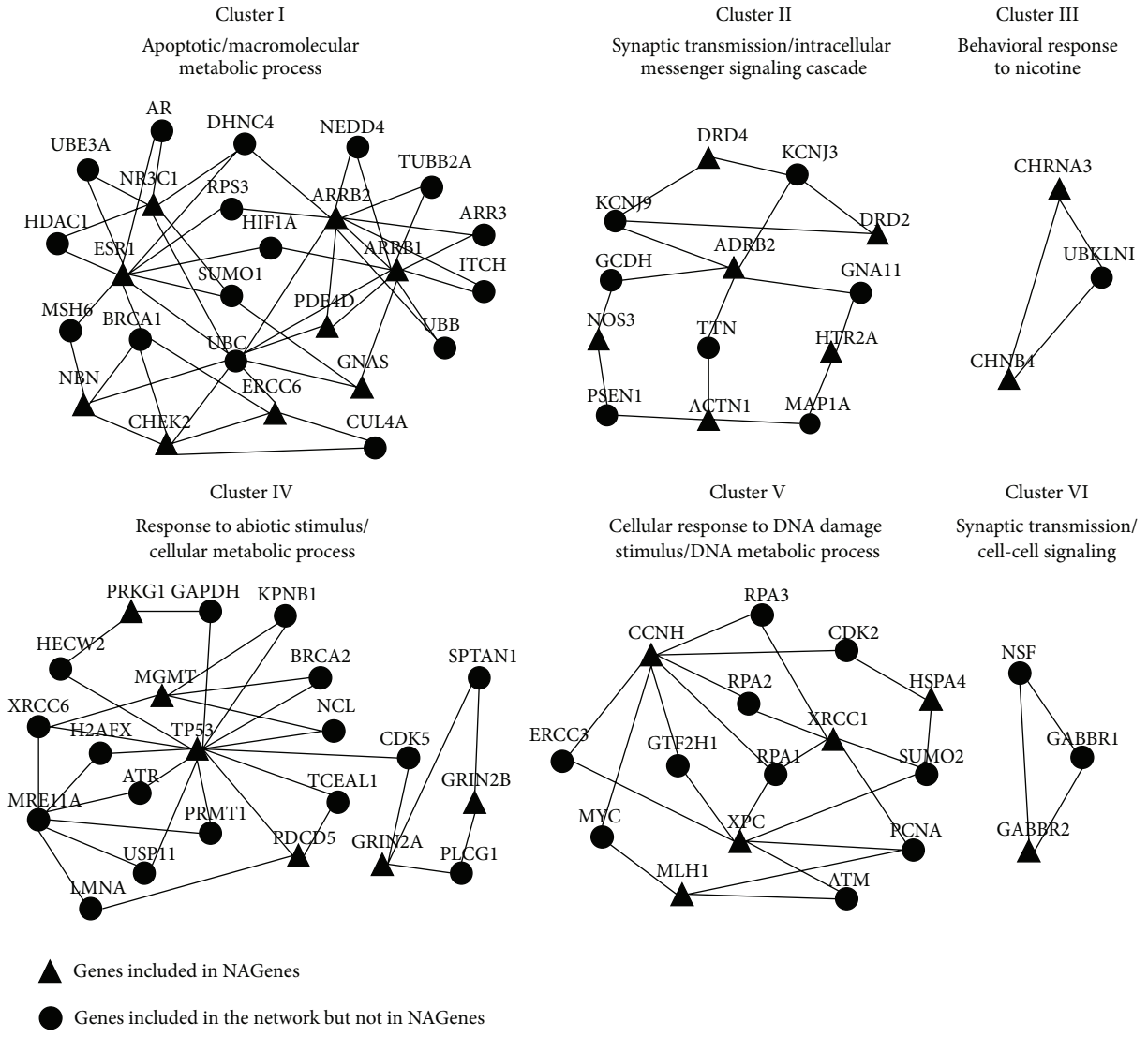


Figure 4: Gene clusters identified by MCODE. NAGenes are shown as triangular nodes and non-NAGenes are ellipse nodes. The functional descriptors of each cluster are based on Gene Ontology term.

Table 1: Gene clusters identified in the nicotine addiction-related network.

Cluster	Cluster function	Score[a]	Nodes	Edges	Gene symbol
I	Apoptotic/macromolecular metabolic process	4.08	25	49	ARRB2, ARRB1, CUL4A, HDAC1, RPS3, ERCC6, GNAS, UBE3A, NBN, CHEK2, BRCA1, ESR1, ARR3, AR, HDAC2, NEDD4, UBB, MSH6, NR3C1, UBC, PDE4D, SUMO1, HIF1A, TUBB2A, ITCH
II	Synaptic transmission/intracellular and second messenger signaling cascade	2.67	13	16	KCNJ9, DRD2, ADRB2, DRD4, NOS3, MAP1A, GCDH, TTN, HTR2A, PSEN1, ACTN1, KCNJ3, GNA11
III	Behavioral response to nicotine	3.00	3	3	UBQLN1, CHRNB4, CHRNA3
IV	Response to abiotic stimulus/cellular metabolic process	3.05	22	32	BRCA2, GAPDH, H2AFX, SPTAN1, PRMT1, PRKG1, MGMT, NCL, HECW2, USP11, ATR, LMNA, GRIN2A, CDK5, TP53, GRIN2B, KPNB1, XRCC6, MRE11A, TCEAL1, PLCG1, PDCD5
V	Cellular response to DNA damage stimulus/DNA metabolic process	3.29	15	23	RPA2, RPA3, CCNH, HSPA4, ERCC3, XRCC1, RPA1, GTF2H1, MLH1, PCNA, MYC, XPC, ATM, CDK2, SUMO2
VI	Synaptic transmission/cell-cell signaling	3.00	3	3	NSF, GABBR1, GABBR2

[a]Score is defined as the product of the cluster density and the number of vertices (proteins) in the cluster (DC × $|V|$).

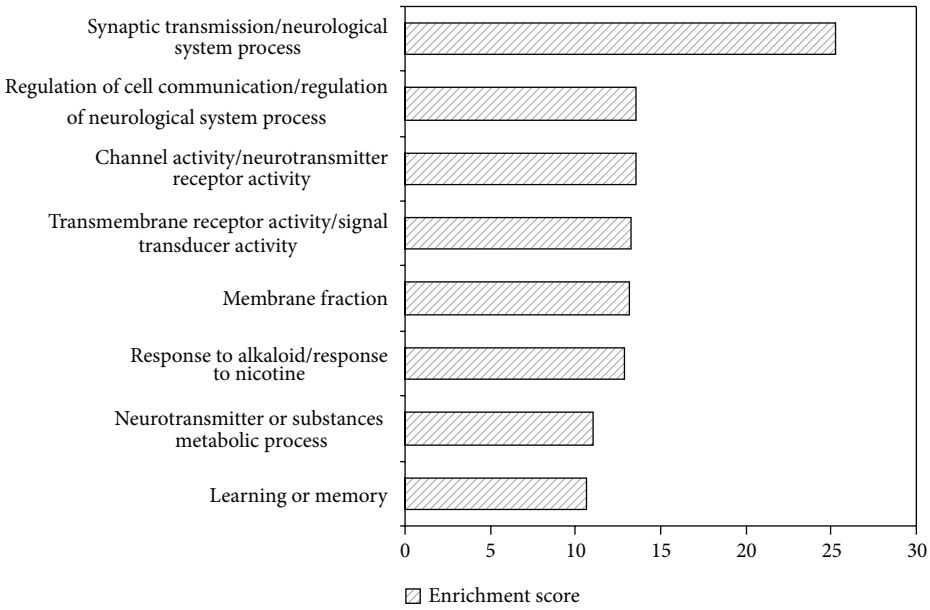


Figure 5: Enriched functional annotation in NAGenes (enrichment score > 10). Detailed information can be seen in supplementary Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/313709.

extremely complex in that they involve many genes and biological functions. Through its direct or indirect interactions with these genes, nicotine can regulate various physiological processes, such as learning and memory, synaptic function, response to stress, and addiction [60–63]. Our results also demonstrated that functional annotation cluster analysis can provide useful insights for intuitive understanding of addiction mechanisms. Furthermore, as neurodevelopment system and neuronal signaling cascades in the brain play important roles in the pathology of nicotine addiction, the genes and pathways related to these biological processes should be the major targets in nicotine addiction study.

4. Conclusions

To achieve better understanding of the molecular mechanisms underlying nicotine addiction, it is necessary to adopt a system biology frame to analyze the candidate genes related to nicotine addiction. In this study, we explored the global network topological characteristics of nicotine addiction. The results revealed that the topological features of NAGenes were significantly different from alcohol genes and cancer genes. Specifically, NAGenes tended to have a more moderate degree and weaker clustering coefficient and they were likely to be in the network margin. Further, integrating the information from the functional modules identified in the global network and annotation cluster analysis, we found that nicotine addiction was involved in many biological functions, such as neurodevelopment, neurotransmitters activity, and various metabolic processes. Our preliminary results present a wealth of potential functional information underlying the mechanism of nicotine addiction and they are valuable for further investigation.

Conflict of Interests

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

Acknowledgments

This project was supported in part by Grants from National Natural Science Foundation of China (Grant no. 31271411) and Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry of China. The authors are grateful to Professor Ming D. Li of the University of Virginia for his help on this study.

References

[1] A. S. Abdullah, F. Qiming, V. Pun, F. A. Stillman, and J. M. Samet, "A review of tobacco smoking and smoking cessation practices among physicians in China: 1987–2010," *Tobacco Control*, vol. 22, no. 1, pp. 9–14, 2013.

[2] M. Sopori, "Effects of cigarette smoke on the immune system," *Nature Reviews Immunology*, vol. 2, no. 5, pp. 372–377, 2002.

[3] P. Jha and R. Peto, "Global effects of smoking, of quitting, and of taxing tobacco," *The New England Journal of Medicine*, vol. 370, no. 1, pp. 60–68, 2014.

[4] P. Jha, "Avoidable global cancer deaths and total deaths from smoking," *Nature Reviews Cancer*, vol. 9, no. 9, pp. 655–664, 2009.

[5] S. Murray, "A smouldering epidemic," *Canadian Medical Association Journal*, vol. 174, no. 3, pp. 309–310, 2006.

[6] B. Adhikari, J. Kahende, A. Malarcher, T. Pechacek, and V. Tong, "Smoking-attributable mortality, years of potential life lost, and productivity losses," *Oncology Times*, vol. 31, no. 2, pp. 40–42, 2009.

[7] J. Zhang, J.-X. Ou, and C.-X. Bai, "Tobacco smoking in China: prevalence, disease burden, challenges and future strategies," *Respirology*, vol. 16, no. 8, pp. 1165–1172, 2011.

[8] CDC, "Cigarette smoking among adults—United States," *Morbidity and Mortality Weekly Report*, vol. 56, no. 44, pp. 1157–1161, 2006.

[9] J. R. Hughes, L. F. Stead, and T. Lancaster, "Antidepressants for smoking cessation," *Cochrane Database of Systematic Reviews*, no. 1, Article ID CD000031, 2007.

[10] C. E. Lerman, R. A. Schnoll, and M. R. Munafò, "Genetics and smoking cessation—improving outcomes in smokers at risk," *American Journal of Preventive Medicine*, vol. 33, no. 6, pp. S398–S405, 2007.

[11] H. H. Maes, P. F. Sullivan, C. M. Bulik et al., "A twin study of genetic and environmental influences on tobacco initiation, regular tobacco use and nicotine dependence," *Psychological Medicine*, vol. 34, no. 7, pp. 1251–1261, 2004.

[12] M. D. Li and M. Burmeister, "New insights into the genetics of addiction," *Nature Reviews Genetics*, vol. 10, no. 4, pp. 225–231, 2009.

[13] J. Sun and Z. Zha, "Functional features, biological pathways, and protein interaction networks of addiction-related genes," *Chemistry & Biodiversity*, vol. 7, no. 5, pp. 1153–1162, 2010.

[14] M. E. M. Benwell and D. J. K. Balfour, "Regional variation in the effects of nicotine on catecholamine overflow in rat brain," *European Journal of Pharmacology*, vol. 325, no. 1, pp. 13–20, 1997.

[15] A. Tammimäki, K. Pietilä, H. Raattamaa, and L. Ahtee, "Effect of quinpirole on striatal dopamine release and locomotor activity in nicotine-treated mice," *European Journal of Pharmacology*, vol. 531, no. 1–3, pp. 118–125, 2006.

[16] H. Gäddnäs, K. Pietilä, T. P. Piepponen, and L. Ahtee, "Enhanced motor activity and brain dopamine turnover in mice during long-term nicotine administration in the drinking water," *Pharmacology Biochemistry and Behavior*, vol. 70, no. 4, pp. 497–503, 2001.

[17] J. Barik and S. Wonnacott, "Indirect modulation by $\alpha 7$ nicotinic acetylcholine receptors of noradrenaline release in rat hippocampal slices: interaction with glutamate and GABA systems and effect of nicotine withdrawal," *Molecular Pharmacology*, vol. 69, no. 2, pp. 618–628, 2006.

[18] P. J. Kenny and S. E. File, "Nicotine regulates 5-HT_{1A} receptor gene expression in the cerebral cortex and dorsal hippocampus," *European Journal of Neuroscience*, vol. 13, no. 6, pp. 1267–1271, 2001.

[19] J. Barik and S. Wonnacott, "Molecular and cellular mechanisms of action of nicotine in the CNS," *Handbook of Experimental Pharmacology*, vol. 192, pp. 173–207, 2009.

[20] D. H. Brunzell, D. S. Russell, and M. R. Picciotto, "*In vivo* nicotine treatment regulates mesocorticolimbic CREB and ERK signaling in C57B1/6J mice," *Journal of Neurochemistry*, vol. 84, no. 6, pp. 1431–1441, 2003.

[21] S. R. Pagliusi, M. Tessari, S. DeVevey, C. Chiamulera, and E. M. Pich, "The reinforcing properties of nicotine are associated with a specific patterning of c-fos expression in the rat brain," *European Journal of Neuroscience*, vol. 8, no. 11, pp. 2247–2256, 1996.

[22] M. Nisell, G. G. Nomikos, K. Chergui, P. Grillner, and T. H. Svensson, "Chronic nicotine enhances basal and nicotine-induced Fos immunoreactivity preferentially in the medial prefrontal cortex of the rat," *Neuropsychopharmacology*, vol. 17, no. 3, pp. 151–161, 1997.

[23] K. Tang, H. Wu, S. K. Mahata, and D. T. O'Connor, "A crucial role for the mitogen-activated protein kinase pathway in nicotinic cholinergic signaling to secretory protein transcription in pheochromocytoma cells," *Molecular Pharmacology*, vol. 54, no. 1, pp. 59–69, 1998.

[24] M. D. Li, J. K. Kane, J. Wang, and J. Z. Ma, "Time-dependent changes in transcriptional profiles within five rat brain regions in response to nicotine treatment," *Molecular Brain Research*, vol. 132, no. 2, pp. 168–180, 2004.

[25] Ö. Konu, J. K. Kane, T. Barrett et al., "Region-specific transcriptional response to chronic nicotine in rat brain," *Brain Research*, vol. 909, no. 1-2, pp. 194–203, 2001.

[26] W. Huang and M. D. Li, "Nicotine modulates expression of miR-140, which targets the 3-untranslated region of dynamin 1 gene (Dnm1)," *International Journal of Neuropsychopharmacology*, vol. 12, no. 4, pp. 537–546, 2009.

[27] M. D. Li and M. Burmeister, "New insights into the genetics of addiction," *Nature Reviews Genetics*, vol. 10, no. 4, pp. 225–231, 2009.

[28] J. Wang, J.-M. Kim, D. M. Donovan, K. G. Becker, and M. D. Li, "Significant modulation of mitochondrial electron transport system by nicotine in various rat brain regions," *Mitochondrion*, vol. 9, no. 3, pp. 186–195, 2009.

[29] J. Sun, P. Jia, A. H. Fanous et al., "Schizophrenia gene networks and pathways and their applications for novel candidate gene selection," *PLoS ONE*, vol. 5, no. 6, Article ID e11351, 2010.

[30] P. Jia, C.-F. Kao, P.-H. Kuo, and Z. Zhao, "A comprehensive network and pathway analysis of candidate genes in major depressive disorder," *BMC Systems Biology*, vol. 5, supplement 3, article S12, 2011.

[31] J. Sun and Z. Zhao, "A comparative study of cancer proteins in the human protein-protein interaction network," *BMC Genomics*, vol. 11, supplement 3, article S5, 2010.

[32] J. Song and M. Singh, "How and when should interactome-derived clusters be used to predict functional modules and protein function?" *Bioinformatics*, vol. 25, no. 23, pp. 3143–3150, 2009.

[33] C.-Y. Li, X. Mao, and L. Wei, "Genes and (common) pathways underlying drug addiction," *PLoS Computational Biology*, vol. 4, no. 1, article e2, 2008.

[34] J. Wang and M. D. Li, "Common and unique biological pathways associated with smoking initiation/progression, nicotine dependence, and smoking cessation," *Neuropsychopharmacology*, vol. 35, no. 3, pp. 702–719, 2010.

[35] X. Liu, M. Liu, X. Li, L. Zhang, R. Fan, and J. Wang, "Prioritizing genes related to nicotine addiction via a multi-source-based approach," *Molecular Neurobiology*, 2014.

[36] J. Wu, T. Vallenius, K. Ovaska, J. Westermarck, T. P. Mäkelä, and S. Hautaniemi, "Integrated network analysis platform for protein-protein interactions," *Nature Methods*, vol. 6, no. 1, pp. 75–77, 2009.

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

[38] J. Dong and S. Horvath, "Understanding network concepts in modules," *BMC Systems Biology*, vol. 1, article 24, 2007.

[39] C. C. Friedel and R. Zimmer, "Inferring topology from clustering coefficients in protein-protein interaction networks," *BMC Bioinformatics*, vol. 7, no. 1, article 519, 2006.

[40] A. Delprato, "Topological and functional properties of the small GTPases protein interaction network," *PLoS ONE*, vol. 7, no. 9, Article ID e44882, 2012.

[41] S. Chavali, F. Barrenas, K. Kanduri, and M. Benson, "Network properties of human disease genes with pleiotropic effects," *BMC Systems Biology*, vol. 4, no. 1, article 78, 2010.

[42] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics*, vol. 27, no. 3, pp. 431–432, 2011.

[43] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[44] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 57, no. 1, pp. 289–300, 1995.

[45] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[46] G. Lima-Mendez, J. Van Helden, A. Toussaint, and R. Leplae, "Reticulate representation of evolutionary and functional relationships between phage genomes," *Molecular Biology and Evolution*, vol. 25, no. 4, pp. 762–777, 2008.

[47] N. Pržulj, D. G. Corneil, and I. Jurisica, "Efficient estimation of graphlet frequency distributions in protein-protein interaction networks," *Bioinformatics*, vol. 22, no. 8, pp. 974–980, 2006.

[48] M. R. Munafò and E. C. Johnstone, "Genes and cigarette smoking," *Addiction*, vol. 103, no. 6, pp. 893–904, 2008.

[49] P. Dasgupta, W. Rizwani, S. Pillai et al., "ARRB1-mediated regulation of E2F target genes in nicotine-induced growth of lung tumors," *Journal of the National Cancer Institute*, vol. 103, no. 4, pp. 317–333, 2011.

[50] C. Nordenvall, P. J. Nilsson, W. Ye, T. M.-L. Andersson, and O. Nyrén, "Tobacco use and cancer survival: a cohort study of 40,230 Swedish male construction workers with incident cancer," *International Journal of Cancer*, vol. 132, no. 1, pp. 155–161, 2013.

[51] G. di Chiara, V. Bassareo, S. Fenu et al., "Dopamine and drug addiction: the nucleus accumbens shell connection," *Neuropharmacology*, vol. 47, no. 1, pp. 227–241, 2004.

[52] R. B. Free, L. A. Hazelwood, D. M. Cabrera et al., "D1 and D2 dopamine receptor expression is regulated by direct interaction with the chaperone protein calnexin," *Journal of Biological Chemistry*, vol. 282, no. 29, pp. 21285–21300, 2007.

[53] N. L. Benowitz, "Nicotine addiction," *The New England Journal of Medicine*, vol. 362, no. 24, pp. 2295–2303, 2010.

[54] Y. Y. Hwang and M. D. Li, "Proteins differentially expressed in response to nicotine in five rat brain regions: identification using a 2-DE/MS-based proteomics approach," *Proteomics*, vol. 6, no. 10, pp. 3138–3153, 2006.

[55] J. K. Kane, Ö. Konu, J. Z. Ma, and M. D. Li, "Nicotine coregulates multiple pathways involved in protein modification/degradation in rat brain," *Molecular Brain Research*, vol. 132, no. 2, pp. 181–191, 2004.

[56] M. Hamada, J. P. Hendrick, G. R. Ryan et al., "Nicotine regulates DARPP-32 (dopamine- and cAMP-regulated phosphoprotein of 32 kDa) phosphorylation at multiple sites in neostriatal neurons," *Journal of Pharmacology and Experimental Therapeutics*, vol. 315, no. 2, pp. 872–878, 2005.

[57] S. C. Pandey, "Effects of chronic nicotine treatment on the expression of phospholipase C isozymes and the alpha subunit of $G_{q/11}$ protein in the rat brain," *Neuroscience Letters*, vol. 212, no. 2, pp. 127–130, 1996.

[58] S. Wonnacott, "Presynaptic nicotinic ACh receptors," *Trends in Neurosciences*, vol. 20, no. 2, pp. 92–98, 1997.

[59] D. L. Marshall, P. H. Redfern, and S. Wonnacott, "Presynaptic nicotinic modulation of dopamine release in the three ascending pathways studied by in vivo microdialysis: comparison of naive and chronic nicotine-treated rats," *Journal of Neurochemistry*, vol. 68, no. 4, pp. 1511–1519, 1997.

[60] E. D. Levin, "Nicotinic systems and cognitive function," *Psychopharmacology*, vol. 108, no. 4, pp. 417–431, 1992.

[61] C. M. Hernandez and A. V. Terry Jr., "Repeated nicotine exposure in rats: effects on memory function, cholinergic markers and nerve growth factor," *Neuroscience*, vol. 130, no. 4, pp. 997–1012, 2005.

[62] F. Dajas-Bailador and S. Wonnacott, "Nicotinic acetylcholine receptors and the regulation of neuronal signalling," *Trends in Pharmacological Sciences*, vol. 25, no. 6, pp. 317–324, 2004.

[63] T. E. Robinson and B. Kolb, "Structural plasticity associated with exposure to drugs of abuse," *Neuropharmacology*, vol. 47, no. 1, pp. 33–46, 2004.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 716945, 12 pages
http://dx.doi.org/10.1155/2015/716945

Research Article

Redesigning Protein Cavities as a Strategy for Increasing Affinity in Protein-Protein Interaction: Interferon-γ Receptor 1 as a Model

Jiří Černý, Lada Biedermannová, Pavel Mikulecký, Jiří Zahradník, Tatsiana Charnavets, Peter Šebo, and Bohdan Schneider

Laboratory of Biomolecular Recognition, Institute of Biotechnology, Academy of Sciences of the Czech Republic, Vídeňská 1083, 142 20 Prague, Czech Republic

Correspondence should be addressed to Bohdan Schneider; bohdan.schneider@gmail.com

Received 2 October 2014; Revised 22 December 2014; Accepted 28 December 2014

Academic Editor: Yuedong Yang

Combining computational and experimental tools, we present a new strategy for designing high affinity variants of a binding protein. The affinity is increased by mutating residues not at the interface, but at positions lining internal cavities of one of the interacting molecules. Filling the cavities lowers flexibility of the binding protein, possibly reducing entropic penalty of binding. The approach was tested using the interferon-γ receptor 1 (IFNγR1) complex with IFNγ as a model. Mutations were selected from 52 amino acid positions lining the IFNγR1 internal cavities by using a protocol based on FoldX prediction of free energy changes. The final four mutations filling the IFNγR1 cavities and potentially improving the affinity to IFNγ were expressed, purified, and refolded, and their affinity towards IFNγ was measured by SPR. While individual cavity mutations yielded receptor constructs exhibiting only slight increase of affinity compared to WT, combinations of these mutations with previously characterized variant N96W led to a significant sevenfold increase. The affinity increase in the high affinity receptor variant N96W+V35L is linked to the restriction of its molecular fluctuations in the unbound state. The results demonstrate that mutating cavity residues is a viable strategy for designing protein variants with increased affinity.

1. Introduction

In studying specificity and affinity of protein-protein interactions, the main focus is traditionally on the structural properties of the interface, for example, complementarity of the residue composition, hydrogen-bonding networks, and the role of hydration [1]. However, there is also a significant contribution of the conformational dynamics to the binding affinity. Analysis of molecular dynamics simulations of 17 protein-protein complexes and their unbound components with quasi-harmonic analysis [2] concluded that the protein flexibility has an important influence on the thermodynamics of binding. Moreover, changes in the protein conformational dynamics may lead to substantial changes in affinity to binding partners without an apparent structural change of the complex. For example, reorganization of the hydrogen bonding networks and solvent bridges of the interacting molecules upon mutation, which was accompanied only by subtle structural changes, leads to radically different binding free energy [3, 4]. A recent work [5] shows that the apparent change in the amino acid dynamics determined by NMR spectroscopy is linearly related to the change in the overall binding entropy and also that changes in side-chain dynamics determined from NMR data can be used as a quantitative estimate of changes in conformational entropy [6, 7]. Also, an analysis of crystallographic B-factors has revealed a significant decrease of flexibility of residues exposed to solvent compared to flexibility of residues interacting with another biomolecule and further compared to their flexibility in the protein core [8]. This "freezing" of atoms upon complexation and in the protein core is only slightly larger for the side chain atoms than for the main chain atoms. Entropic cost specific for side-chain freezing has been computationally evaluated as a small, but important contribution to the thermodynamics

of binding [9, 10]. These results indicate that changes in amino acid conformational entropy upon binding contribute significantly to the free energy of protein-protein association.

However important the interaction interface is for the affinity, the interaction is influenced by the whole composition of the cognate molecules, so that modulation of affinity can be achieved by changing other residues than residues at the interface. One such possible alternative approach would be filling cavities in one of the binding partners, thus influencing the stability and dynamics of the interacting proteins [11–14]. Thermodynamic consequences of introducing cavity-filling mutations have been discussed for residues at the interaction interface [15–17] showing that filling the interfacial cavity increases affinity due to both gain in binding enthalpy and a loss in binding entropy, the latter being attributed to a loss of conformational degrees of freedom. It has been shown that interactions between the internal "core" residues is responsible for the folding and thermal stability of a protein [18]. Here, we decided to test whether the protein-protein affinity could be increased by mutations not on the interface, but in cavities inside one of the cognate protein molecules.

This study follows our previous article [21] in which we designed mutations increasing the affinity of human interferon-γ receptor 1 (IFNγR1) towards its natural cognate molecule interferon-γ (IFNγ), an important protein of innate immunity [22, 23]. Here, we retain this model system and the main contours of the protocol but replace the search for interface mutations by searching for mutations in the receptor cavities in order to further increase its interaction affinity to IFNγ and our computer analysis revealed four such cavity mutants. Combining one of these cavity mutations with the best variant designed in our previous study led to a sevenfold increase in affinity compared to the wild-type receptor. We show that the affinity increase in this mutant is related to the restricted flexibility of amino acids in the unbound state of IFNγR1.

2. Materials and Methods

2.1. Outline of the Protocol. Our computational predictions are based on the analysis of crystal structures of complexes between IFNγ and the extracellular part of IFNγR1, namely, the structures of PDB codes 1fg9 [19] and 1fyh [20] that contain four crystallographically independent IFNγ/IFNγR1 complexes. Throughout the paper, IFNγR1 residues are numbered as in UniProt entry P15260. We used the empirical force field implemented in the software FoldX [24] to search for mutations within the positions lining the internal cavities of IFNγR1 molecule that would increase its stability and/or its affinity to IFNγ. All designed mutants of IFNγR1 were subsequently expressed and purified and their affinity to a "single-chain" form of IFNγ (IFNγSC, [25]) was measured. Individual steps of the computational protocol as well as experimental procedures are described below.

2.2. In Silico Design of Variants. The program 3V [26] was used to identify internal cavities in all four available structures of IFNγR1 molecules complexed with IFNγ. In total, 52 cavity-lining residues, which were identified as encapsulating the cavities in at least one of the four structures, were extracted using the VMD program [27]. Each of 52 amino acid residues identified as lining the internal receptor cavities was mutated in all four crystal IFNγ/IFNγR1 complexes to 20 amino acid residues using the "`positionscan`" and "`analyzecomplex`" FoldX keywords. This represented $52 \times 4 \times 20$ mutations (including self-mutations leading to $\Delta\Delta G = 0$). Three types of changes of free energy ($\Delta\Delta G$) were calculated using the program FoldX:

(1) "$\Delta\Delta G$ of folding of IFNγR1 in complex" gauged the influence of mutations on the stability of the whole IFNγ/IFNγR1 complex;

(2) "$\Delta\Delta G$ of folding of free IFNγR1" estimated the effect of mutations on the stability of the isolated receptor;

(3) "$\Delta\Delta G$ of binding" of complex between IFNγR1 and IFNγ estimated the change of the interaction between the receptor molecule and the rest of the complex.

2.3. Modeling. IFNγR1 models are based on PDB structures 1fg9 [19] and 1fyh [20]. Missing residues in both structures were added using Modeller suite of programs [28]. The lowest energy loop models were used for further calculations.

2.4. Molecular Dynamics (MD) Simulations. MD simulations were run using GROMACS suite of programs to test the stability and dynamic properties, including analysis of values of root means square fluctuations (RMSF) [29] and the effect of variable geometry on prediction of changes of interaction free energy ($\Delta\Delta$Gs), of the IFNγ/IFNγR1 complexes (PDB codes 1fyh and 1fg9). More detailed protocol of MD and FoldX calculations follows.

2.5. Protocol of Molecular Dynamics (MD) Calculations. For the MD simulations the following setup was used: protonation state was determined by `pdb2gmx` program using parameters provided by the OpenMM [30] Zephyr [31] program. Implicit solvation (GBSA, $\varepsilon = 78.3$, with collision interval of 10.99 fs) was used in combination with parm96 force field [32]. OpenMM Zephyr implementation of GPU accelerated version of GROMACS [29] suite of programs was used to simulate the systems. The initial crystal structures were optimized and the simulation was propagated at 300 K with the time step of 2 fs. RMSF (root-mean square fluctuations) of atoms in the analyzed proteins were calculated from the 100 ns trajectory to estimate flexibility of residues; they were calculated by `g_rmsf` program in 5 ns windows.

2.6. Construction, Expression, and Purification of Recombinant IFNγR1 Variants. We followed the protocols from our previous study [21] for all proteins produced in this study. All selected IFNγR1 variants were prepared, expressed, and successfully purified to homogeneity by the following protocol.

Codon-optimized synthetic gene (GenScript) encoding extracellular domain of human IFNgR1 (residues 18–245) was cloned into the pET-28b(+) vector (Novagen) using

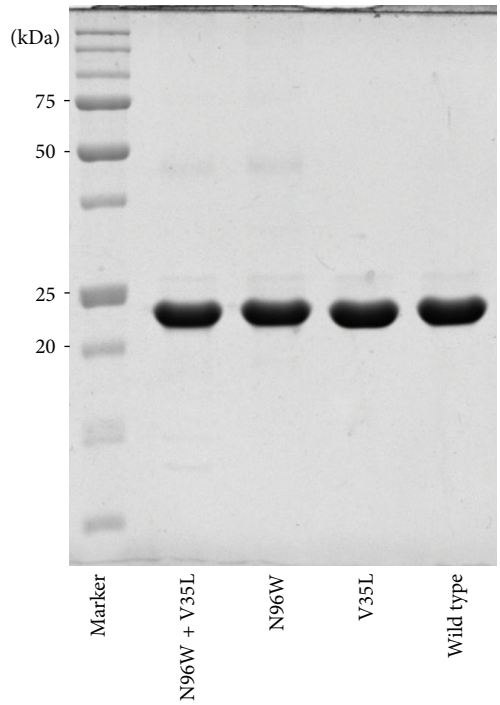


Figure 1: Nonreducing 12.5% SDS-PAGE gel of selected monomeric refolded recombinant His-tagged IFNγR1 variants. Proteins were extracted from inclusion bodies by 8 M urea, further purified on Ni-NTA agarose, and dialyzed, and monomeric fraction was separated on gel filtration column (see above). IFNγR1 with C-terminal His-Tag migrates at a molecular mass of 23 kDa when analyzed on nonreducing SDS-PAGE gel.

NcoI and *XhoI* restriction enzymes in frame with N-terminal start codon and C-terminal HisTag. The QuikChange II Site-Directed Mutagenesis Kit (Agilent Technologies) was used for mutating the IFNγR1 gene according to manufacturer's manual using primers listed below. Primers were designed by web-based PrimerX program (http://www.bioinformatics.org/primerx/).

The recombinant IFNγR1 variants were expressed in *Escherichia coli* BL21 (λDE3) in LB medium containing 60 μg/mL of kanamycin at 37°C for 4 hours after induction by 1 mM IPTG. Harvested cells by centrifugation (8,000 g, 10 min, 4°C) were disrupted by ultrasound in 50 mM Tris buffer pH 8 and centrifuged at 40,000 g, 30 min, 4°C, and inclusion bodies were dissolved in 50 mM Tris buffer pH 8 containing 8 M urea and 300 mM NaCl to extract protein that was further affinity-purified on Ni-NTA agarose (Qiagen) in the same buffer. Protein was eluted from resin by 250 mM Imidazole pH 8 in previous buffer and refolded by dialysis against 100 mM Tris-HCl pH 8, 150 mM NaCl, 2.5 mM EDTA, 0.5 mM Cystamine, and 2.5 mM Cysteamine overnight at 4°C. Final purification of monomeric receptor variants was performed at 4°C on a HiLoad 16/600 Superdex 200 pg (GE Healthcare) equilibrated by PBS buffer pH 7.4 (Figure 1). Monodispersity of the purified receptor protein was verified by dynamic light scattering (DLS) using Malvern Zetasizer Nano ZS90 instrument (data not shown).

2.7. Primers. Mutagenesis primers are designed for the introduction of single residue substitution into IFNγR1 WT. Mutated nucleotides are underlined. We have the following:

V35L

Forward: 5′-GTCCCGACCCCGACCAACTTGACGATT-GAAAGTTACAAC-3′

Reverse: 5′-GTTGTAACTTTCAATCGTCAAGTTGGT-CGGGGTCGGGAC-3′

A114E

Forward: 5′-GAAAGAATCAGCGTATGAAAAATCGGA-AGAATTCGCC-3′

Reverse: 5′-GGCGAATTCTTCCGATTTTTCATACGC-TGATTCTTTC-3′

D124N

Forward: 5′-CGCCGTGTGCCGTAATGGCAAAATCG-3′

Reverse: 5′-CGATTTTGCCATTACGGCACACGGCG-3′

H222Y

Forward: 5′-CTGAAGGCGTTCTGTATGTCTGGGGTG-TC-3′

Reverse: 5′-GACACCCCAGACATACAGAACGCCTTC-AG-3′

2.8. Construction, Expression, and Purification of IFNγSC. Recombinant interferon gamma in so-called single chain form (IFNγSC) described by [25] was cloned into pET-26b(+) vector (Novagen) using *NdeI* and *XhoI* restriction enzymes in frame with N-terminal start codon not to have no peptide leader nor tag.

The recombinant IFNγSC was expressed in *E. coli* BL21 (λDE3) in LB medium containing 60 μg/mL of kanamycin at 30°C for 4 hours after induction by 1 mM IPTG. Harvested cells by centrifugation (8,000 g, 10 min, 4°C) were disrupted by ultrasound in 20 mM Na-Phosphate buffer pH 7.3 and centrifuged at 40,000 g, 30 min, 4°C, and soluble fraction was further purified on SP Sepharose HP (GE Healthcare) using linear gradient of NaCl and further purified to homogeneity by gel filtration in same procedure as IFNγR1 receptor (see above).

2.9. Biophysical Characterization of the Studied Proteins. Melting temperatures of the receptor variants were measured using fluorescence-based thermal shift assay and for selected mutants by CD melting experiments. Interactions between IFNγR1 variants and IFNγSC were measured by the technique of surface plasmon resonance (SPR) as discussed in our previous study [21]. Experimental procedures are detailed below.

Table 1: Cavities in the four molecules of the IFNγR1 receptor in crystal structures 1fg9 [19] and 1fyh [20]. The receptor molecules are labeled by chain ID (chains C and D from 1fg9 and chains B and E from 1fyh). Figure 2 shows cavities 1–8 as they project into the chain C of 1fg9.

	Surface [Å^2]*	Number of residues lining the cavity†	Residues selected for mutation	Cavity observed in IFNγR1 chain of	
				1fg9	1fyh
1	134	7	V35, A114	C D	—
2	133	5	—	—	B E
3	470	14	D124	C D	—
4	262	9	H222	C D	B E
5	120	6	—	C D	E
6	165	7	—	C D	E
7	177	7	—	D	B E
8	138	5	—	C	B

*Surface calculated with a probe radius of 0.25 Å for cavities combined from all relevant receptor chains.
†Some residues are shared by neighboring cavities.

2.10. CD Measurements. CD spectra were recorded using "Chirascan-plus" (Applied Photophysics) spectrometer in steps of 1 nm over the wavelength range of 190–260 nm. Samples at a concentration of 0.2 mg/mL were placed into 0.05 cm path-length quartz cell to the thermostated holder and individual spectra were recorded at the temperature of 25°C. The CD signal was expressed as the difference between the molar absorption of the right- and left-handed circularly polarized light and the resulting spectra were buffer subtracted. To analyze the ratio of the secondary structures we used the CDNN program provided with Chirascan CD spectrometer [33]. For CD melting measurements, samples at a concentration of 1.5 mg/mL were placed into 10 mm path-length quartz cell to the thermostated holder and CD signal at 280 nm was recorded at 1°C increment at rate of 1.0°C/min over the temperature range of 25 to 65°C with an averaging time of 10 seconds. CD melting curves were normalized to relative values between 1.0 and 0.0.

2.11. Thermostability of the IFNγR1 Variants by Thermal-Based Shift Assay. Melting temperature (T_m) curves of the WT and selected variants were obtained from fluorescence-based thermal shift assay (TSA) using fluoroprobe. Experiment was performed in "CFX96 Touch Real-Time PCR Detection System" (Bio-Rad) using FRET Scan Mode. The concentration of fluorescent SYPRO Orange dye (Sigma Aldrich) was 8-fold dilution from 5000-fold stock and protein concentration was 2 μL in final volume of 25 μL. As a reference we used only buffer (PBS buffer pH 7.4) without protein. Thermal denaturation of proteins was performed in capped "Low Tube Strips, CLR" (Bio-Rad) and possible air bubbles in samples were removed by centrifugation immediately before the assay. The samples were heated from 20°C to 75°C with stepwise increment of 0.5°C per minute and a 30 s hold step for every point, followed by the fluorescence reading. Data subtraction by reference sample was normalized and used for first derivative calculation to estimate the melting temperature.

2.12. SPR Measurements. His-tagged receptor molecules were diluted to concentration of 10 μg/mL in PBST running buffer (PBS pH 7.4, 0.005% Tween 20) and immobilized on a HTG sensor chip activated with Ni^{2+} cations at a flow rate 30 μL/min for 60 s to gain similar surface protein density. Purified IFNγSC was diluted in running buffer to concentrations ranging from 0.1 to 9 nM and passed over the sensor chip for 90 seconds at a flow rate 100 μL/min (association phase). Dissociation was measured in the running buffer for 10 min at the same flow rate. Correction for nonspecific binding of IFNγSC to the chip surface was done by subtraction of the response measured on uncoated interspots and reference channel coated with His-tagged Fe-regulated protein D (FrpD) from Neisseria meningitides [34]. Data were processed in the ProteOn Manager software (version 3.1.0.6) and the doubly referenced data were fitted to the 1 : 1 "Langmuir with drift" binding model.

3. Results and Discussion

3.1. Internal Cavities Identified in IFNγR1. The cavity analysis revealed generally different number and size of cavities for each IFNγR1 crystal structure; their characteristics are listed in Table 1; their location in a representative receptor molecule (PDB entry 1fg9, chain C [19]) is highlighted in Figures 2(a) and 2(b). All amino acid residues lining cavities in all four IFNγR1 proteins complexed with IFNγ were combined, resulting in 52 residues used in subsequent *in silico* analysis.

3.2. In Silico Design of Variants. All 52 amino acids lining the cavities of the receptor molecule were subject to the mutation analysis by FoldX. The resulting $\Delta\Delta G$ values indicated potential for mutation leading to increasing the receptor affinity to IFNγ. The mutations were ordered by their $\Delta\Delta G$ values and the first 50 best mutations from each crystal structure (200 mutations in total) were further analyzed. Of these 200 mutations, twelve positions were predicted in all four or at least three crystal structures. The twelve promising positions are highlighted in orange and yellow in Figure 2(c). Following the previous study [21], where we observed significant differences between $\Delta\Delta G$ predicted directly from the crystal structures and from structures after molecular dynamics (MD) relaxation, we performed short

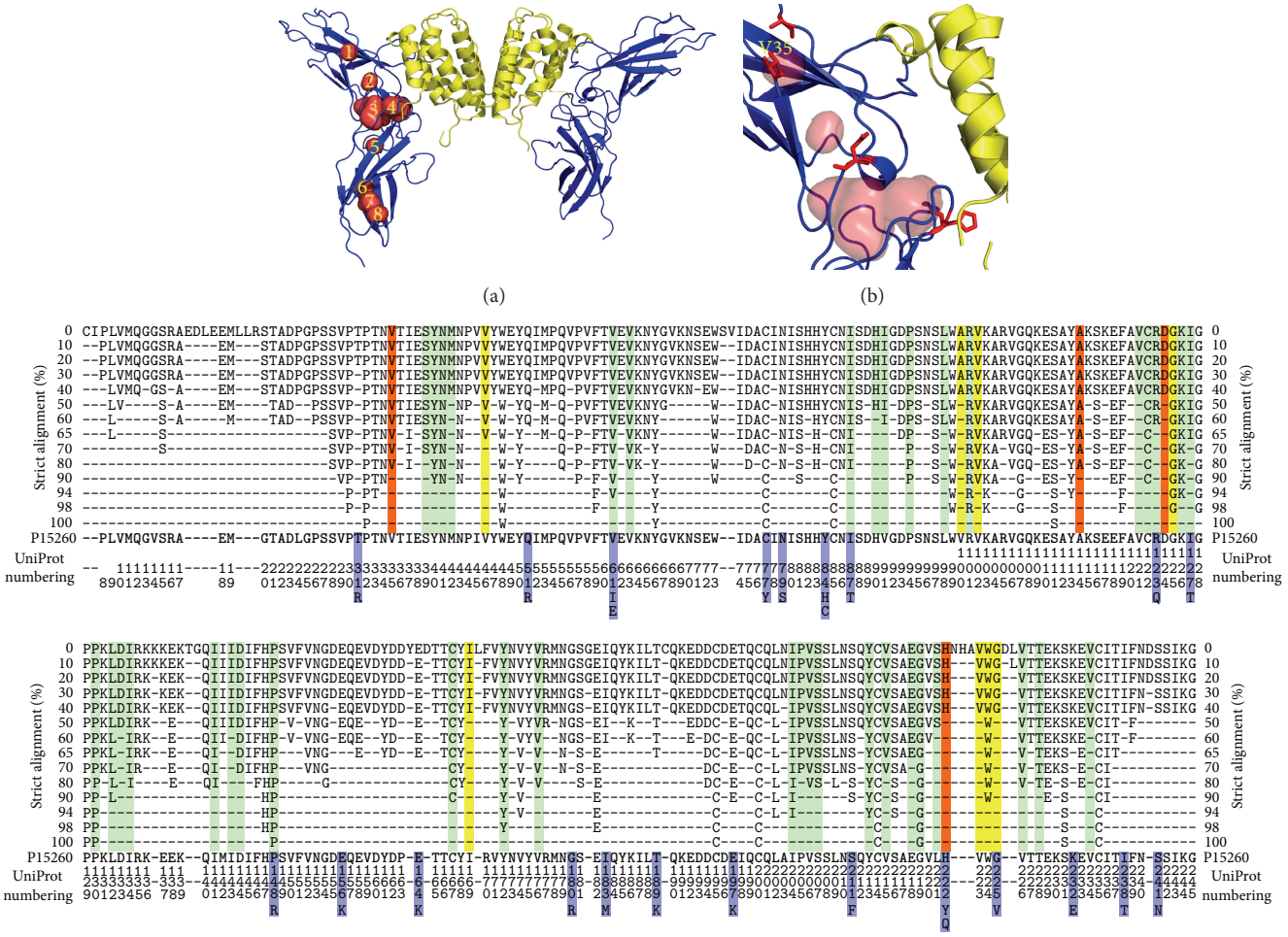


Figure 2: (a) The complex between IFNγ and the extracellular part of its receptor 1 (IFNγR1) from crystal structure of PDB code 1fg9 [19]. The two IFNγR1 molecules are drawn as blue cartoon and IFNγ homodimer as yellow cartoon. The eight identified cavities in the receptor molecule are shown as numbered red surfaces. (b) A close-up of the mutated cavities. The receptor cavities are drawn as red surface and residues selected for mutations as red sticks; valine 35 is labeled. (c) Residue conservancy calculated by strict alignment of 32 sequences of the extracellular part of IFNγR1 from 19 species. The residues lining the cavities and not suitable for mutation are highlighted in green, those selected by FoldX as mutable in yellow, and the residues selected for mutations after MD simulations are in red (they are also listed in Table 1). Blue highlights show IFNγR1 mutants occurring naturally in humans. Percentages of the conservation are shown on the left and right sides; analyzed sequence (residues 6–245 of the UniProt entry P15260) is shown at the bottom of the alignment.

(10 ns) MD simulations of the four crystal structures of complexes between wild type IFNγR1 and IFNγ, and repeated the FoldX mutation analysis on 500 snapshots extracted from these MD trajectories. After averaging of the predicted $\Delta\Delta G$ values for the twelve selected positions, we made the final selection of the four candidate mutations. The averaged $\Delta\Delta G$ values resulting from these calculations for structure 1fg9, receptor chain C, are summarized in Figure 3. The final selection of the four variants is listed in Table 2 together with the changes of their binding free energies averaged over 500 MD snapshots from each of the four IFNγ/IFNγR1 complexes in crystal structures 1fg9 and 1fyh.

Finally, the four consensus candidate mutations, which resulted as the best replacements of the WT sequence, were expressed, and characterized by SPR, CD, and thermal-based shift assay. The relative affinities of these four cavity-filling single mutants are shown in Figure 4(a) together with relative affinities of the double mutants combining the four cavity-filling mutations with mutation N96W.

As Table 2 and in detail Figure 3 show, the $\Delta\Delta G$ calculations revealed only modest potential gains in interaction affinity, probably because of small cavity volumes as well as the fact that they are often lined by evolutionary highly conserved residues. As opposed to the interface mutations, where the predicted $\Delta\Delta G$s of IFNγR1 stability and binding to IFNγ served as a sufficient criterion for the selection of affinity increasing mutations, there was no clear-cut rule for selecting internal cavity mutations that would result in improved interaction energy. We thus decided to test experimental consequences of combination of three types of $\Delta\Delta G$ values

Table 2: Predicted changes of free energy changes ($\Delta\Delta G$) of the four selected IFNγR1 variants with cavity-lining mutations relative to the wild type receptor. All energy values are in kcal/mol.

Variant	$\Delta\Delta G$ of folding of IFNγR1 in complex*	$\Delta\Delta G$ of folding of free IFNγR1†	$\Delta\Delta G$ of binding of IFNγR1/IFNγ complex‡	Sequence conservation¶
V35L	−0.88	−0.85	−0.02	80%
A114E	0.28	0.46	−0.20	60%
D124N	0.65	0.88	−0.21	40%
H222Y	−0.72	−0.69	0.15	40%

*$\Delta\Delta G$ of folding of IFNγR1 bound to IFNγ measures the influence of mutations on the stability of the whole complex.
†$\Delta\Delta G$ of folding of IFNγR1 alone represents changes of the stability of the isolated receptor.
‡$\Delta\Delta G$ of binding of the whole complex between IFNγR1 and IFNγ estimates the change of the affinity between the receptor molecule and the rest of the complex.
¶Sequence conservation of amino acid residues at positions 35, 114, 124, and 222. It was based on the global alignment of 32 sequences of the extracellular part of IFNγR1 (Figure 2(c)).

[1]		GLY	ALA	VAL	LEU	ILE	SER	THR	CYS	MET	ASN	GLN	LYS	ARG	HIS	PRO	ASP	GLU	PHE	TYR	TRP
VAL	35	2.8	2.0	0.0	−0.9	−0.4	2.9	1.6	1.3	0.1	2.0	2.5	3.5	5.1	4.7	1.2	3.4	3.5	4.3	7.3	10.6
VAL	46	3.8	2.2	0.0	−0.1	−0.3	3.1	1.8	1.8	0.6	2.4	3.0	4.1	6.4	6.4	2.0	3.6	3.7	4.0	6.6	9.1
VAL	100	5.6	3.7	0.0	0.3	−0.3	4.2	2.4	2.7	0.8	3.6	3.9	5.4	7.6	6.2	5.0	5.5	5.1	4.0	6.8	9.8
VAL	102	5.2	3.3	0.0	1.2	−0.4	4.0	2.2	2.5	1.8	3.6	4.1	7.1	11.9	9.4	4.8	4.9	4.9	7.5	11.2	15.6
ALA	114	1.0	0.0	−0.2	0.1	0.1	0.3	0.2	0.1	0.3	0.6	0.2	0.2	0.7	3.3	2.3	1.1	0.3	0.7	1.0	1.9
ASP	124	3.0	2.2	2.5	1.4	2.5	2.3	2.7	2.0	1.7	0.7	1.4	1.8	2.1	2.3	5.7	0.0	1.5	1.4	1.6	2.5
GLY	125	0.0	2.0	6.0	6.4	7.7	2.9	5.6	3.0	4.7	5.7	6.8	8.1	10.1	31.3	6.2	7.1	7.1	12.0	14.1	21.8
ILE	169	5.1	3.7	1.1	0.1	0.0	4.7	3.2	2.9	0.3	3.0	3.2	4.1	5.5	3.9	1.8	4.2	3.6	1.9	4.7	7.0
HIS	222	0.7	0.1	0.8	−0.3	1.1	−0.3	0.6	0.4	−0.3	−0.6	0.5	−0.1	0.3	0.0	2.9	−0.1	0.5	−1.1	−0.7	1.1
VAL	223	2.5	2.0	0.0	0.7	0.3	3.7	1.4	2.3	0.9	3.0	3.2	3.8	6.3	14.2	7.3	4.6	4.9	7.6	11.5	15.6
TRP	224	5.5	4.7	3.5	2.8	3.1	5.5	4.9	4.5	2.4	4.8	4.2	4.2	4.0	3.3	4.6	5.9	5.2	1.1	1.5	0.0
GLY	225	0.0	1.5	3.3	2.0	3.4	2.0	3.3	1.8	1.6	2.1	2.5	2.6	2.9	4.7	4.3	2.9	3.0	2.4	2.6	2.9

[2]		GLY	ALA	VAL	LEU	ILE	SER	THR	CYS	MET	ASN	GLN	LYS	ARG	HIS	PRO	ASP	GLU	PHE	TYR	TRP
VAL	35	2.8	2.0	0.0	−0.9	−0.4	2.9	1.6	1.3	0.1	2.0	2.5	3.6	5.3	4.5	1.2	3.4	3.5	4.3	7.3	10.7
VAL	46	5.0	3.0	0.0	−0.2	−0.5	4.1	2.4	2.4	0.5	3.2	3.8	5.1	8.1	7.6	2.9	4.8	4.8	4.4	7.9	11.5
VAL	100	5.7	3.8	0.0	0.3	−0.3	4.2	2.4	2.7	0.8	3.7	4.0	5.5	7.7	5.9	5.0	5.5	5.1	4.1	6.8	9.7
VAL	102	5.2	3.3	0.0	1.2	−0.4	4.0	2.2	2.5	1.8	3.6	4.1	7.1	11.9	9.5	4.8	4.9	4.9	7.5	11.2	15.7
ALA	114	1.0	0.0	−0.2	0.2	0.2	0.3	0.2	0.1	0.4	0.7	0.3	0.3	0.7	3.4	2.3	1.2	0.5	0.8	1.1	2.0
ASP	124	2.4	1.6	2.0	0.7	1.7	1.8	2.1	1.6	0.9	0.9	1.3	1.0	1.4	1.5	4.8	0.0	1.4	1.0	1.2	1.8
GLY	125	0.0	2.0	6.0	6.4	7.8	2.9	5.6	3.0	4.8	5.8	6.8	8.2	10.2	32.2	6.2	7.1	7.2	12.1	14.2	21.9
ILE	169	5.1	3.7	1.1	0.1	0.0	4.7	3.2	2.9	0.3	3.0	3.2	4.2	5.6	3.8	1.8	4.2	3.7	1.9	4.7	7.0
HIS	222	−0.1	−0.6	0.5	−0.4	0.6	−1.0	0.2	−0.1	−0.4	−0.7	0.2	−0.5	0.0	0.0	2.3	−0.3	−0.2	−0.9	−0.7	0.6
VAL	223	0.9	0.6	0.0	0.2	−0.2	1.7	0.4	1.2	0.1	1.2	1.1	0.5	0.9	1.0	5.6	1.6	1.2	0.1	0.3	0.6
TRP	224	2.7	2.1	1.3	1.2	1.1	2.5	2.1	1.9	0.9	2.2	1.9	1.6	2.0	2.0	2.2	2.4	2.0	0.3	0.6	0.0
GLY	225	0.0	1.2	1.9	0.8	1.7	1.2	1.7	1.1	0.6	0.9	0.9	0.8	1.0	1.0	2.8	1.0	1.0	0.4	0.5	0.3

[3]		GLY	ALA	VAL	LEU	ILE	SER	THR	CYS	MET	ASN	GLN	LYS	ARG	HIS	PRO	ASP	GLU	PHE	TYR	TRP
VAL	35	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−0.1	−0.2	0.0	0.0	0.0	0.0	0.0	0.0	−0.1
VAL	46	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
VAL	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
VAL	102	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ALA	114	0.0	0.0	0.0	−0.1	−0.1	0.0	0.0	0.0	−0.1	−0.1	−0.1	−0.1	0.0	−0.1	0.0	−0.1	−0.2	−0.1	−0.1	−0.1
ASP	124	0.2	0.1	0.0	−0.1	0.1	0.1	0.0	−0.1	−0.1	−0.2	−0.2	−0.1	−0.1	0.1	0.2	0.0	−0.1	−0.1	−0.1	0.1
GLY	125	0.0	0.0	0.0	0.0	−0.1	0.0	0.0	0.0	0.0	0.0	0.0	−0.1	−0.1	−0.1	−0.1	0.0	0.0	−0.1	−0.1	−0.1
ILE	169	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HIS	222	1.3	1.2	0.9	0.7	0.8	1.1	0.9	0.7	0.4	0.3	0.5	0.4	0.2	0.0	1.2	0.5	0.6	0.1	0.2	0.8
VAL	223	1.7	1.4	0.0	0.1	0.4	2.0	0.8	1.2	0.3	1.5	1.6	2.0	3.2	6.1	2.2	2.7	2.9	5.7	8.2	10.4
TRP	224	3.3	3.1	2.6	1.8	2.2	3.3	3.0	3.0	1.4	2.9	2.3	2.4	2.0	1.3	2.8	3.6	3.1	0.9	0.8	0.0
GLY	225	0.0	0.3	1.1	0.6	1.2	0.6	1.2	0.4	0.3	0.8	0.9	0.9	0.9	2.9	1.3	1.4	1.3	1.1	1.1	1.4

Figure 3: Color-coded values of free energy changes ($\Delta\Delta G$) of mutating the twelve cavity-lining residues of IFNγR1. $\Delta\Delta G$ values were calculated using the program FoldX for 500 MD snapshots and averaged. Red colored matrix fields indicate stabilization, blue ones destabilization. Shown are $\Delta\Delta G$ values calculated for PDB 1fg9 [19]; receptor chain C. analogical matrices are calculated for 1fg9 receptor chain D, and for receptor chains B and E from the structure 1fyh [20]. (1) "$\Delta\Delta G$ of folding of IFNγR1 in complex" gauged the influence of mutations on the stability of the whole IFNγ/IFNγR1 complex. (2) "$\Delta\Delta G$ of folding of free IFNγR1" estimated the effect of mutations on the stability of the isolated receptor. (3) "$\Delta\Delta G$ of binding" of complex between IFNγR1 and IFNγ made an estimate of change of the interaction between the receptor molecule and the rest of the complex.

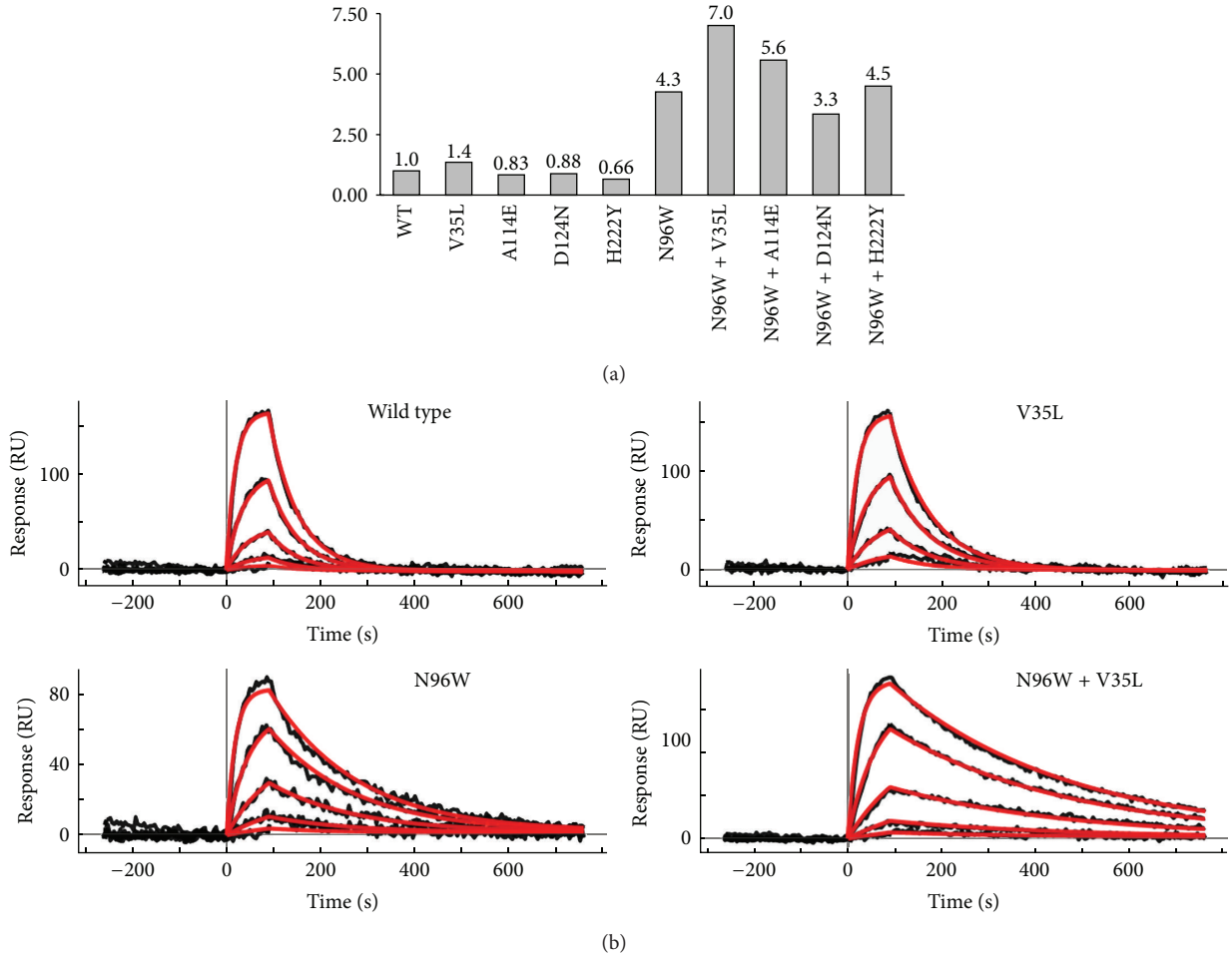


Figure 4: Affinities of the IFNγR1 wild type (WT) and mutants to IFNγSC obtained from SPR measurements. (a) Graph represents relative affinities of IFNγR1 variants compared to WT. All selected "cavity" single amino acid mutation variants bind to the IFNγSC with similar affinity as WT, but the V35L variant has slightly higher affinity itself and further increases the affinity of the "interface" mutant N96W if combined together. (b) SPR sensorgrams showing the interaction between IFNγSC and selected IFNγR1 variants. The V35L variant behaves similarly as WT displaying fast association and dissociation phases. Two variants (N96W and N96W + V35L) with higher affinities compared to WT bind IFNγSC with slower dissociation phase, thus increasing the affinity. Measured SPR signal is in black and calculated fitted curves are in red; concentrations of IFNγSC used for SPR measurements were as follows: 0.1, 0.3, 1.0, 3.0, and 9.0 nM.

calculated from the MD snapshots. To identify potentially favorable mutations, we combined $\Delta\Delta G$ values of folding ($\Delta\Delta G$ types (1) and (2) in the *in silico* protocol described in Materials and Methods) and of binding (type (3)). The first two mutations, V35L and H222Y, were predicted to increase $\Delta\Delta G$ of folding to a similar extent for both the complexed and free IFNγR1 ($\Delta\Delta G$ (1) and (2)), while calculated values of their $\Delta\Delta G$ of binding were virtually zero. The other two selected mutations, A114E and D124N, were predicted to slightly improve $\Delta\Delta G$ of binding while both types of their $\Delta\Delta G$ of folding were destabilizing. In the latter case, $\Delta\Delta G$ of folding of free IFNγR1 (type 2) was more unfavorable than $\Delta\Delta G$ of folding of complexed IFNγR1 (type 1). This means that the complex is predicted to be relatively more stable compared to the free IFNγR1.

3.3. Experimental Determination of the Affinities between IFNγR1 Variants and IFNγSC. Computer-designed IFNγR1 variants were expressed and purified and their affinities to IFNγSC were determined by SPR measurements; relative affinities are plotted in Figure 4(a); SPR sensograms are depicted in Figure 4(b). The calculated K_d values showed that the four selected "cavity" single amino acid mutation variants bind to the IFNγSC with similar affinity as WT; a modest increase was observed for the V35L variant. In line with our previous work, we decided to test to what extent the effect of two distant point mutations is additive. To this end, we combined the four cavity mutants designed here with the variant with the highest affinity designed previously, N96W. The results were quite encouraging: while affinity of one double mutant (N96W + H222Y) is neutral and one (N96W + D124) affinity actually decreased, two double mutants, N96W with A114E and V35L, had affinity increased compared to WT. The affinity increase of one of the double mutants, N96W + V35L, is significant, seven times higher than affinity of WT.

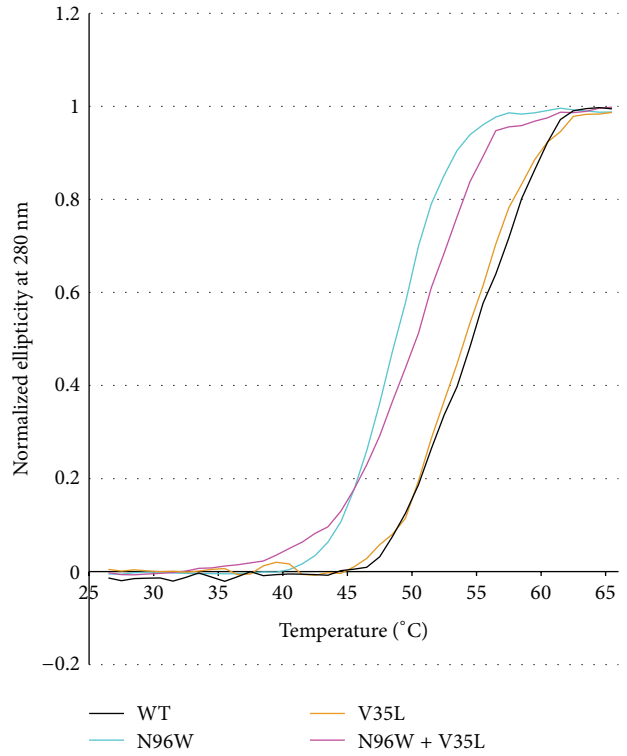


Figure 5: Normalized melting curves of IFNγR1 variants measured by temperature-dependent near ultraviolet circular dichroism (CD) spectra. Each data point is from the intensity measured at 280 nm. IFNγR1 WT, V35L, N96W, and N96W + V35L variants were measured in PBS buffer between 25 and 65°C at steps 1°C/minute. The melting temperature (T_m) of IFNγR1 variants was determined as 54°C for WT, 53°C for V35L, 50°C for N96W + V35L, and 48°C for N96W, respectively.

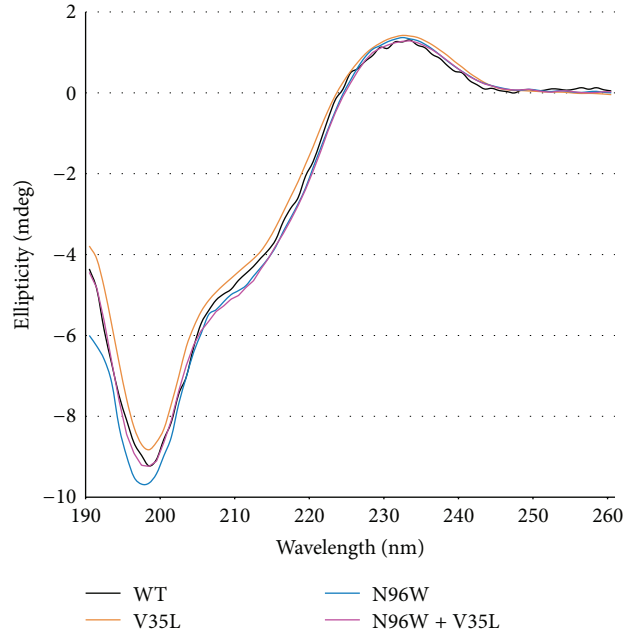


Figure 6: Circular dichroism (CD) spectra of IFNγR1 variants (WT, N96W, V35L, and N96W + V35L) measured in water at 25°C. CD melting curves for the same variants are shown in Figure 5.

The thermal stability (Figure 5) and secondary structure (Figure 6) of four IFNγR1 variants, V35L, N96W, N96W + V35L, and WT, were studied by CD and their melting temperatures were confirmed by thermal-based shift assay (Figure 7); the CD-measured melting temperatures are 53, 48, 50, and 54°C, respectively. Both variants with the highest affinity, N96W and N96W + V35L, have melting temperatures lower than WT, so that mutation from asparagine to tryptophan at the position 96 apparently causes a decrease of IFNγR1 thermal stability. However, the CD spectra of all four proteins are highly similar (Figure 6); their analysis provided virtually identical composition of the secondary structure elements dominated by the beta-sheet fractions indicating that no global structural rearrangements were caused by the mutations and the fold of these four variants is most likely the same. Moreover, the spectra are in agreement with the spectrum measured previously [35] for WT of IFNγR1.

3.4. Analysis of Internal Dynamics of the IFNγR1 Variants. To test how a cavity-filling mutation changes the flexibility of the receptor molecule in unbound and complexed states we analyzed root-mean square fluctuations (RMSF) of the selected variants. Comparison of RMSF sorted by their

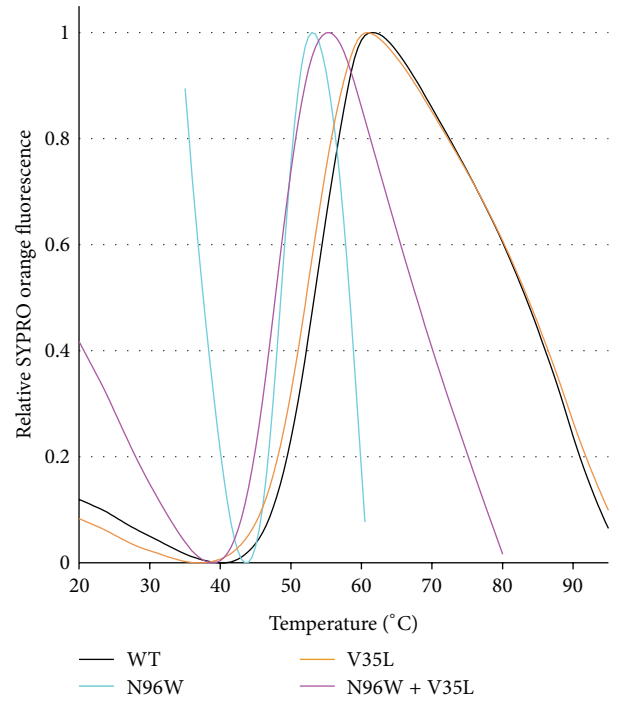
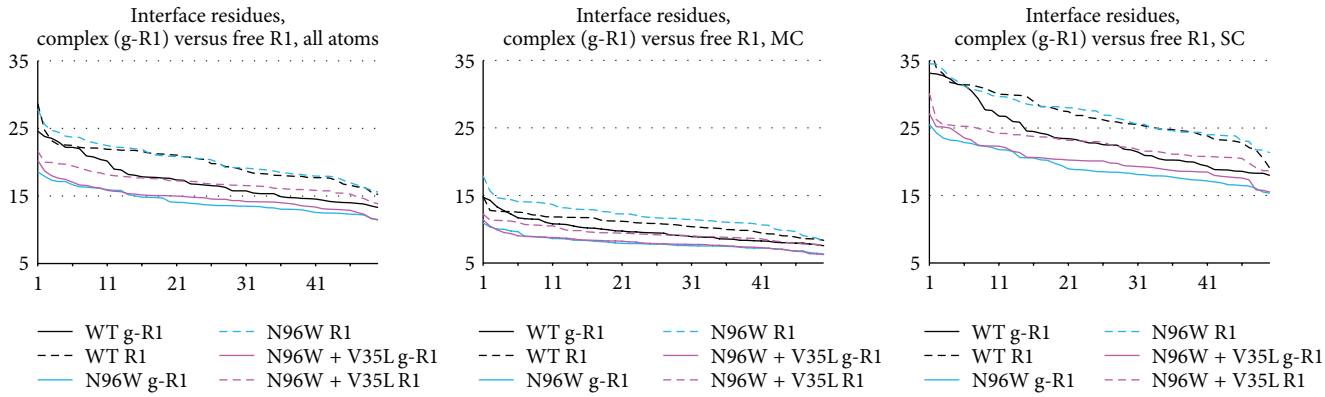
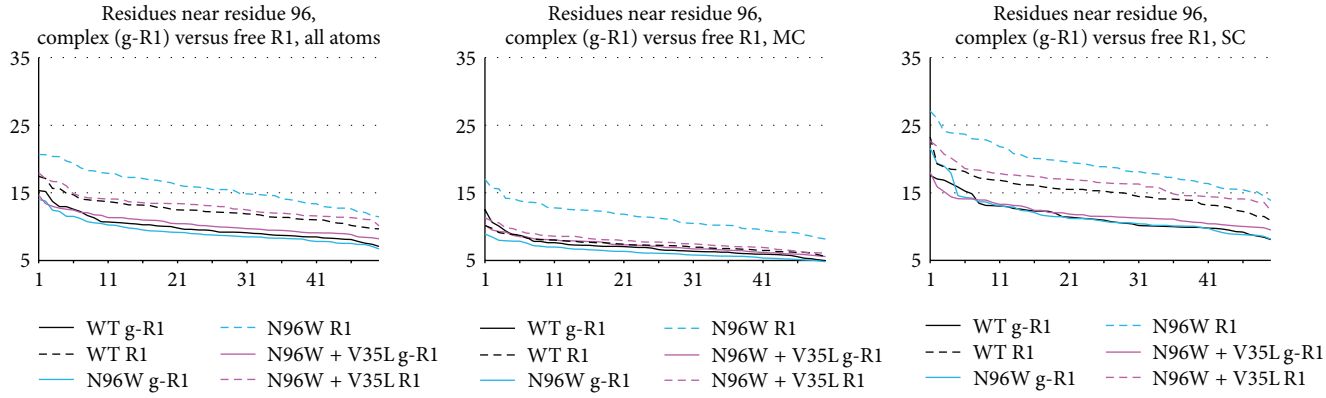


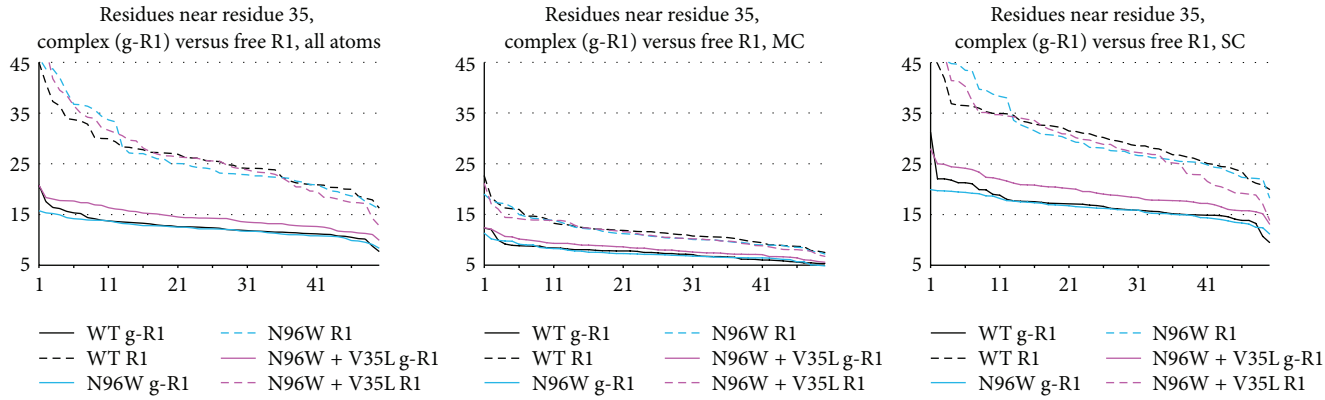
Figure 7: Melting temperatures of selected IFNγR1 variants determined by thermal-based shift assay. Plotted are normalized data of reference-subtracted fluorescence intensities of IFNγR1 WT, V35L, N96W, and N96W + V35L. The melting temperatures (T_m) of IFNγR1 variants were determined from the first derivatives of the curves plotted in the figure: 55°C for WT, 53°C for V35L, 49°C for N96W, and 48°C for N96W + V35L. The T_m values determined by temperature-dependent CD spectra and thermal-based shift assay are within 1°C the same.



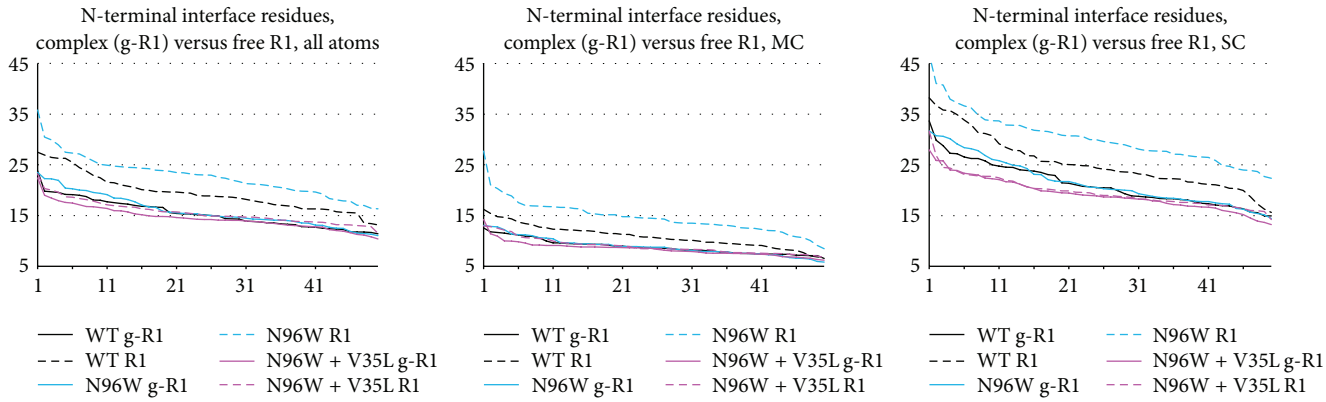
(a)



(b)

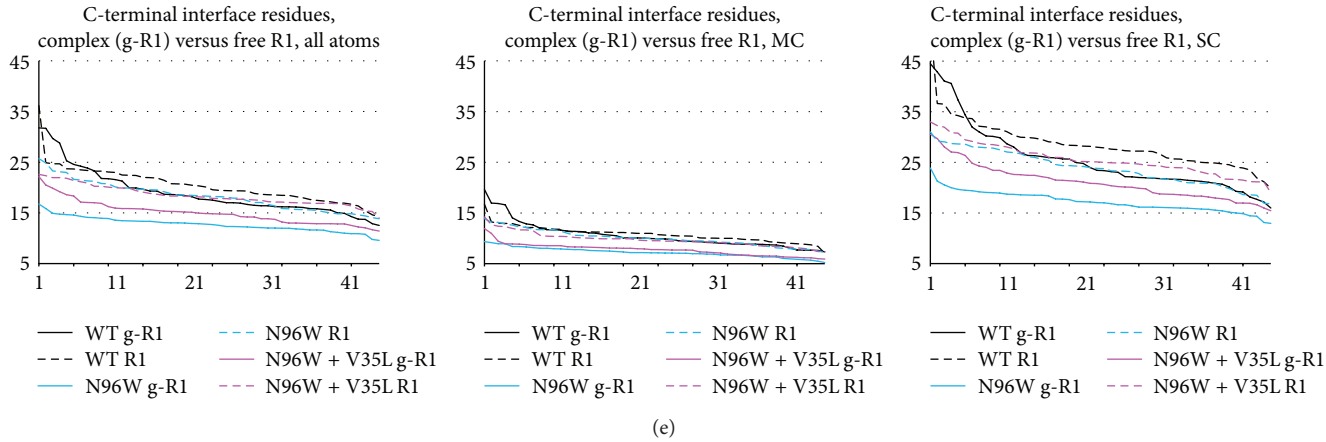


(c)



(d)

Figure 8: Continued.



(e)

Figure 8: Ranked RMSF values collected at the last 50 ns of the 100 ns MD simulations of WT, N96W, and N96W + V35L variants of IFNγR1. Solid lines labeled g-R1 denote RMSF values of the IFNγ/IFNγR1 complex; dashed lines labeled R1 denote values of IFNγR1 alone. The RMSF values are on the y-axis; the rank of the values (1–50) is on the x-axis. Shown are RMSF values of all atoms, main chain atoms (MC), and side chain atoms (SC) for the following residues: (a) all 40 interface residues (i.e., residue numbers 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 93, 95, 96, 97, 99, 115, 116, 118, 123, 164, 165, 166, 168, 170, 171, 186, 189, 190, 191, 192, 193, 197, 220, 221, 222, 223, 224, 225, 226, and 227); (b) residues within 6 Å of residue 96 (i.e., residue numbers 65, 66, 67, 91, 92, 93, 94, 95, 96, 97, 98, 119, 120, 121, and 224); (c) residues within 6 Å of residue 35 (i.e., residue numbers 32, 33, 34, 35, 36, 37, 46, 47, 48, 49, 100, 101, 102, 114, 115, 116, and 117); (d) the interface residues from the N-terminal domain (i.e., residues 64 to 123); (e) the interface residues from the C-terminal domain (i.e., residues 164 to 227).

values, "ranked RMSF," for WT, N96W, and N96W + V35L, are plotted in Figure 8 (solid lines for IFNγ/IFNγR1 complexes, dashed lines for IFNγR1 alone). These plots revealed significant differences between dynamics of the variants as is detailed below.

(1) The interface residues of N96W and WT are more flexible in the free receptor than in the complex, while the flexibility of the interface residues of N96W + V35L is similar for the free and complexed receptor (Figures 8(a) and 8(d)). This indicates entropically more favorable binding of the N96W + V35L variant compared to the other two variants.

(2) Interestingly, the origin of this behavior is different in the N-terminal and C-terminal domains of the IFNγR1 molecule: in the N-terminal domain (Figure 8(d)), the flexibility of the interface residues of all variants is similar in the bound state, while being different in unbound state; they are most flexible in N96W and the least in N96W + V35L. In the C-terminal domain (Figure 8(e)), the flexibility of the three variants is similar in their free states, but it differs in the bound state between N96W, which has the lowest flexibility, and WT with the highest flexibility.

(3) The V35L mutation stiffens the receptor nonlocally and makes especially the C-terminal interface residues more flexible in the bound state compared to the N96W mutant (Figure 8(e)).

(4) To sum up, the V35L mutation brought flexibility of the free and complexed receptor closer together, indicating reduced entropy penalty of binding and resulting in the higher affinity of the N96W + V35L double mutant compared to N96W mutant.

Filling the cavity by hydrophobic groups as in the V35L mutation is stabilizing but not as much as would be implied by $\Delta\Delta G$ of the removal of the corresponding hydrophobic group to water. A compensatory effect lowering a potential increase of the protein and/or complex stability has been observed previously [13] and a comparable decrease of stabilization was also predicted here by FoldX. Filling of a cavity may stabilize the interaction by several mechanisms, for example, by reducing the entropic penalty of complexation by stiffening interacting molecules in the free state, or indirectly by destabilization of the intermediate molten globule state rather than by stabilization of the folded protein [36]. These compensatory effects further illustrate complexity of protein-protein interactions (and/or folding) and the known limits of computational approaches to increasing protein-protein affinity [37].

An important issue potentially affecting reliability of FoldX predictions is the flexibility of the receptor molecule. The first round of FoldX $\Delta\Delta G$ calculations based on the static crystal structures suggested one additional mutation, G225Y, as potentially increasing receptor affinity to IFNγ. Although further calculations using structures of snapshots from the MD simulations did not confirm this prediction, we expressed and characterized this mutation. The experimental data were in agreement with the MD-based prediction showing much lower binding affinity compared to the WT (the ratio of the respective K_d values was 0.4), and also the N96W + G225Y double mutant had a fairly low binding affinity (compared to WT, the ratio of the respective K_d values was 3.1, which is lower than for the N96W mutant). This observation can be explained by the structural properties of the receptor molecule. The loop region of IFNγR1 containing the G225 residue is flexible and any residue at the position 225 is thus only a fraction of time in the geometry, in which it may

increase the binding affinity. An important role of flexibility at the C-terminal part of the interacting IFNγ and IFNγR1 is well illustrated by a study of IFNγ modified at its C-terminal side [38].

3.5. Sequence Conservation of Mutable Residues. We checked sequence conservation for the 12 positions selected by the FoldX calculations for potential cavity-filling mutations. Global alignment of 32 sequences of the extracellular part of IFNγR1 from various organisms by Kalign as implemented in program Ugene [39] (Figure 2(c)) shows conservation between 40 and 98% for these positions; the position V35 is well conserved (80%). The independence of sequence conservation and its potential for stabilizing mutation filling-up protein cavity ("mutability") contrasts with previously observed tight correlation between conservation and mutability for receptor residues interacting with IFNγ [21]: we tested several mutations of the interface residues S97 and E118, which were conserved at the 90% level (Figure 2(c)), namely, S97X (X = L, N, W) and E118X (X = M, F, Y, W), and they did not bind IFNγSC at all (unpublished SPR data) despite the fact that binding of these mutants to IFNγ was predicted to be stronger than that of WT.

3.6. Relationship Between FoldX $\Delta\Delta G$ Values and Naturally Occurring IFNγR1 Variants. Interesting, albeit indirect, validation of the present FoldX predictions of $\Delta\Delta G$ of mutations can be found among naturally occurring IFNγR1 single-point mutations collected in the database of single nucleotide polymorphism (dbSNP) [40]. The database contains 25 nucleotide mutations at 22 unique positions of the extracellular part of the IFNγ receptor, which is studied here; these 22 positions are marked blue in Figure 2(c). Most of the $\Delta\Delta G$ predictions for these natural mutants show neutral effect on the stability of free IFNγR1 and on its complex with IFNγ. This is in agreement with the fact that only two of the natural mutants exhibit deleterious effects or are represented by a pathological phenotype.

4. Conclusions

We present a new computational strategy for designing higher affinity variants of a binding protein and show that it is possible to increase the affinity of a protein-protein interaction by mutations not at the interface, but in the interior cavities of a binding partner. The mutations were selected at positions lining internal cavities of one binding partner, and an *in silico* protocol identified mutations that would fill the protein cavities and increase the stability of the complex. We showed that the selection of such cavity mutations in interferon-γ receptor 1 (IFNγR1) could be performed based on a combination of simple empirical force-field calculations and MD simulations. The mechanism by which the cavity mutations cause affinity increase is shown to be restriction of molecular fluctuations, which can be related to reduced entropy penalty upon binding [6, 7]. IFNγR1 WT and all computationally designed receptor mutants were expressed, purified, and refolded, and the affinity towards the cognate protein, IFNγSC, was measured by SPR. While single mutants showed roughly the same affinity as WT, double mutants combining cavity mutations with the best interface mutation obtained previously [21] were successful in further increasing the binding affinity.

The results demonstrate that mutating cavity residues is a viable strategy for designing protein variants with increased binding affinity. The comparison of computational data and experiments helped to further improve our understanding of forces governing protein-protein interactions. The newly obtained high-affinity binders of IFNγ could be developed into a new diagnostic tool. The significance of the present work can be seen in the fact that small $\Delta\Delta G$ gains of cavity mutants led to significant increase of affinity when combined with more conventional mutations influencing the interface.

Conflict of Interests

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

Acknowledgments

Support from Grant P305/10/2184 from the Czech Science Foundation is greatly acknowledged. This study was supported by BIOCEV CZ.1.05/1.1.00/02.0109 from the ERDF, Biotechnological expert CZ.1.07/2.3.00/30.0020, and by institutional Grant RVO 86 652 036.

References

[1] P. L. Kastritis and A. M. J. J. Bonvin, "Molecular origins of binding affinity: seeking the Archimedean point," *Current Opinion in Structural Biology*, vol. 23, no. 6, pp. 868–877, 2013.

[2] R. Grünberg, M. Nilges, and J. Leckner, "Flexibility and conformational entropy in protein-protein binding," *Structure*, vol. 14, no. 4, pp. 683–693, 2006.

[3] T. N. Bhat, G. A. Bentley, G. Boulot et al., "Bound water molecules and conformational stabilization help mediate an antigen-antibody association," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 3, pp. 1089–1093, 1994.

[4] Y. Urakubo, T. Ikura, and N. Ito, "Crystal structural analysis of protein-protein interactions drastically destabilized by a single mutation," *Protein Science*, vol. 17, no. 6, pp. 1055–1065, 2008.

[5] K. K. Frederick, M. S. Marlow, K. G. Valentine, and A. J. Wand, "Conformational entropy in molecular recognition by proteins," *Nature*, vol. 448, no. 7151, pp. 325–329, 2007.

[6] M. S. Marlow, J. Dogan, K. K. Frederick, K. G. Valentine, and A. J. Wand, "The role of conformational entropy in molecular recognition by calmodulin," *Nature Chemical Biology*, vol. 6, no. 5, pp. 352–358, 2010.

[7] A. J. Wand, "The dark energy of proteins comes to light: conformational entropy and its role in protein function revealed by NMR relaxation," *Current Opinion in Structural Biology*, vol. 23, no. 1, pp. 75–81, 2013.

[8] B. Schneider, J. C. Gelly, A. G. de Brevern, and J. Cerny, "Local dynamics of proteins and DNA evaluated from crystallographic B factors," *Acta Crystallographica D: Biological Crystallography*, vol. 70, part 9, pp. 2413–2419, 2014.

[9] C. Wang, O. Schueler-Furman, and D. Baker, "Improved side-chain modeling for protein-protein docking," *Protein Science*, vol. 14, no. 5, pp. 1328–1339, 2005.

[10] C. Cole and J. Warwicker, "Side-chain conformational entropy at protein-protein interfaces," *Protein Science*, vol. 11, no. 12, pp. 2860–2870, 2002.

[11] M. Bueno, N. Cremades, J. L. Neira, and J. Sancho, "Filling small, empty protein cavities: structural and energetic consequences," *Journal of Molecular Biology*, vol. 358, no. 3, pp. 701–712, 2006.

[12] T. Ohmura, T. Ueda, K. Ootsuka, M. Saito, and T. Imoto, "Stabilization of hen egg white lysozyme by a cavity-filling mutation," *Protein Science*, vol. 10, no. 2, pp. 313–320, 2001.

[13] M. Tanaka, H. Chon, C. Angkawidjaja, Y. Koga, K. Takano, and S. Kanaya, "Protein core adaptability: crystal structures of the cavity-filling variants of *Escherichia coli* rnase HI," *Protein and Peptide Letters*, vol. 17, no. 9, pp. 1163–1169, 2010.

[14] T. Koudelakova, R. Chaloupkova, J. Brezovsky et al., "Engineering enzyme stability and resistance to an organic cosolvent by modification of residues in the access tunnel," *Angewandte Chemie—International Edition*, vol. 52, no. 7, pp. 1959–1963, 2013.

[15] S. Atwell, M. Ultsch, A. M. de Vos, and J. A. Wells, "Structural plasticity in a remodeled protein-protein interface," *Science*, vol. 278, no. 5340, pp. 1125–1128, 1997.

[16] Y. Kawasaki, E. E. Chufan, V. Lafont et al., "How much binding affinity can be gained by filling a cavity?" *Chemical Biology and Drug Design*, vol. 75, no. 2, pp. 143–151, 2010.

[17] L. Morellato-Castillo, P. Acharya, O. Combes et al., "Interfacial cavity filling to optimize CD4-mimetic miniprotein interactions with HIV-1 surface glycoprotein," *Journal of Medicinal Chemistry*, vol. 56, no. 12, pp. 5033–5047, 2013.

[18] J. Černý, J. Vondrášek, and P. Hobza, "Loss of dispersion energy changes the stability and folding/unfolding equilibrium of the trp-cage protein," *The Journal of Physical Chemistry B*, vol. 113, no. 16, pp. 5657–5660, 2009.

[19] D. J. Thiel, M.-H. Le Du, R. L. Walter et al., "Observation of an unexpected third receptor-molecule in the crystal structure of human interferon-γ receptor complex," *Structure*, vol. 8, no. 9, pp. 927–936, 2000.

[20] M. Randal and A. A. Kossiakoff, "Crystallization and preliminary X-ray analysis of a 1:1 complex between a designed monomeric interferon-gamma and its soluble receptor," *Protein Science*, vol. 7, no. 4, pp. 1057–1060, 1998.

[21] P. Mikulecký, J. Černý, L. Biedermannová et al., "Increasing affinity of interferon-γ receptor 1 to interferon-γ by computer-aided design," *BioMed Research International*, vol. 2013, Article ID 752514, 12 pages, 2013.

[22] K. Schroder, P. J. Hertzog, T. Ravasi, and D. A. Hume, "Interferon-gamma: an overview of signals, mechanisms and functions," *Journal of Leukocyte Biology*, vol. 75, no. 2, pp. 163–189, 2004.

[23] E. C. Borden, G. C. Sen, G. Uze et al., "Interferons at age 50: past, current and future impact on biomedicine," *Nature Reviews Drug Discovery*, vol. 6, no. 12, pp. 975–990, 2007.

[24] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, "The FoldX web server: an online force field," *Nucleic Acids Research*, vol. 33, no. 2, pp. W382–W388, 2005.

[25] A. Landar, B. Curry, M. H. Parker et al., "Design, characterization, and structure of a biologically active single-chain mutant of human IFN-γ," *Journal of Molecular Biology*, vol. 299, no. 1, pp. 169–179, 2000.

[26] N. R. Voss and M. Gerstein, "3V: cavity, channel and cleft volume calculator and extractor," *Nucleic Acids Research*, vol. 38, no. 2, pp. W555–W562, 2010.

[27] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[28] B. Webb and A. Sali, "Protein structure modeling with MODELLER," *Methods in Molecular Biology*, vol. 1137, pp. 1–15, 2014.

[29] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 3, pp. 435–447, 2008.

[30] P. Eastman and V. S. Pande, "OpenMM: a hardware-independent framework for molecular simulations," *Computing in Science & Engineering*, vol. 12, no. 4, pp. 34–39, 2010.

[31] M. S. Friedrichs, P. Eastman, V. Vaidyanathan et al., "Accelerating molecular dynamic simulation on graphics processing units," *Journal of Computational Chemistry*, vol. 30, no. 6, pp. 864–872, 2009.

[32] P. A. Kollman, "Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules," *Accounts of Chemical Research*, vol. 29, no. 10, pp. 461–469, 1996.

[33] G. Bohm, R. Muhr, and R. Jaenicke, "Quantitative analysis of protein far UV circular dichroism spectra by neural networks," *Protein Engineering*, vol. 5, no. 3, pp. 191–195, 1992.

[34] E. Sviridova, L. Bumba, P. Rezacova et al., "Crystallization and preliminary crystallographic characterization of the iron-regulated outer membrane lipoprotein FrpD from *Neisseria meningitidis*," *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, vol. 66, part 9, pp. 1119–1123, 2010.

[35] M. Fountoulakis and R. Gentz, "Effect of glycosylation on properties of soluble interferon gamma receptors produced in prokaryotic and eukaryotic expression systems," *Nature Biotechnology*, vol. 10, no. 10, pp. 1143–1147, 1992.

[36] T. Sengupta, Y. Tsutsui, and P. L. Wintrode, "Local and global effects of a cavity filling mutation in a metastable serpin," *Biochemistry*, vol. 48, no. 34, pp. 8233–8240, 2009.

[37] T. S. Chen and A. E. Keating, "Designing specific protein-protein interactions using computation, experimental library screening, or integrated methods," *Protein Science*, vol. 21, no. 7, pp. 949–963, 2012.

[38] E. Saesen, S. Sarrazin, C. Laguri et al., "Insights into the mechanism by which interferon-γ basic amino acid clusters mediate protein binding to heparan sulfate," *Journal of the American Chemical Society*, vol. 135, no. 25, pp. 9384–9390, 2013.

[39] K. Okonechnikov, O. Golosova, M. Fursov, and UGENE Team, "Unipro ugene: a unified bioinformatics toolkit," *Bioinformatics*, vol. 28, no. 8, pp. 1166–1167, 2012.

[40] S. T. Sherry, M.-H. Ward, M. Kholodov et al., "DbSNP: the NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, 2001.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 253854, 11 pages
http://dx.doi.org/10.1155/2015/253854

Research Article

An Improved Method for Completely Uncertain Biological Network Alignment

Bin Shen,[1] Muwei Zhao,[1] Wei Zhong,[2] and Jieyue He[1]

[1]*School of Computer Science and Engineering, MOE Key Laboratory of Computer Network and Information Integration, Southeast University, Nanjing 210096, China*
[2]*Division of Math and Computer Science, University of South Carolina Upstate, Spartanburg, SC 29303, USA*

Correspondence should be addressed to Jieyue He; jieyuehe@seu.edu.cn

Received 30 September 2014; Revised 27 December 2014; Accepted 2 January 2015

Academic Editor: Yuedong Yang

With the continuous development of biological experiment technology, more and more data related to uncertain biological networks needs to be analyzed. However, most of current alignment methods are designed for the deterministic biological network. Only a few can solve the probabilistic network alignment problem. However, these approaches only use the part of probabilistic data in the original networks allowing only one of the two networks to be probabilistic. To overcome the weakness of current approaches, an improved method called completely probabilistic biological network comparison alignment (C_PBNA) is proposed in this paper. This new method is designed for complete probabilistic biological network alignment based on probabilistic biological network alignment (PBNA) in order to take full advantage of the uncertain information of biological network. The degree of consistency (agreement) indicates that C_PBNA can find the results neglected by PBNA algorithm. Furthermore, the GO consistency (GOC) and global network alignment score (GNAS) have been selected as evaluation criteria, and all of them proved that C_PBNA can obtain more biologically significant results than those of PBNA algorithm.

1. Introduction

The development of biological experiment technology has generated more and more biological network data such as protein-protein interaction and gene transcriptional regulatory network data, which brings considerable number of pieces of information about the interactions and relationships between biological organisms. For this reason scientists carried out a lot of research in this area. Comparative analysis, namely, biological network alignment, is an important method in biological network research. Biological networks alignment can simply be described as the analysis of biological networks by comparing the data to find the correlation between structure and function of organisms and thus to help the study of biological development and evolution. This study demonstrates great potentials to discover basic functions and to reveal essential mechanisms for various biological phenomena, by understanding biological systems not at individual component level but at a system-wide level [1, 2].

Ogata et al. first proposed the graph comparison approach to identify local similarities between two graphs, which allows gaps and mismatch of nodes and edges and is especially suitable for detecting biological features in 2000 [3]. They used the above-mentioned comparative method to discover the relationship between enzymes and positions of their corresponding gene encodings in the entire genome. After analyzing these results, they found the local structure similarities corresponding to functionally related enzyme clusters. Thereafter, the graph comparison research attracted many scholars' research interests in this field. Kelley et al. in 2003 introduced the value of the concept called BLASTE into protein interaction network and thereby described a new way to detect the highly conserved pathway and the highly conserved functional module in the two networks [4]. Subsequently, Koyutürk et al. took advantage of the duplication/divergence model to translate protein-protein interaction (PPI) network comparison into the maximum weight subgraphs problems and used the greedy method to solve the problem [5]. In 2007, Singh et al. proposed

the IsoRank algorithm, which converted the problem to a constraint-based optimization objective function problem. Then they also introduced an algorithm called IsoRankN, which was an approach similar to the PageRank-Nibble algorithm, to align multiple PPI network [6]. The Match-and-Split algorithm proposed by Narayanan and Karp, 2007, with the idea of dividing and conquering strategy divided biological networks into submodules. This approach deals with biological networks alignment by comparing smaller modules [7]. In 2009, Klau [8] normalized the problem to an optimization problem and solved the problem with the integer linear programming method. So far, most of the researches are focused on determining biological network data.

However due to the size, density, and redundancy of interacting molecules in the network and even errors in biological experiments [9], interactions in biological networks are probabilistic events. For example, in a living cell, DNA binding proteins are believed to be in equilibrium between the bound and unbound states, thus introducing uncertainties in protein-DNA interactions. Similar circumstance holds for protein-protein interactions, which are crucial to cellular functions both in assembling protein machinery and in signaling cascades. Therefore we abstract the biological networks into the uncertain networks whose edges are denoted by the values, respectively. Our solution is closer to realistic situation. Incorporation of uncertain information will bring more challenges to the biological networks alignment and analysis.

To the best of our knowledge, there are only two biological network alignment algorithms that can deal with uncertain network. Weighted IsoRankN [10] based on the IsoRank was proposed to deal with the probabilistic case. But the probability information in Weighted IsoRankN was considered as "weight" rather than the true "probability." Essentially, the Weighted IsoRankN algorithm merely simplifies the probability graph into the deterministic diagram. Hence a majority of pieces of information were neglected via this measure. PBNA (probabilistic biological network alignment) [11] proposed by Todor et al. in 2013 was an advanced version of the IsoRank algorithm. The core of this algorithm was to replace the determining variables in the IsoRank with a random variable so as to establish a model for biological network alignment problem. Then this probability algorithm was optimized by using conditional probability distribution knowledge to reduce its complexity. However, the PBNA approach requires at least one of the networks participating in alignment be determined diagram. In other words, if the participating networks are all uncertain networks, one of them will be considered automatically as a deterministic graph. Clearly, the neglect of the probability information of the networks may lead to the deviation.

In order to simplify the discussion in the rest of the cases, "deterministic network alignment" refers to the algorithm in which two participators are all deterministic network and "part probabilistic network alignment" refers to the algorithm in which one of participators is probability graph and "complete probabilistic network alignment" refers to the algorithm in which both participators are uncertain graphs.

In this paper, we develop a method called "complete probabilistic biological network alignment" (C_PBNA) based on "part probabilistic biological network alignment (PBNA)." Our approach can take full advantage of the information for the uncertain network alignment with two uncertain networks. Finally, we conduct 122 groups of contrast tests based on uncertain protein interactions network data preprocessed by Todor et al. from MINT database. We use the agreement to tell the difference between two algorithms. The biological significance of the network comparison is quantified by global network alignment score, gene ontology consistency, and functional coherence of the alignments. The experiment results indicate that C_PBNA can find the results neglected by PBNA algorithm. Furthermore, C_PBNA can obtain more biologically significant results than those of PBNA algorithm.

The rest of the paper is organized as follows. In Section 2, we describe the C_PBNA algorithm. In Section 3, we apply the C_PBNA algorithm into MINT [12] and analyze the results. In Section 4, the conclusions are given.

2. Methods

The C_PBNA proposed in the paper is an advanced biological networks alignment algorithm derived from the PBNA [9]. Both of the algorithms are based on the framework of IsoRank for deterministic network. The PBNA approach takes the uncertainties of the networks into consideration. However, the precondition of PBNA is that at least one of the biological networks participating in alignment must be deterministic network. Our approach can deal with alignment of two uncertain biological networks. Furthermore it can directly deal with the deterministic and partially probabilistic situation. In the following sections, we start by analyzing the probabilistic biological networks which are dealt with by the C_PBNA algorithm then and the results discovered by C_PBNA. Our whole approach is described in five sections: (1) Probabilistic Biological Network; (2) Probabilistic Support Matrix; (3) Probabilistic Model of the Eigenvector; (4) Extracting Alignment Results; (5) C_PBNA Algorithm.

2.1. Probabilistic Biological Network. Traditional deterministic biological network can be characterized by a two-tuple $G = (V, E)$, where V denotes the vertex and E denotes the set of graph edges. Different types of biological networks correspond to different graphs; for instance, PPI network (protein-protein interaction network) can be abstracted into an undirected graph with the vertices tagged by labels denoting different proteins and the edges denoting the interaction relationship of proteins.

It is important to note that biological networks usually are indeterminate; for instance, the interactions of proteins often exist at certain probability. Therefore, we consider the uncertain biological network as a network in which the proteins are denoted by determinate nodes and the interactions of proteins are denoted by edges with a probability value.

Uncertain network is characterized by a three-tuple $g = (V, E, \Pr_E)$, where V, E denote the vertex set and edge set, respectively [13]. Consider

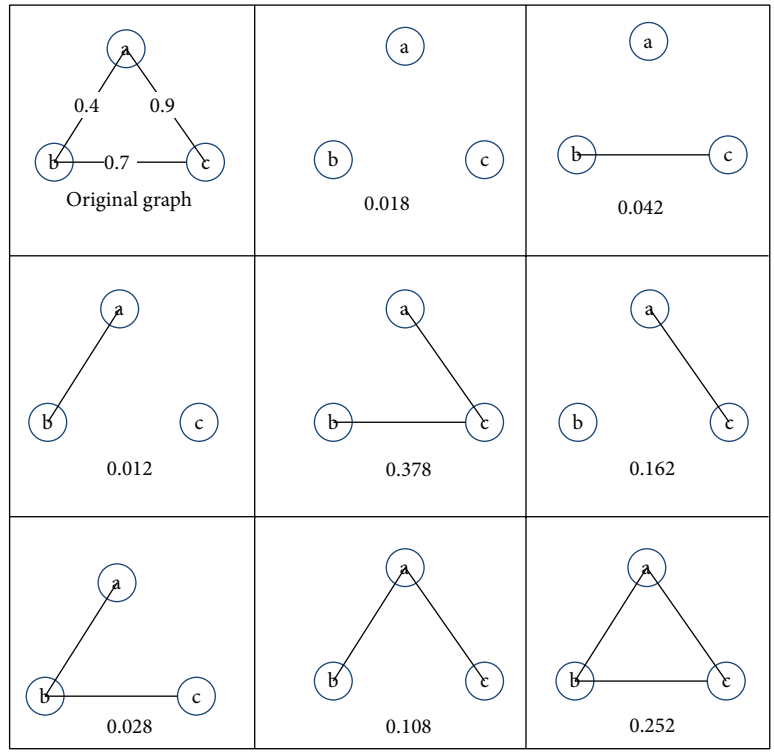


FIGURE 1: Original uncertain graph, as well as its eight-implication graph.

$$\Pr_E : E \rightarrow [0,1]. \tag{1}$$

Note that $\Pr_E$ is a function which denotes a value probability in $[0,1]$ for each edge $e = (u,v)$; specifically $\Pr_e = 1$ indicates that edge $e = (u,v)$ definitely exists:

$$\Pr(g \Longrightarrow G) = \prod_{e \in E'} \Pr_E(e) \prod_{e \in E \cap (V' \times V') \setminus E'} (1 - \Pr_E(e)). \tag{2}$$

$g = (V, E, \Pr_E)$ denotes uncertain graph and $G = (V', E')$ denotes deterministic graph, respectively. Particularly, uncertain graph g contains deterministic G, which is abbreviated as $g \Rightarrow G$, when and only when $V' = V$ and $E' \subseteq E \cap (V' \times V')$, where $E \cap (V' \times V')$ denotes an edge set in which two endpoints of the edge are both in the vertex set V'.

Example 1. Observe that original uncertain graph in Figure 1 has three probability sides. Hence, it contains 8 kinds of different deterministic networks with different probability. This means that in a probabilistic network with $|E|$ edges there are actually $2^{|E|}$ deterministic networks which occur at a certain probability.

Note that, with the uncertainty added, the complexity of the graph increases greatly. For instance, the MINT [12] network data used in the experiments contain 2^{313} implication graphs for the maximum organism containing 313 edges. Precise comparison seems almost impossible for such a large amount of data.

2.2. Probabilistic Support Matrix. Firstly, our approach proposed in the paper is based on the primal framework of IsoRank algorithm which aimed at deterministic graph alignment. One of the core ideas of IsoRank is that the similarity between two vertices may be determined by all the neighborhood vertices' similarity. First of all, we introduce the simple case of pairwise global network alignment (GNA).

For deterministic networks $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, the similarity score R_{ij} between vertexes v_i and v_j can be calculated by

$$R_{ij} = \sum_{u \in N(i)} \sum_{v \in N(j)} \frac{1}{d_u d_v} R_{uv}, \tag{3}$$

where $v_i \in V_1$, $v_j \in V_2$, $N(i)$, $N(j)$, respectively, denote the neighbor vertexes set of v_i and v_j, and d_u, d_v, respectively, denote the degrees of v_u and v_v. We assume that $m = |V_1|$, $n = |V_2|$, all similarity scores R_{ij} $(0 \le i \le m, 0 \le j \le n)$ constituting an $m \times n$ dimensional similarity score vector R. R can be seen as a vector form converted from an $m \times n$ matrix. Therefore (4) can be rewritten in matrix form: $R = AR$, where

$$A[i,j][u,v] = \begin{cases} \dfrac{1}{d_u d_v}, & \text{if } (v_i, v_u) \in E_1, (v_j, v_v) \in E_2 \\ \dfrac{1}{mn}, & \text{if } d_u d_v = 0 \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $A[i,j][u,v]$ is a $(mn) \times (mn)$ matrix with double indexed row and column. And $A[i,j][u,v]$ denotes

Table 1: Degree distribution of nodes.

D	$P(D)$
0	$\prod_{e\in E}(1-\Pr_E(e))$
1	$\sum_{e_i\in E}\Pr_E(e_i)\prod_{e\in E/e_i}(1-\Pr_E(e))$
⋮	⋮
$d^{\max}$	$\prod_{e\in E}\Pr_E(e)$

the element of $[i, j]$ rows and $[u, v]$ column of the matrix A. The term $1/mn$ denotes that point d_u or point d_v is an acnode. As can be seen, formula, $R = AR$, indicates that vector R is a characteristic feature vector of matrix A when the eigenvalue is 1.

The important improvement of PBNA algorithm is to replace the deterministic variable in original algorithm with a random variable so as to simplify the model by calculating the expectation $E(A)$ instead of A itself. It should be stressed that, due to the complexity in calculating desired $E(A)$, PBNA alignment algorithm requires that one of the graphs must be determined. Considering this idea as a reference, we propose C_PBNA algorithm which can be extended to the network alignment problem in which two graphs, G_1 and G_2, are both uncertain graphs. Hence, the degrees of uncertain graph nodes set of v_u and v_v are both uncertain rather than deterministic values. The degree values d_v, d_u are denoted by discrete random variables D_v, D_u respectively, and then (4) can be rewritten as

$$A[i,j][u,v] = \begin{cases} \dfrac{1}{D_u D_v}, & \text{if } (v_i, v_u) \in E_1,\ (v_j, v_v) \in E_2 \\ \dfrac{1}{mn}, & \text{if } D_u D_v = 0 \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where D_v, D_u are discrete distribution: $P(D_v = k)$, $k = 0, 1, \ldots, d^{\max}$, D_v is the degree distribution of node v_v. We assume that E_v is a set of edges connecting to point v_v; hence, $P(D_v)$ can be obtained via probabilistic graphical models shown in Table 1.

Clearly, adding uncertain information increases the complexity of the algorithm. As a result, the time complexity for calculating each node degree distribution sequence increase to exponential time, because the neighbor points degrees in the matrix A as an item subject to the discrete distribution rather than a certain value. Therefore, based on the core idea of literature [9], we use $E(A)$ instead of A involved in the calculation. The following section summarizes the calculation arriving at expectation $E(A)$ of matrix A.

2.3. Probabilistic Model of the Eigenvector. We start by discussing (5) in the first case. Clearly, as discussed earlier, $(v_i, v_u) \in E_1$, $(v_j, v_v) \in E_2$; hence v_u and v_v have at least one connecting edge. So bring $D_u = k_1$ ($k_1 = 1, \ldots, d_u^{\max}$), $D_v = k_2$, ($k_2 = 1, \ldots, d_v^{\max}$) into (5) as follows:

$$\begin{aligned} &E[A[i,j][u,v]] \\ &= \sum_{k_1\in D_u, k_2\in D_v} \frac{1}{k_1 k_2} \\ &\quad \cdot P\left(A[i,j][u,v] = \frac{1}{k_1 k_2} \mid (v_i, v_u) \in E_1, (v_j, v_v) \in E_2\right). \end{aligned} \tag{6}$$

Because of assuming that the edges of the network G_1 and G_2 are independent events, so the D_u and D_v are independent too, we can derive as follows:

$$\begin{aligned} E[A[i,j][u,v]] = &\sum_{k_1\in D_u, k_2\in D_v} \frac{1}{k_1 k_2} \\ &\cdot P(D_u = k_1 \mid (v_i, v_u) \in E_1) \\ &\cdot P(D_v = k_2 \mid (v_j, v_v) \in E_2). \end{aligned} \tag{7}$$

In the next case of (5), the probability is denoted by $P(D_u D_v = 0)$. Note that D_u and D_v are also independent; we can get (8), after some manipulations as follows: $E[A[i, j][u, v]] = (1/mn)P(D_u D_v = 0)$

$$E[A[i,j][u,v]] = \frac{1}{mn}\left(P(D_u = 0) + P(D_v = 0)\right). \tag{8}$$

Similarly, substituting the results of (7) and (8) for P_0 and $P_{k_1k_2}$, respectively, we obtain

$$E(A[i,j][u,v]) = \frac{1}{mn} \times P_0 + \sum_{k_1=1}^{d_u^{\max}} \sum_{k_2=1}^{d_v^{\max}} \frac{1}{k_1 k_2} \cdot P_{k_1 k_2}, \tag{9}$$

where the probability distributions of D_u and D_v are calculated as discussed in Table 1; thus we get the $E(A)$. However, as discussed above in Section 2.2, calculating $P(D_u = k)$ directly means that the computational complexity of the algorithm can reach $O(2^{d_v + d_u})$. In order to reduce the high complexity, we use the probability generating function introduced in the literature.

Definition 2 (see [14]). Assume that X is a discrete random integer variable ranging from 0 to N; therefore the probability generating function (PGF) of X may be defined as a polynomial of z:

$$Q_X(z) = E[z^X] = \sum_{k=0}^{N} P(X = k) z^k. \tag{10}$$

As we see in Definition 2, the coefficient distribution sequence corresponds to discrete random variables distribution of X in probability generating function. Clearly, as long as the probability generating function is calculated, the probability distribution will be obtained easily. Moreover, the probability generating function may be calculated by Theorem 3.

Table 2: Degree distribution of node a.

N_a	0	1	2
$P(N_a)$	0.06	0.58	0.36

Theorem 3 (see [14]). *Suppose that $G = (V, E, \Pr_E)$ is an uncertain graph and E_v denotes a set of edges connecting with v endpoint; hence, the degree of v is a discrete random variable whose probability generating function is shown as*

$$Q_{N_v}(z) = \prod_{e \in E_v} (1 - p_e + p_e z). \tag{11}$$

For example, Figure 1 shows an uncertain network, in which there are two edges connecting with the node and the appearance probability of those two edges is 0.4 and 0.9, respectively. As a result, the probability generating function of degree distribution for node a is $Q_{N_a}(z) = (0.6 + 0.4z)(0.1 + 0.9z) = 0.06 + 0.58z + 0.36z^2$; then we can easily get the distribution of degree for node a as in Table 2.

Therefore, we can calculate the probability generating function of the node degree distribution for v and then obtain node distribution sequence via the probability generating function. The computational complexity of this process can decrease Naive Approach Complexity from $O(2^{d_v^{\max} + d_u^{\max}})$ to $O((d_v^{\max} d_u^{\max})^2)$.

Our ultimate objective is the conditional probability distribution of the degree. In other words, the purpose is to calculate the distribution sequence of node v with the presupposition that there exists an edge connecting to node v and edge $e \in E_v$. As a result, the probability generating function of the conditional probability can be obtained by simply dividing $(1 - p_e + p_e z)$.

Now we can easily calculate the probability generating function of the conditional probability $P(D_v = k \mid e \in E_v)$ $k = 1, 2, \ldots, d_v^{\max}$. And we can get the support matrix $E(A)$ within the time expected in the polynomial equation (9). In particular, the sequence similarity method is also added to the score vector according to the literature [5] as

$$R = \alpha E(A) R + (1 - \alpha) E, \tag{12}$$

where E is the vector denoting normalization of sequence similarity score BLAST and α is a constant used for balancing influence of topological similarity and sequence similarity on calculating the pairwise similarity. Finally, we use a power iteration method [15, 16] to calculate R and record all similarity score. See Algorithm 2.

2.4. Extracting Alignment Results. After calculating similarity vector R, the last step of our model is to extract the final alignment results from vector R. In order to extract the final alignment results, we introduce a breadth first searching approach by using maximum weight bipartite matching technique [17, 18]. First of all, we introduce the concept of bipartite graph and maximum weight bipartite matching.

Bipartite graph: a graph $G_{12} = (V_{12}, E_{12})$ is bipartite if there exists $V_{12} = V_1 \cup V_2$ with $V_1 \cap V_2 = \emptyset$. And for each edge $e \in E_{12}$, the two end vertices must belong to the two different subsets V_1 and V_2.

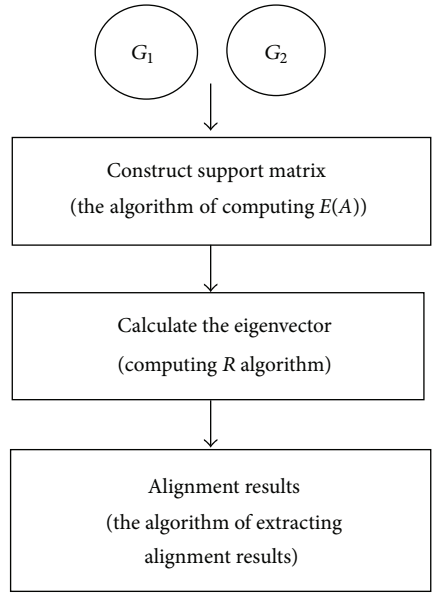


Figure 2: The framework of C_PBNA.

Maximum weight bipartite matching: given a bipartite graph $G_{12} = (V_{12}, E_{12})$ with bipartition (V_1, V_2) and weight function $w : E \to R$ find matching of maximum weight where the weight of matching M is given by $w(M) = \sum_{e \in E_{12}} w(e)$.

The process of extracting results from the feature vector R is the process to find a max-weight matching M in G_{12}.

Let us call a function $y : (V_1 \cup V_2) \to R$ a potential if $y(i) + y(j) \leq w(i, j)$ for each $i \in V_1$, $j \in V_2$. The value of potential is $\sum_{v \in V_1 \cup V_2} y(v)$. It can be seen that the cost of each perfect matching is at least the value of each potential. The Hungarian method finds perfect matching and a potential with equal value which proves the optimality of both. In fact it finds perfect matching of tight edges: an edge e_{ij} is called tight for a potential if $y(i) + y(j) = w(i, j)$. See Algorithm 3.

2.5. C_PBNA Algorithm. The C_PBNA algorithm can be roughly divided into three steps, constructing the support matrix, calculating the eigenvector of the matrix, and extracting alignment results, as in Figure 2. These steps will give detailed descriptions by Algorithms 1, 2, and 3.

First we build probabilistic support matrix based on the conclusions of Section 2.2 and calculate $E(A)$ based on formula (9) in Section 2.3. The pseudocode can be seen in Algorithm 1. Secondly, we calculate the feature vector R by using an iterative approach denoted as in Algorithm 2. Thirdly, we extract optimal comparison by interpreting R as encoding a bipartite graph and finding the maximum weight bipartite matching, which is denoted as in Algorithm 3.

In Algorithm 1 we have the following.

(1) *Line 1–Line 4.* Construct the PGF for every node in probabilistic networks G_1 and G_2.

Input: Probabilistic graph $G_1 = (V_1, E_1)$,
Probabilistic graph $G_2 = (V_2, E_2)$
Output: $E(A)$
// Construct PGF of V_1V_2
(1) for all $u \in V_1, v \in V_2$ do:
(2) construct PGF of u
(3) construct PGF of v
(4) end for
// Compute every entry in $E(A)$
(5) for all $E(A[i, j][u, v]) \in E(A)$ do:
// Compute P_0 using (10)
(6) $P_{01} = 1$
(7) for all $e \in \varepsilon_u$ do:
(8) $P_{01} \times = (1 - p_e)$
(9) end for
(10) $P_{02} = 1$
(11) for all $e \in \varepsilon_v$ do:
(12) $P_{02} \times = (1 - p_e)$
(13) end for
(14) $P_0 = P_{01} + P_{02}$
//Compute $P_{k_1k_2}$ using (12)
(15) $S = 0$
(16) for all $k_1 = 1 \rightarrow d_u^{\max}$, $k_2 = 1 \rightarrow d_v^{\max}$ do:
//Compute PGF of u' and v'
(17) $Q^i_{D_u} = Q_{D_u}/(1 - P(i, u) + P(i, u)z)$
(18) $Q^j_{D_v} = Q_{D_v}/(1 - P(j, v) + P(j, v)z)$
//Use PGF of u' and v'
//to get conditional Probabilistic distribution
//according to Theorem 3
(19) $P_{k_1k_2} = Q^i_{D_u}||_{k_1-1} \times Q^j_{D_v}||_{k_2-1}$
(20) $S+ = P_{k_1k_2}/k_1k_2$
(21) end for
(22) $E[A[i, j][u, v]] = 1/mn \times P_0 + S$
(23) end for

Algorithm 1: The algorithm of computing $E(A)$.

(2) *Line 5–Line 14*. Calculate the probability P_0 corresponding to the values of $1/mn$.

(3) *Line 15–Line 21*. Calculate the $1/k_1k_2$ of probability $P_{k_1k_2}$ corresponding to $Q^i_{D_u}\|_{k-1}$ which represents $k-1$ coefficient of the probability generating function $Q^i_{D_u}$.

After getting the desired $E(A)$, we use an iterative approach to calculate the feature vector R. Set each of values R_{ij} in the eigenvalue R equal to constant $1/mn$, the original variable R called R_0. E indicates the normalized vector of sequence homologies. The α values have been studied in the literature [5, 16], so we directly use the best value 0.6. ε is a sufficiently small constant; iteration will eventually converge to approximate similarity score vector R. The process calculation of R is shown in Algorithm 2.

In Algorithm 2 we have the following.

(1) *Line 1–Line 4*. Set the initial value of the feature vector R_0.

Input: $E(A), \varepsilon$
Output: R
//initialize R_0 of V_1V_2
(1) for all $i \in V_1$, $j \in V_2$ do:
(2) $S_{ij} = 1/mn$
(3) end for
(4) $R_0 = S$
//Power Iteration Method Computation of R
(5) $k = 0$
(6) loop do:
(7) $R_{k+1} \leftarrow \alpha E(A)R_k + (1 - \alpha)E$
(8) $\delta = \|R_{k+1} - R_k\|$
(9) while $\delta > \varepsilon$

Algorithm 2: Computing R algorithm.

Table 3: Experimental environment.

Experimental environment	
Programming environment	QT, C++
Library function	QT and OGDF library function
Hardware environment	CPU clock speed of 3.3 GHz, memory of 4 G

(2) *Line 5–Line 9*. There is iterative calculation until the two values of feature vector difference are less than the set value of ε.

After getting the feature vector R, the last step of the C_PBNA algorithm is extracted by final comparison results from R as shown in Algorithm 3. C_PBNA adopted the method mentioned in Section 2.4; this method is to find perfect matching M.

In Algorithm 3 we have the following.

(1) *Line 1–Line 3*. Build bipartite graph G_{12} and compute the weight matrix by the feature vector R.

(2) *Line 3–Line 5*. Set the initial value of the y, E_{12}, M.

(3) *Line 7–Line 11*. Find an optimal augmenting path cover (v_u, v_v) by the max-flow min-cut theorem and then update the feasible labeling y.

(4) *Line 12–Line 14*. Update E_{12}, M until M is perfect matching.

3. Experiments and Results

The experiments in this research include two main parts. The first part shows that C_PBNA algorithm can obtain the results which are neglected by PBNA. Further, the second part of the experiments proves that results of C_PBNA are more biologically significant using GOC and GNAS (global network alignment score) as evaluation standards.

Experimental environment described in Table 3 indicates the conditions of conducting the experiment designed in this study. Besides, we make use of QT (a cross-platform application framework) library function directly to deal

Input: R Probabilistic graph $G_1 = (V_1, E_1)$, Probabilistic graph $G_2 = (V_2, E_2)$
Output: Maximum Weight Bipartite Matching M
//initialize Matrix W, G_{12}
(1) $G_{12} = \text{Build}BG(G_1, G_2)$
(2) W = matrix (W_{ij}) of weights on the edges of $1 - R$ with partite sets V_1 and V_2.
(3) $y \leftarrow 0$
(4) $E \leftarrow$ set of tight edges
(5) $M \leftarrow$ max cardinality matching for graph $G_{12} = (V_{12}, E_{12})$
//repeat find a perfect matching M
(6) while M is not a perfect matching do
(7) let $G = (V_{12}, E)$
(8) let $S \subseteq A$ be such that $|S| > |N(S)|$
(9) let $\varepsilon = \min_{a \in S, b \in B \setminus N(S)} \{w(a, b) - y(a) - y(b)\}$
(10) $\forall a \in S \quad y(a) = y(a) + \varepsilon$
(11) $\forall b \in N(S) \quad y(b) = y(b) - \varepsilon$
(12) update E_{12}, M
(13) end while
(14) return M

Algorithm 3: Extracting alignment results.

with problems associated with array, matrix, and sorting in PBNA and C_PBNA algorithm. The QT library function is available at http://qt-project.org/. In addition, the OGDF library function which can be obtained at http://ogdf.net/ is used to read and query biological network data.

The uncertain dataset used in the experiment obtained from the MINT database is network data of protein-protein interactions preprocessed by Todor et al. [9, 10]. As a result, providing that MINT network is of enough biological importance, the network information offered by KEGG database is divided into several smaller networks. Then, only the network with more than 10 nodes remains. Finally, we obtain 198 protein-protein interaction networks coming from 10 organisms. Table 4 shows statistical information of this network.

There are 198 networks comparing with each other, which will result in $C_{198}^2 = 19503$ groups of experiments. However the networks through KEGG are divided into a set of vertices function associated with proteins, and KEGG used a label to mark the set of proteins. The proteins from different sets share less similarity, which makes little sense to do the network comparison. Therefore we can get 122 groups of comparison experiments from the KEGG database.

3.1. Coherence Comparison of C_PBNA and PBNA. In order to prove that C_PBNA can discover the results neglected by PBNA, agreement evaluation criterion [9] is introduced in this research.

The definition of agreement is based on the same dataset. Hence, the proportion of common results discovered by both C_PBNA and PBNA in all alignments is shown as

$$\frac{|\text{Alignments in common}|}{|\text{All alignments}|}. \tag{13}$$

The score of this evaluation criterion is between 0 and 1. The larger the score is, the more common results these two algorithms have. Particularly, it shows that the results of both methods agree perfectly if the evaluation criterion equals 1 while it means that the results of these two algorithms are completely different if the evaluation criterion equals 0.

Figure 3: Agreement.

In this research, 122 groups of experiments are conducted and the result agreements of C_PBNA and PBNA algorithms are figured out in each group of experiment. Finally, we get 122 agreements plotted by the number of experiments on the horizontal axis and agreements on the vertical axis as shown in Figure 3.

The ordinate values are the 122 agreement values after sorting, and the abscissa values are the serial number of the experiment.

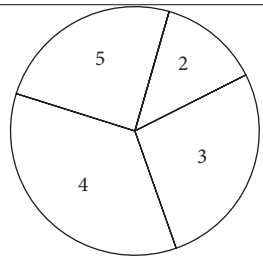
Table 5 shows the detailed agreement statistics of 122 groups of experimentation. In particular, the left Pie Chart is divided into 4 parts corresponding to the percentage of each category and Category 1 is not described in the Pie Chart due to its percentage of 0. For instance, Category 5 in Table 5 indicates that there are 30 experiments with the agreement

Table 4: Experimental data.

Organism	Number of networks	Number of proteins		Number of interactions	
		Average	Max	Average	Max
Cel	7	14.00	22	9.57	21
Dme	7	17.14	28	12.42	26
Eco	6	16.83	27	21.16	26
Hpy	1	11.00	11	7.00	7
Has	83	36.50	96	46.55	168
Mmu	43	16.23	40	11.16	33
Rno	13	14.69	30	11.00	22
Sce	34	32.91	106	80.32	313
Spo	3	11.00	11	10.00	10
Tpa	1	20.00	20	21.00	21

Table 5: Agreement statistics.

Category	Agreement	Quantity	Percentage
1	0–0.2	0	0%
2	0.2–0.4	16	13.1%
3	0.4–0.6	33	27.0%
4	0.6–0.8	43	35.2%
5	0.8–1	30	24.6%
Total	0–1	122	100%



between 0.8 and 1, which accounts for 24.6% of the total experiments.

The general distribution of the agreements is shown intuitively in Figure 4. We can see that the Agreement values are distributed within the range from 0.2 to 0.95. From the Pie Chart, we can further see that agreement scores less than 0.8 experiments accounted for 75% of the experiments, among which only 14% of the total experiment is less than 0.4 points. It indicates that both the C_PBNA results and PBNA results have many overlapping parts but have noticeable difference at the same time. The reason is that one of the most basic differences is that C_PBNA concludes all of the uncertain information while the PBNA method only utilizes half the uncertain information.

Therefore, we may draw a conclusion that neglecting uncertain information could lead to deviation. In addition, all the above shows that much more innovative result can be obtained through C_PBNA algorithm. The corresponding biological significance will be demonstrated by the following experiments.

3.2. Gene Ontology Consistency Comparisons of C_PBNA and PBNA. Gene ontology consistency (GOC) has been generally used to measure the biological significance of alignment results and we use it to evaluate biological meaning of alignment result by

$$\mathrm{GOC} = \sum_{\langle u,v\rangle \in V_{12}} \frac{|\mathrm{GO}(u) \cap \mathrm{GO}(v)|}{|\mathrm{GO}(u) \cup \mathrm{GO}(v)|}, \tag{14}$$

where GO(u) denotes the set of GO terms which label a protein u in gene ontology database.

Then, the GOC of each pair of proteins in alignment results is calculated, respectively. The bigger the GOC is, the more similar function these proteins have; especially, the maximum of GOC is 1 which means that these proteins have totally the same function. All GO data in this study comes from GO Consortium [19] and literature [14].

Similarly, in order to get alignment results of C_PBNA and PBNA, respectively, the GOC (the value of GOC is between 0 and 1) is calculated for each pair of proteins in 122 groups of experiments. Finally, we get 122 groups of results, among which there are 2 GOCs in each group. The distribution of the GOC is shown in Figure 4 and Table 6.

Table 6: GOC statistical data.

Category	Diversity	Quantity	Percentage
1	<−10%	0	0%
2	−10%–0%	26	21.3%
3	0%–10%	79	64.8%
4	>10%	17	13.9%
Total	−∞–+∞	122	100%

In Figure 4, x-coordinate denotes GOC value of C_PBNA algorithm and y-coordinate denotes GOC value of PBNA algorithm. Table 6 shows 122 groups of GOC value and the diversity of every two GOC values in each group.

As we can see in Table 6, for most of the results the value of x-coordinate is larger than y-coordinate. For instance, it includes 96 groups of experiments, 78.7% of the total experiments, in Category 3 and Category 4. Furthermore, the x-coordinate is higher than the y-coordinate more than 10% in 17 groups of experiment. These all indicate that in most of experiments C_PBNA algorithm may discover more

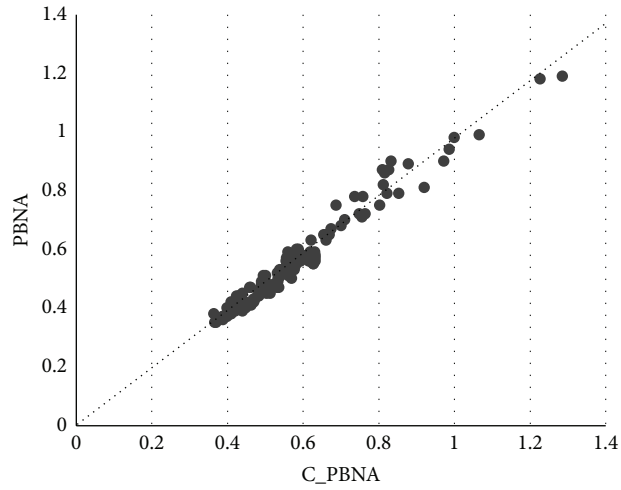


Figure 4: GOC statistics of PBNA and C_PBNA.

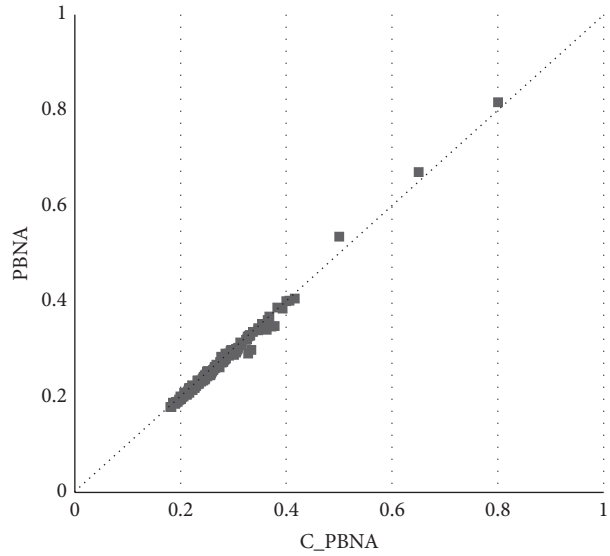


Figure 5: Functional coherence of alignments using PBNA and our method C_PBNA.

biologically significant results based on the same evaluation standard GOC.

In conclusion, C_PBNA and PBNA can obtain diverse results for biological networks alignment which is proved through the first experiment. Moreover, C_PBNA is demonstrated in the second experiment to be superior to PBNA in discovering biologically significant results since it uses all uncertain information while the PBNA algorithm neglects some uncertain information in biological networks.

3.3. Functional Coherence of the Alignments. The functional coherence of the alignments is motivated by the lack of automated and direct measures of ortholog-list quality. Comparing with GOC, the functional coherence of the alignments reports the average of the medians instead of the sum. And it maps each GO term to one or more of a standardized set of GO terms.

We can see in Figure 5 that PBNA and C_PBNA have very similar functional coherence values with only a few minor differences. One of the reasons is that the functional coherence function computes the similarity of a standardized set of GO terms instead of the aligned proteins directly. The other reason is that it reports the average of the medians, so it cannot tell whether a mapping has many highly similar terms or not. Since the median of a distribution is not an accurate representation of the entire distribution, the result it returns is not sensitive enough to tell the difference between different alignments.

3.4. GNAS Valuation Comparison of C_PBNA and PBNA. As one of the biological networks alignment algorithms, GNAS (global network alignment score) [20] is adopted as evaluation criterion in this paper defined in (15). Specifically the larger value of GNAS indicates more conserved interactions and higher sequence similarity:

$$\text{GNAS} = \alpha \times |E| + (1 - \alpha) \times \sum \text{seq}(u, v). \quad (15)$$

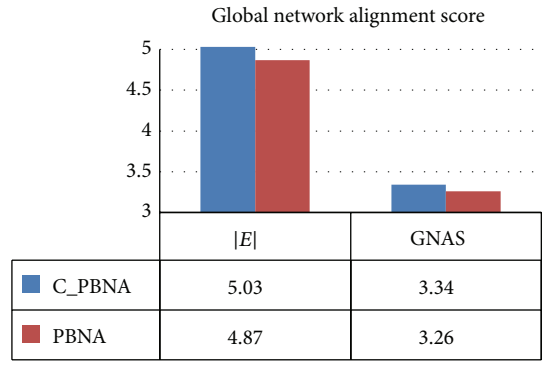


Figure 6: Comparison of $|E|$ and GNAS from C_PBNA and PBNA.

In this formula, $\text{seq}(u, v)$ denotes sequence similarity values of the two nodes; $|E|$ denotes the number of edges of G_{12} (for the definition of G_{12}, please refer to Section 2.4). Based on the common dataset, two groups of GNAS with 122 values in each group are obtained through C_PBNA and PBNA, respectively. The average values of GNAS and $|E|$ are showed in Figure 6.

As we can see in Figure 6, the values of $|E|$ and GNAS obtained from C_PBNA are superior to those from PBNA since C_PBNA adopts full uncertain information which increases the amount of conserved interactions.

3.5. Time Analysis. The running time of PBNA and C_PBNA is evaluated in this experiment. The most time-consuming step for both algorithms is constructing similarity matrix, which takes about 90% of the entire running time. Therefore, it is reasonable that we measure only this step's running time in order to evaluate the entire algorithm time efficiency. The results are shown in Table 7.

Table 7 indicates that the time spent in constructing similarity in C_PBNA is longer than its counterpart in PBNA,

Table 7: PBNA and C_PBNA algorithm time statistics.

Method	Average (second)	Max (second)
PBNA	125.4	545.5
C_PBNA	490.1	9014.6

because C_PBNA deals with both probabilistic networks, which takes more information into consideration. Network comparison with two uncertain networks is much more complex as we can see in Sections 2.1 and 2.2. When computing $E(A)$, in fact, the complexity of C_PBNA is $O((d_v^{\max} d_u^{\max})^2)$ while the complexity of PBNA is $O((d_v^{\max})^2)$ by PGF method mentioned above in Section 2.3. Although C_PBNA spends more time dealing with both probabilistic networks, its time performance is still acceptable. Furthermore, the results which we get have more biological significance, and C_PBNA can deal with the probabilistic networks directly instead of the preprocessing and transforming of the data into deterministic network.

4. Conclusions

Biological networks alignment is an important topic in bioinformatics. However, the network data has its inherent complexity of and the combine optimizes features of biological networks alignment are not clear, which make relevant algorithm study extremely challenging. A majority of the classic biological networks alignment algorithms are based on deterministic network while the alignment method for probabilistic networks is still under discussion.

In this paper, we propose a complete probabilistic model and a complete probabilistic biological algorithm for network comparison. Our approach has several advantages. First, our approach is based on complete probabilistic network, which takes the uncertainties of both networks instead of the single one into consideration. Consequently, our approach can take full advantage of the uncertainties properties of network comparison. Second, we model the network alignment using two probability matrices. Therefore, the uncertainties can be quantified by the probabilities of connections in the networks. As a result, our approach is capable of comparing two networks which both have uncertain properties. Third, we use a unified probabilistic model for different types of network alignment (deterministic, part probabilistic, and complete probabilistic), unlike other alignments which use different methods for different types of networks. Finally, the evaluation criteria including GOC and GNAS are used in the experiments to demonstrate that the results of C_PBNA and PBNA are different and that the results of the former algorithm are more biologically significant.

Usually, affinity propagation in probabilistic networks is random and probability factors have not been taken into consideration in this paper, and the effect of these factors on results will remain an open problem for the future research. The computational time increases as a result of using more probability information, which is a subject we will study in the next step.

Conflict of Interests

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

Acknowledgment

This work was supported by the Natural Science Foundation of Jiangsu Province under Grant no. BK2012742.

References

[1] X. L. Guo, L. Gao, and X. Chen, "Models and algorithms for alignment of biological networks," *Journal of Software*, vol. 21, no. 9, pp. 2089–2106, 2010.

[2] S. Zhang, G. X. Jin, X.-S. Zhang, and L. N. Chen, "Discovering functions and revealing mechanisms at molecular level from biological networks," *Proteomics*, vol. 7, no. 16, pp. 2856–2869, 2007.

[3] H. Ogata, W. Fujibuchi, S. Goto, and M. Kanehisa, "A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters," *Nucleic Acids Research*, vol. 28, no. 20, pp. 4021–4028, 2000.

[4] B. P. Kelley, R. Sharan, R. M. Karp et al., "Conserved pathways within bacteria and yeast as revealed by global protein network alignment," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 20, pp. 11394–11399, 2003.

[5] M. Koyutürk, Y. Kim, U. Topkara, S. Subramaniam, W. Szpankowski, and A. Grama, "Pairwise alignment of protein interaction networks," *Journal of Computational Biology*, vol. 13, no. 2, pp. 182–199, 2006.

[6] R. Singh, J. B. Xu, and B. Berger, "Pairwise global alignment of protein interaction networks by matching neighborhood topology," in *Proceedings of the 11th Annual International Conference on Research in Computational Molecular Biology (RECOMB '07)*, pp. 16–31, April 2007.

[7] M. Narayanan and R. M. Karp, "Comparing protein interaction networks via a graph match-and-split algorithm," *Journal of Computational Biology*, vol. 14, no. 7, pp. 892–907, 2007.

[8] G. W. Klau, "A new graph-based method for pairwise global network alignment," *BMC Bioinformatics*, vol. 10, supplement 1, article S59, 2009.

[9] E. Hirsh and R. Sharan, "Identification of conserved protein complexes based on a model of protein network evolution," *Bioinformatics*, vol. 23, no. 2, pp. e170–e176, 2007.

[10] R. Singh, J. B. Xu, and B. Berger, "Global alignment of multiple protein interaction networks with application to functional orthology detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 35, pp. 12763–12768, 2008.

[11] A. Todor, A. Dobra, and T. Kahveci, "Probabilistic biological network alignment," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 1, pp. 109–121, 2013.

[12] A. Chatraryamontri, A. Ceol, L. M. Palazzi et al., "MINT: the molecular interaction database," *Nucleic Acids Research*, vol. 35, no. 1, pp. D572–D574, 2007.

[13] Z. Zou, J. Li, H. Gao, and S. Zhang, "Mining frequent subgraph patterns from uncertain graph data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 9, pp. 1203–1218, 2010.

[14] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[15] G. H. Golub and C. van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, Md, USA, 2006.

[16] R. Singh, J. Xu, and B. Berger, "Global alignment of multiple protein interaction networks with application to functional orthology detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 35, pp. 12763–12768, 2008.

[17] C. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Dover, 1998.

[18] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[19] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[20] A. E. Aladağ and C. Erten, "SPINAL: scalable protein interaction network alignment," *Bioinformatics*, vol. 29, no. 7, pp. 917–924, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 867516, 9 pages
http://dx.doi.org/10.1155/2015/867516

Research Article

Detecting Protein-Protein Interactions with a Novel Matrix-Based Protein Sequence Representation and Support Vector Machines

Zhu-Hong You,[1] Jianqiang Li,[1] Xin Gao,[2] Zhou He,[3] Lin Zhu,[4] Ying-Ke Lei,[4] and Zhiwei Ji[4]

[1]*College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China*
[2]*Department of Medical Imaging, Suzhou Institute of Biomedical Engineering and Technology, Suzhou, Jiangsu 215163, China*
[3]*College of Information Science and Engineering, Guilin University of Technology, Guilin, Guangxi 541004, China*
[4]*School of Electronics and Information Engineering, Tongji University, Shanghai 200092, China*

Correspondence should be addressed to Jianqiang Li; lijq@szu.edu.cn and Xin Gao; xingaosam@yahoo.com

Received 1 October 2014; Revised 9 January 2015; Accepted 9 January 2015

Academic Editor: Yuedong Yang

Proteins and their interactions lie at the heart of most underlying biological processes. Consequently, correct detection of protein-protein interactions (PPIs) is of fundamental importance to understand the molecular mechanisms in biological systems. Although the convenience brought by high-throughput experiment in technological advances makes it possible to detect a large amount of PPIs, the data generated through these methods is unreliable and may not be completely inclusive of all possible PPIs. Targeting at this problem, this study develops a novel computational approach to effectively detect the protein interactions. This approach is proposed based on a novel matrix-based representation of protein sequence combined with the algorithm of support vector machine (SVM), which fully considers the sequence order and dipeptide information of the protein primary sequence. When performed on yeast PPIs datasets, the proposed method can reach 90.06% prediction accuracy with 94.37% specificity at the sensitivity of 85.74%, indicating that this predictor is a useful tool to predict PPIs. Achieved results also demonstrate that our approach can be a helpful supplement for the interactions that have been detected experimentally.

1. Introduction

Since detection of protein interactions is of fundamental importance to understand the molecular mechanism in biological systems, many researchers have focused on this area in postgenome era [1, 2]. Over the past decades, high-throughput experimental techniques, such as yeast two-hybrid (Y2H) system [3, 4] and mass spectrometry (MS), involving genome-wide detection of PPIs, have been developed to generate large amounts of interaction data. However, these traditional experimental methods are time-consuming and expensive, especially for genome-wide scale. In addition, the high-throughput biological experiment usually suffers from high rates of both false negatives and false positives [5]. Combining the experimental techniques with computational model is a promising direction to better understand the mechanisms of protein interactions at the molecular level and to unravel the global picture of PPIs in the cell [6, 7]. Hence, it is of great practical significance to build low cost protein detection systems and establish the reliable computational methods to facilitate the detection of PPIs.

So far, a variety of computational methods have been developed to effectively and accurately predict protein interactions [2, 8–10]. The computational approaches for in silico prediction can be roughly categorized into genome based approaches, network topology based approaches, literature knowledge based methods, and structure based approaches [11]. In addition, there are also some approaches that integrate interaction information from several different biological data sources [9, 10].

However, the aforementioned approaches cannot be implemented if prior information about the proteins is

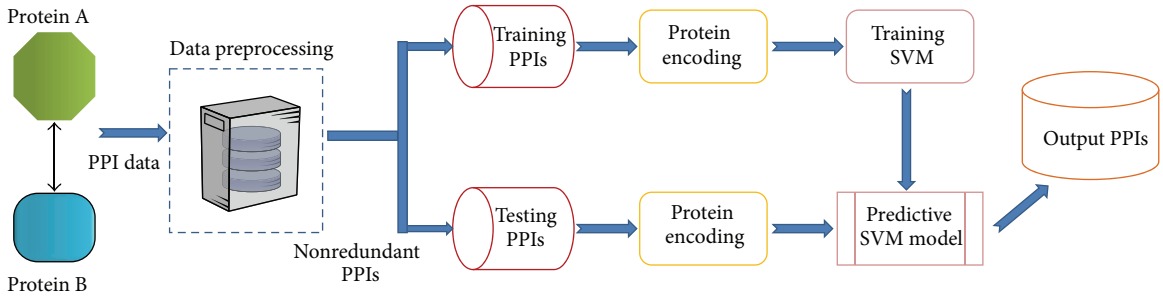


Figure 1: The schematic diagram for detecting protein-protein interactions by integrating experimental PPI data with SVM model.

not available [12]. Recently, the sequence-based approaches which derive information directly from protein amino acids sequence are of particular interest [13, 14]. Prediction of protein interactions from only protein sequence is a much more universal way [15, 16]. The previous works demonstrate that the RNA and protein sequences alone contain sufficient information [17, 18]. The previous researches demonstrated that the information of protein amino acid sequences is sufficient to predict PPIs. Although the sequence-based approaches can yield a high prediction accuracy of 80%~88%, it is necessary to design the novel approaches to further improve the prediction performance compared with the existing methods.

In recent years, many efforts have been made aiming to develop accurate approaches for identifying PPIs based on protein sequence information [19, 20]. Shen et al. built a prediction model by employing the conjoint triad feature extraction and support vector machine. When applied to predicting *human* PPIs, this method yields a high prediction accuracy of about 84% [21]. Because the conjoint triad method did not take the neighboring effect into account and protein interactions usually occur in the discontinuous amino acids segments in the sequence, Guo et al. proposed an approach based on SVM and autocovariance feature representation which extract the interactions information in the discontinuous amino acids segments in the sequence [22]. Their approach reached a prediction accuracy of 86.55%, when applied to predicting *saccharomyces cerevisiae* PPIs. Lately, You et al. developed a novel ensemble learning model to predict *Saccharomyces cerevisiae* PPIs from protein primary sequences directly [23]. In this study, the protein pairs retrieved from the database of interacting proteins (DIP) were encoded into feature vectors by using four kinds of protein sequences information. Focusing on dimension reduction, an effective feature extraction method PCA was then employed to construct the most discriminative new feature set. Finally, multiple extreme learning machines were trained and then aggregated into a consensus classifier by majority voting. The experimental results show that it is a very promising scheme for PPIs prediction.

In this study, we report a novel sequence-based method for the prediction of interacting protein pairs using a matrix-based protein sequence descriptors combined with support vector machine (SVM) algorithm. More specifically, we first represent each protein sequence as a feature matrix, from which a novel matrix-based protein descriptor is extracted to numerically characterize each protein sequence. Then we characterize a protein pair in different feature vectors by coding the vectors of two proteins in this protein pair. Finally, an SVM model is established using these feature vectors of the protein pair as input. To evaluate the prediction performance, the proposed method was applied to *Saccharomyces cerevisiae* and *Helicobacter pylori* PPI datasets. The experiment results show that our method can achieve 90.06% and 85.91% prediction accuracy with 94.37% and 83.33% specificity at the sensitivity of 85.74% and 85.27%, respectively. Achieved results demonstrate that the approach can be a helpful supplement for the interactions that have been detected experimentally.

2. Materials and Methodology

In this section, we outline the main idea behind the proposed method. The schematic diagram intuitively showing how to detect protein interactions using experimental PPIs data with computational model is given in Figure 1. Firstly, we briefly discuss the PPIs datasets which is employed in the study (the source code and the datasets are freely available at http://sites.google.com/site/zhuhongyou/data-sharing/ for academic use). Next we propose the novel matrix-based protein representation method. Finally, we briefly describe the computational model, SVM, used in this study.

2.1. Golden Standard Datasets. We evaluated the proposed method with two real PPIs datasets. The first one was collected from *Saccharomyces cerevisiae* core subset of database of interacting proteins (DIP). After the redundant protein pairs which contain a protein with fewer than 50 residues or have ≥40% sequence identity were deleted, the remaining 5,594 protein pairs comprise the golden standard positive dataset. The selection of golden standard negative dataset has an important impact on the prediction performance, and it can be artificially inflated by a bias towards dominant samples in the positive data. For golden standard negative dataset, we followed the previous work [22] assuming that the proteins in different subcellular compartments do not interact with each other.

After strictly following the steps in Guo's work, we finally obtained 5,594 protein pairs as the golden standard negative

dataset. By combining the above two golden standard positive and negative PPI datasets, the final whole PPI dataset consists of 11,188 protein pairs, where nearly half are from the positive dataset and half are from the negative dataset. The second one is a small-scale *Helicobacter pylori* PPIs dataset, which is composed of 2,916 protein pairs (1,458 interacting pairs and 1,458 noninteracting pairs) as described by Martin et al. [24].

2.2. Representing Proteins with Descriptors from Primary Protein Sequences. To successfully use the machine learning algorithm to detect PPIs from primary protein amino acids sequences, one of the computational challenges is to effectively characterize a protein sequence by a fixed length feature vector in which the important information content of proteins is fully encoded [25]. In this study, we propose a novel matrix-based protein sequence representation approach for predicting PPIs. Firstly, the protein sequence is transformed into a sparse matrix, which considered the properties of one amino acid and its vicinal amino acids and regarded any two continuous amino acids as a unit. Then the protein features are extracted from the obtained sparse matrix.

A protein sequence can be represented as a series of amino acids by their single character codes A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, and V. Consider a protein sequence with L amino acid residues:

$$S_1\, S_2\, S_3\, S_4\, S_5\, S_6\, S_7, \ldots, S_L, \tag{1}$$

where S_1 denotes the amino acid at protein chain position 1, S_2 denotes the amino acid at protein chain position 2, and so forth. L denotes the length of the protein sequence. We scan the protein sequence from left to right by stepping each two vicinal amino acids at a time, which considers the properties of one amino acid and its vicinal amino acid and regards any two continuous amino acids as a unit. Here the number of all possible pairs of amino acids (dipeptides) that can be extracted from the protein sequence is 400, that is, AA, AR, AN, ..., YV, and VV.

For step j ($j = 1, 2, 3, \ldots, L-1$), if the "S_jS_{j+1}" is the ith type of dipeptide, then we set the element $a_{ij} = 1$. The rest can be done in the same manner and then a protein sequence can be transformed into a 400 by $L-1$ matrix (see Table 1), namely, M, as follows:

$$M = \left(a_{ij}\right)_{400\times(L-1)},$$
$$a_{ij} = \begin{cases} 1, & \text{if } S_jS_{j+1} = \text{dipeptide}\,(i) \\ 0, & \text{others,} \end{cases} \tag{2}$$

where L is the length of protein sequence, $i = 1, 2, 3, \ldots, 400$, $j = 1, 2, 3, \ldots, L-1$, and dipeptide(i) denotes the ith type of dipeptides listed in Table 1. Here, each column of the matrix M is a unit vector, in which only one element is 1 and the others are all 0. We can see from Table 1 that the occurrence position of all kinds of dipeptides along the protein sequence is contained in the column of the matrix M. Meanwhile, the row of the matrix M denotes the ith kind of dipeptide appearing at the jth position within the protein sequence.

Table 1: The matrix-based representation for a protein amino acid sequence.

	S_1S_2	S_2S_3	S_3S_4	S_4S_5	$\cdots$	$S_{L-1}S_L$
AA	a_{11}	a_{12}	a_{13}	a_{14}	$\cdots$	$a_{1,L-1}$
AR	a_{21}	a_{22}	a_{23}	a_{24}	$\cdots$	$a_{2,L-1}$
AN	a_{31}	a_{32}	a_{33}	a_{34}	$\cdots$	$a_{3,L-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
VV	$a_{400,1}$	$a_{400,2}$	$a_{400,3}$	$a_{400,4}$	$\cdots$	$a_{400,L-1}$

Generally speaking, the matrix M transformed from protein amino acid sequence embodies the essential information including the information of its sequence order and sequence length of the protein sequence. Thus, given a protein primary sequence, we can design a matrix-based protein descriptor to represent it, which is capable of facilitating PPIs detections.

Low-rank approximation (LRA) is an important matrix analysis method, in which the cost function measures the fit between a given sparse matrix and an approximating matrix (the optimization variable), subject to a constraint that the approximating matrix has reduced rank [26]. Here, using LRA upon the obtained protein feature matrix, we derive a matrix-based descriptor to represent the protein sequence. For a feature matrix M, which denotes a $400 * (L-1)$ matrix, the LRA of the data can be written as follows:

$$\min_{\widehat{M}} \quad \left\|M - \widehat{M}\right\|_F \tag{3}$$

$$\text{Subject to:} \quad \text{rank}\left(\widehat{M}\right) \le r, \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm. The above minimization problem has analytic solution in terms of the singular value decomposition (SVD) of the data matrix M.

Let $M = U\Sigma V^T \in R^{m\times n}$ be the SVD of M and partition $U, \Sigma =: \text{diag}(\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_{400})$, and N as follows:

$$U =: \begin{bmatrix} U_1 & U_2 \end{bmatrix},$$
$$\Sigma =: \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, \tag{5}$$
$$V =: \begin{bmatrix} V_1 & V_2 \end{bmatrix},$$

where Σ_1 is a $r \times r$ matrix, U_1 is $m \times r$, and V_1 is $n \times r$. Then the rank-r matrix is obtained as follows:

$$\widehat{M}^* = U_1\Sigma_1V_1^T, \tag{6}$$

where $\|M - \widehat{M}^*\|_F = \min_{\text{rank}(\widehat{M})\le r}\|M - \widehat{M}\|_F = \sqrt{\sigma_{r+1}^2 + \sigma_{r+2}^2 + \cdots + \sigma_m^2}$.

Then we compute the square root of the reduced matrix Σ_1 to obtain $\Sigma_1^{1/2}$ with dimensions r-by-r. Finally, we can get a $400 * r$ matrix $U_1\Sigma_1^{1/2}$, which contains the information of protein sequence order. It should be noticed that the feature matrix M for different protein sequences sometime have different columns with each other, which shows that these

protein primary sequences are of nonequal length. However, the $U_1\Sigma_1^{1/2}$ for different protein sequences are $400 * r$ matrix.

We build a vector (row matrix) from the obtained matrix $U_1\Sigma_1^{1/2}$ by concatenating all rows, from 1 to 400, of matrix $U_1\Sigma_1^{1/2}$. Therefore, the matrix-based protein descriptor consists of a total of $400 * r$ descriptor values; that is, a $400 * r$ dimensional vector has been built to represent the protein sequence. Considering the trade-off between the overall prediction accuracy and computational complexity for extracting protein sequence descriptors, the optimal rank is $k = 4$. Thus, we set k to 4 in this study. A representation of an interaction pair is formed by concatenating the descriptors of two protein sequences in this protein pairs.

2.3. Support Vector Machine. Machine learning has been seen as useful and reliable in many applications. Various machine learning techniques can be employed to predict the PPIs. Among them, support vector machine (SVM) is one of the popular learning algorithms based on statistical learning theory [27]. Here we give a brief introduction to the basic idea of SVM.

The goal of the SVM algorithm is to find an optimal hyperplane that separates the training samples by a maximal margin, with all positive samples lying on one side and all negative samples lying on the other side. Suppose that we are given a training dataset of N instance-labeled pairs $X = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ with input data $x_i \in R^n$ and labeled output data $y_i \in \{+1, -1\}$. The SVM algorithm solves the quadratic optimization problem as minimizing the function as below:

$$\min_{w,b,\xi} \quad \frac{\langle w \cdot w \rangle}{2} + C\sum_{i=1}^{N}\xi_i \tag{7}$$

subject to

$$\begin{aligned} y_i\left(\langle w \cdot x_i \rangle + b\right) &\geq 1 - \xi_i, \\ \xi_i &\geq 0, \\ (i = 1, 2, 3, \ldots, N), \end{aligned} \tag{8}$$

where w is the normal vector of hyperplane; b is the bias of hyperplane; C is the penalty factor; ξ_i is the slack variable.

Since $\|w\|^2$ is convex, minimizing (7) under linear constraints (8) can be solved with Lagrange multipliers. Further, the aforementioned optimization problem can be transferred to a dual form as maximizing the function

$$L(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j \left\langle x_i \cdot x_j \right\rangle \tag{9}$$

subject to

$$\begin{aligned} &\sum_{i=1}^{N} y_i \alpha_i = 0, \\ &0 \leq \alpha_i \leq C, \\ &i = 1, 2, 3, \ldots, l, \end{aligned} \tag{10}$$

where $C \geq 0$, $\alpha_i = [\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_l]^T$, and $\alpha_i \geq 0$, $(i = 1, 2, 3, \ldots, l)$ are coefficients corresponding to x_i. x_i with nonzero α_i is called support vector.

In real applications, the training samples are not linearly separable in its original space. Usually, the training samples x_i are mapped into a high-dimensional feature space through some nonlinear function ϕ. Then SVM finds a linear separating hyperplane with the maximal margin in this higher-dimensional space. Furthermore, $K(x_i, x_j) = \phi(x_i)^T \cdot \phi(x_j)$ is called the kernel function. Actually, the flexibility and classification power of SVM reside in its kernel functions, since they make it possible to discriminate within challenging datasets. Typical kernel functions for SVM include polynomial function, linear function, sigmoid function, and radial basis function (RBF):

polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + \gamma)^D$, $\gamma > 0$;

linear: $K(x_i, x_j) = x_i^T x_j$;

sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + B)$;

radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$, $\gamma > 0$;

here, D, B, and γ are kernel parameters which are set a priori.

If we replace samples x_i with their mapping in the feature space $\phi(x_i)$, (9) becomes

$$L(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j K\left(x_i, x_j\right) \tag{11}$$

and the decision function becomes

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{N_S}\alpha_i y_i K\left(x_i, x\right) + b\right), \tag{12}$$

where N_S is the number of SV, $x = [x_1, x_2, x_3, \ldots, x_l]$ is the input sample, and α_i and y_i are Lagrange multipliers.

3. Results and Discussion

In the section, we describe our simulation methodology and present the experimental results that evaluate the effectiveness of our schemes. The proposed sequence-based PPI predictor was implemented using MATLAB platform. For SVM algorithm, the LIBSVM implementation available from http://www.csie.ntu.edu.tw/~cjlin/libsvm/ was utilized, which was originally developed by Chang and Lin [28]. As the kernels, four kinds of kernel functions, radial basis function (RBF), polynomial function, linear function, and sigmoid function, were selected to implement the experiment. The optimized parameters for the SVM were obtained with a grid search approach. In the simulation, all the experiments were carried out on a computer with 3.1 GHz 2-Core CPU, 12 GB memory, and Windows operating system.

Table 2: Comparing the prediction performance by the proposed method and some state-of-the-art works on the *yeast* dataset. Here, N/A means not available.

Model	Test set	SN (%)	PPV (%)	ACC (%)	MCC (%)
Proposed method	SVM	**85.74 ± 0.94**	**93.84 ± 0.98**	**90.06 ± 0.64**	**82.03 ± 1.03**
Guos' work	ACC	89.93 ± 3.68	88.87 ± 6.16	89.33 ± 2.67	N/A
	AC	87.30 ± 4.68	87.82 ± 4.33	87.36 ± 1.38	N/A
Zhous' work	SVM + LD	87.37 ± 0.22	89.50 ± 0.60	88.56 ± 0.33	77.15 ± 0.68
Yangs' work	Cod1	75.81 ± 1.20	74.75 ± 1.23	75.08 ± 1.13	N/A
	Cod2	76.77 ± 0.69	82.17 ± 1.35	80.04 ± 1.06	N/A
	Cod3	78.14 ± 0.90	81.86 ± 0.99	80.41 ± 0.47	N/A
	Cod4	81.03 ± 1.74	90.24 ± 1.34	86.15 ± 1.17	N/A

3.1. Measures for the Prediction Performance. In the study, fivefold cross-validation technique has been employed to evaluate the performance of the proposed model. In the fivefold cross-validation technique, the whole dataset is randomly divided into five subsets, where each subset consists of nearly equal number of interacting and noninteracting protein pairs. Four subsets are used for training and the remaining set for testing. This process is repeated five times so that each subset is used once for testing. The performance of method is average performance of method on five sets.

Several evaluation measures have been used in the study to measure the predictive ability of the proposed method. The parameters are as follows: (1) the overall prediction accuracy (ACC) is the percentage of correctly identified interacting and noninteracting protein pairs; (2) the sensitivity (SN) is the percentage of correctly identified interacting protein pairs; (3) the specificity (SP) is the percentage of correctly identified noninteracting protein pairs; (4) the positive predictive value (PPV) is the positive prediction value; (5) the negative predictive value (NPV) is the negative prediction value; (6) the F-score is a weighted average of the PPV and sensitivity, where an F-score reaches its best value at 1 and worst score at 0; (7) the Matthew correlation coefficient (MCC) is more stringent measure of prediction accuracy accounting for both under- and overpredictions. These parameters are defined as follows:

$$\begin{aligned}
\mathrm{ACC} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}},\\
\mathrm{SN} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\\
\mathrm{SP} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},\\
\mathrm{PPV} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\\
\mathrm{NPV} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}},\\
F1 &= 2\times\frac{\mathrm{SN}\times\mathrm{PPV}}{\mathrm{SN}+\mathrm{PPV}},\\
\mathrm{MCC} &= \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}},
\end{aligned} \tag{13}$$

where true positive (TP) is the number of true PPIs that are predicted correctly; false negative (FN) is the number of true PPIs that are predicted to be noninteracting pairs; false positive (FP) is the number of true noninteracting pairs that are predicted to be PPIs; and true negative (TN) is the number of true noninteracting pairs that are predicted correctly.

The above-mentioned parameters rely on the selected threshold. The area under the ROC curve (AUC), which is threshold-independent for evaluating the performances, can be easily calculated according to the following formula [29]:

$$\mathrm{AUC} = \frac{S_0 - n_0\left(n_0+1\right)/2}{n_0\times n_1}, \tag{14}$$

where n_0 and n_1 denote the number of positive and negative samples, respectively, and S_0 is the sum of the ranks of all positive samples in the list of all samples ranked in increasing order by estimated probabilities belonging to positive. AUC values can give us a good insight into performance comparison of different prediction methods. Although the AUC is threshold-independent, an appropriate threshold must be selected for the final decision. For the classifier which outputs a continuous numeric value to represent the confidence or probability of a sample belonging to the predicted class, adjusting the classification threshold will lead to different confusion matrices which decide different ROC points [21].

3.2. Prediction Performance of Proposed Model. We evaluated the performance of the proposed model using the DIP PPIs data as investigated in Guo et al. [22]. To guarantee that the experimental results are valid and can be generalized for making predictions regarding new data, the fivefold cross-validation is utilized to evaluate the performance of the proposed method. The whole PPI dataset is randomly divided into five subsets of roughly equal size, and each subset consists of nearly equal number of interacting and noninteracting protein pairs. Four out of these five subsets are used for training and the remaining one for test. This process is repeated five times such that each subset is used once and only once for test. The results are then averaged over the five runs to ensure the highest level of fairness.

The prediction performance of SVM predictor with matrix-based protein sequence representation across five runs is shown in Table 2. It can be observed from Table 2 that high prediction accuracy 90.06% is obtained for the proposed model. To better investigate the prediction ability

Table 3: Comparing the prediction performance by the proposed method and amino acid dipeptide composition method on the yeast dataset.

Methods	Kernel	Mean/std.	Testing							
			ACC	SN	SP	PPV	NPV	$F1$	MCC	AUC
The proposed method	Sigmoid	Mean	0.8734	0.8379	0.9092	0.9032	0.8474	0.8693	0.7784	0.9385
		Variance	0.0073	0.0093	0.0078	0.0087	0.0063	0.0088	0.0111	0.0071
	Gaussian	Mean	**0.9006**	**0.8574**	**0.9437**	**0.9384**	**0.8689**	**0.8961**	**0.8203**	**0.9528**
		Variance	0.0064	0.0094	0.0095	0.0098	0.0048	0.0076	0.0103	0.0064
	Polynomial	Mean	0.8963	0.8517	0.9408	0.9351	0.8639	0.8915	0.8134	0.9506
		Variance	0.0079	0.0072	0.0112	0.0118	0.0050	0.0085	0.0124	0.0061
	Linear	Mean	0.8642	0.8267	0.9016	0.8938	0.8389	0.8589	0.7646	0.9238
		Variance	0.0048	0.0098	0.0114	0.0103	0.0073	0.0052	0.0068	0.0038
AADC method	Sigmoid	Mean	0.6776	0.6726	0.6825	0.6792	0.6760	0.6758	0.5630	0.7343
		Variance	0.0088	0.0194	0.0098	0.0107	0.0136	0.0133	0.0062	0.0129
	Gaussian	Mean	0.8654	0.8349	0.8959	0.8892	0.8443	0.8612	0.7666	0.9292
		Variance	0.0065	0.0104	0.0047	0.0041	0.0119	0.0058	0.0095	0.0087
	Polynomial	Mean	0.8514	0.8196	0.8833	0.8754	0.8305	0.8465	0.7465	0.7540
		Variance	0.0063	0.0144	0.0078	0.0072	0.0110	0.0077	0.0090	0.3751
	Linear	Mean	0.8409	0.8150	0.8668	0.8597	0.8240	0.8367	0.7320	0.9021
		Variance	0.0060	0.0050	0.0146	0.0128	0.0070	0.0049	0.0080	0.0030

of our model, we also calculated the values of sensitivity, precision, MCC, and AUC. From Table 2, we can see that our model gives good prediction performance with an average sensitivity value of 85.74%, precision value of 93.84%, MCC value of 82.03%, and AUC value of 95.28%. Further, it can also be seen from Table 2 that the standard deviation of sensitivity, precision, accuracy, MCC, and AUC is as low as 0.0094, 0.0098, 0.0064, 1.03, and 0.0064, respectively.

We further compared our method with those of Guo et al. [22], Zhou et al. [30], and Yang et al. [31], where the SVM, SVM, and KNN were performed with the conventional autocovariance, local descriptor, and local descriptor representation as the input feature vectors, respectively. From Table 2, we can see that the performance of all of these methods with different machine learning models and sequence-based feature representation methods are lower than ours, which indicates the advantages of our method. To sum up, we can readily conclude that the proposed approach generally outperforms the previous model with higher discrimination power for predicting PPIs based on the information of protein sequences. Therefore, we can see clearly that our model is a much more appropriate method for predicting new protein interactions compared with the other methods. Consequently, it makes us more convinced that the proposed method can be very helpful in assisting the biologist to contribute to the design and validation of experimental studies and in the prediction of interaction partners.

3.3. Comparison between the Proposed Model and AADC Method. The amino acid dipeptide composition (AADC) is a representation method for protein sequences that count the frequency of occurrence of adjacent pairs of amino acids. Similar to the proposed matrix-based protein sequence representation method, AADC only needs the information of protein amino acids; no attention is paid to the physicochemical properties of amino acids or other pieces of biological information about proteins. To demonstrate the performance of the proposed model, we further compared the proposed protein feature representation methods with AADC method.

The prediction performance of SVM predictor with the aforementioned two protein sequence representation across five runs is shown in Table 3. It can be observed from Table 3 that high prediction accuracy of 90.06% is achieved for the proposed model with Gaussian kernel function. To better investigate the prediction ability of our model, we also calculated the values of sensitivity, specificity, PPV, NPV, F-score, MCC, and AUC. From Table 3, we can see that our model gives good prediction performance with an average sensitivity value of 85.74%, specificity value of 94.37%, PPV value of 93.84%, NPV value of 86.89%, F-score value of 89.61%, MCC value of 82.03%, and AUC value of 95.28%. Further, it can also be seen from Table 3 that the standard deviation of accuracy, sensitivity, specificity, PPV, NPV, F-score, MCC, and AUC is as low as 0.0064, 0.0094, 0.0095, 0.0098, 0.0048, 0.0076, 0.0103, and 0.0064, respectively. The performance of the proposed model with other kernel functions including sigmoid function, polynomial function, and linear function is also demonstrated in Table 3.

In addition, the prediction performance of AADC based model is shown in Table 3. The AUC of the AADC model with Gaussian kernel is 0.9292, which is lower than that of the proposed model. The overall accuracy, sensitivity, specificity, PPV, NPV, $F1$ score, and MCC of AADC model are, respectively, 86.54%, 83.49%, 89.59%, 88.92%, 84.43%, 86.12%, and 76.66% as illustrated in Table 3. Hence, it can be seen that almost all evaluation measures of the proposed model are better than those of AADC method.

We also conduct experiment to characterize the sensitivity (i.e., the size of true positives that can be detected by our method) and specificity (i.e. 1 – false positive rate) of the proposed approach for different activation functions

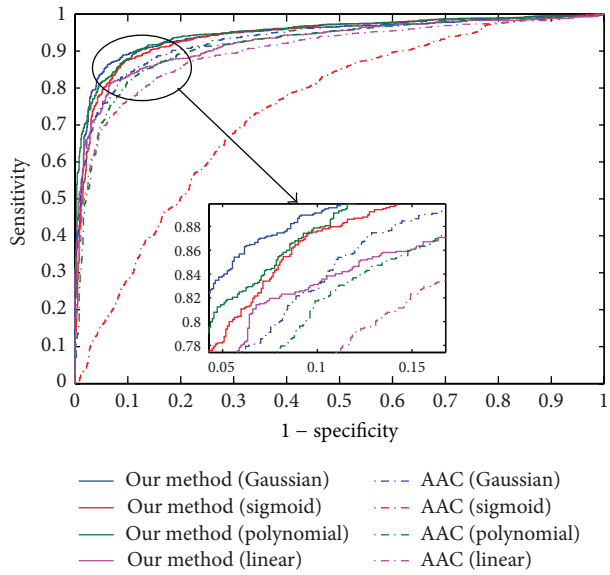


Figure 2: The ROC (receiver operator characteristic) curve illustrating the performance of different activation functions. The curve presents the true positive rate (sensitivity) against the false positive rate (1 – specificity).

(see Figure 2). The results in Figure 2 are reported using receiver operator characteristic (ROC) curves, which plot the achievable sensitivity at a given specificity (1 – false positive rate). Good performance is reflected in curves with a stronger bend towards the upper-left corner of the ROC graph (i.e., high sensitivity is achieved with a low false positive rate). We found that proposed method achieved over 89 percent detection rate with less than 10 percent false positive rate. The results demonstrate that the proposed matrix-based model can successfully classify positive and negative samples in all five activation functions that we investigated. Our algorithm can perfectly classify interacting and noninteracting protein pairs with only a few exceptions.

To sum up, considering the high efficiency as well as the good performance we can readily conclude that the proposed approach generally outperforms the AADC model with higher discrimination power for predicting PPIs based on the information of protein sequences. Therefore, we can see clearly that our model is a much more appropriate method for predicting new protein interactions compared with the other methods.

3.4. Comparing the Prediction Performance between Our Method and Other Existing Methods. In order to highlight the advantage of our model, it was also tested by *Helicobacter pylori* dataset. This dataset gives a comparison of proposed method with several previous works including phylogenetic bootstrap [32], signature products [24], HKNN [33], and boosting [34]. The methods of phylogenetic bootstrap, signature products, and HKNN are based on individual classifier system to infer PPIs, while the methods of boosting belong to ensemble-based classifiers.

The average prediction results of 10-fold cross-validation over five different approaches are demonstrated in Table 4.

Table 4: Performance comparison of different methods on the *H. pylori* dataset. Here, N/A means not available.

Methods	SN (%)	PE (%)	ACC (%)	MCC (%)
Phylogenetic bootstrap	69.8	80.2	75.8	N/A
HKNN	86	84	84	N/A
Signature products	79.9	85.7	83.4	N/A
Boosting	80.37	81.69	79.52	70.64
Proposed method	**85.27**	**83.33**	**85.91**	**75.53**

From Table 4, we can see that the average prediction performance, that is, sensitivity, precision, accuracy, and MCC achieved by proposed predictor, are 85.27%, 83.33%, 85.91%, and 75.53%, respectively. It clearly shows that our method outperforms all other individual classifier-based methods and the ensemble classifier systems (i.e., boosting). All these results demonstrate that the proposed method not only achieves accurate performance, but also substantially improves precision in the prediction of PPIs.

4. Conclusions

In this paper, we proposed an efficient and accurate learning technique, which utilizes the information of protein amino acid sequence order and distribution, for accurate identification PPIs at considerably high speed. It is well known that the order and distributions of dipeptide possess more pieces of information than those of amino acid dipeptide composition (AADC), so the main advantage is that this algorithm can extract more pieces of information hidden in protein primary sequences than AADC can. Then, the application of SVM predictor ensures reliable recognition with minimum error. Experimental results demonstrated that the proposed method performed significantly well in distinguishing interacting and noninteracting protein pairs. It was observed that the proposed method achieved the mean classification accuracy of 90.06% using fivefold cross-validation. Meanwhile, comparative study was conducted on the proposed method and other existing methods. The experimental results showed that our method outperformed these works in terms of classification accuracy.

Conflict of Interests

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

Acknowledgments

This work is supported in part by the National Science Foundation of China, under Grants 61102119, 61373086, 61202347, and 61401385, and in part by Fundamental Research Funds for the Central Universities, under Grant no. CDJZR12180012. The authors would like to thank all the guest editors and anonymous reviewers for their constructive advices.

References

[1] N. J. Krogan, G. Cagney, H. Yu et al., "Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*," *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.

[2] Q. C. Zhang, D. Petrey, L. Deng et al., "Structure-based prediction of protein-protein interactions on a genome-wide scale," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[3] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.

[4] K. B. Stibius and K. Sneppen, "Modeling the two-hybrid detector: experimental bias on protein interaction networks," *Biophysical Journal*, vol. 93, no. 7, pp. 2562–2566, 2007.

[5] Z.-H. You, Y.-K. Lei, J. Gui, D.-S. Huang, and X. Zhou, "Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data," *Bioinformatics*, vol. 26, no. 21, Article ID btq510, pp. 2744–2751, 2010.

[6] Y. Yang, H. Zhao, J. Wang, and Y. Zhou, "SPOT-Seq-RNA: predicting protein–RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction," in *Protein Structure Prediction*, pp. 119–130, Springer, New York, NY, USA, 2014.

[7] J.-F. Yu, X. Sun, and J.-H. Wang, "TN curve: a novel 3D graphical representation of DNA sequence based on trinucleotides and its applications," *Journal of Theoretical Biology*, vol. 261, no. 3, pp. 459–468, 2009.

[8] Z.-H. You, J.-Z. Yu, L. Zhu, S. Li, and Z.-K. Wen, "A MapReduce based parallel SVM for large-scale predicting protein-protein interactions," *Neurocomputing*, vol. 145, pp. 37–43, 2014.

[9] X.-M. Zhao, Y. Wang, L. Chen, and K. Aihara, "Protein domain annotation with integration of heterogeneous information sources," *Proteins: Structure, Function and Genetics*, vol. 72, no. 1, pp. 461–473, 2008.

[10] X.-M. Zhao, X. Li, L. Chen, and K. Aihara, "Protein classification with imbalanced data," *Proteins: Structure, Function and Genetics*, vol. 70, no. 4, pp. 1125–1132, 2008.

[11] Y. Yang, E. Faraggi, H. Zhao, and Y. Zhou, "Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates," *Bioinformatics*, vol. 27, no. 15, pp. 2076–2082, 2011.

[12] Y. Yang and Y. Zhou, "Specific interactions for ab initio folding of protein terminal regions with secondary structures," *Proteins: Structure, Function and Genetics*, vol. 72, no. 2, pp. 793–803, 2008.

[13] H. Lin, "The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 252, no. 2, pp. 350–356, 2008.

[14] W. Chen, P.-M. Feng, H. Lin, and K.-C. Chou, "IRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Research*, vol. 41, no. 6, article e68, 2013.

[15] X.-Y. Pan, Y.-N. Zhang, and H.-B. Shen, "Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features," *Journal of Proteome Research*, vol. 9, no. 10, pp. 4992–5001, 2010.

[16] Z.-H. You, S. Li, X. Gao, X. Luo, and Z. Ji, "Large-scale protein-protein interactions detection by integrating big biosensing data with computational model," *BioMed Research International*, vol. 2014, Article ID 598129, 9 pages, 2014.

[17] J.-F. Yu, X. Sun, and J.-H. Wang, "A novel 2D graphical representation of protein sequence based on individual amino acid," *International Journal of Quantum Chemistry*, vol. 111, no. 12, pp. 2835–2843, 2011.

[18] J.-F. Yu, J.-H. Wang, and X. Sun, "Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation," *MATCH: Communications in Mathematical and in Computer Chemistry*, vol. 63, no. 2, pp. 493–512, 2010.

[19] X.-M. Zhao, Y. Wang, L. Chen, and K. Aihara, "Gene function prediction using labeled and unlabeled data," *BMC Bioinformatics*, vol. 9, no. 1, article 57, 2008.

[20] X.-M. Zhao, L. Chen, and K. Aihara, "Protein function prediction with high-throughput data," *Amino Acids*, vol. 35, no. 3, pp. 517–530, 2008.

[21] J. Shen, J. Zhang, X. Luo et al., "Predicting protein-protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 11, pp. 4337–4341, 2007.

[22] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences," *Nucleic Acids Research*, vol. 36, no. 9, pp. 3025–3030, 2008.

[23] Z.-H. You, Y.-K. Lei, L. Zhu, J. Xia, and B. Wang, "Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis," *BMC Bioinformatics*, vol. 14, supplement 8, article S10, 2013.

[24] S. Martin, D. Roe, and J.-L. Faulon, "Predicting protein-protein interactions using signature products," *Bioinformatics*, vol. 21, no. 2, pp. 218–226, 2005.

[25] D.-S. Huang, X.-M. Zhao, G.-B. Huang, and Y.-M. Cheung, "Classifying protein sequences using hydropathy blocks," *Pattern Recognition*, vol. 39, no. 12, pp. 2293–2300, 2006.

[26] I. Markovsky and K. Usevich, "Software for weighted structured low-rank approximation," *Journal of Computational and Applied Mathematics*, vol. 256, pp. 278–292, 2014.

[27] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[28] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[29] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001.

[30] Y. Z. Zhou, Y. Gao, and Y. Y. Zheng, "Prediction of protein-protein interactions using local description of amino acid sequence," in *Advances in Computer Science and Education Applications, Part II*, M. Zhou and H. H. Tan, Eds., vol. 202 of *Communications in Computer and Information Science*, pp. 254–262, 2011.

[31] L. Yang, J.-F. Xia, and J. Gui, "Prediction of protein-protein interactions from protein sequence using local descriptors," *Protein and Peptide Letters*, vol. 17, no. 9, pp. 1085–1090, 2010.

[32] J. R. Bock and D. A. Gough, "Whole-proteome interaction mining," *Bioinformatics*, vol. 19, no. 1, pp. 125–134, 2003.

[33] L. Nanni, "Hyperplanes for predicting protein-protein interactions," *Neurocomputing*, vol. 69, no. 1–3, pp. 257–263, 2005.

[34] M.-G. Shi, J.-F. Xia, X.-L. Li, and D.-S. Huang, "Predicting protein-protein interactions from sequence using correlation coefficient and high-quality interaction dataset," *Amino Acids*, vol. 38, no. 3, pp. 891–899, 2010.