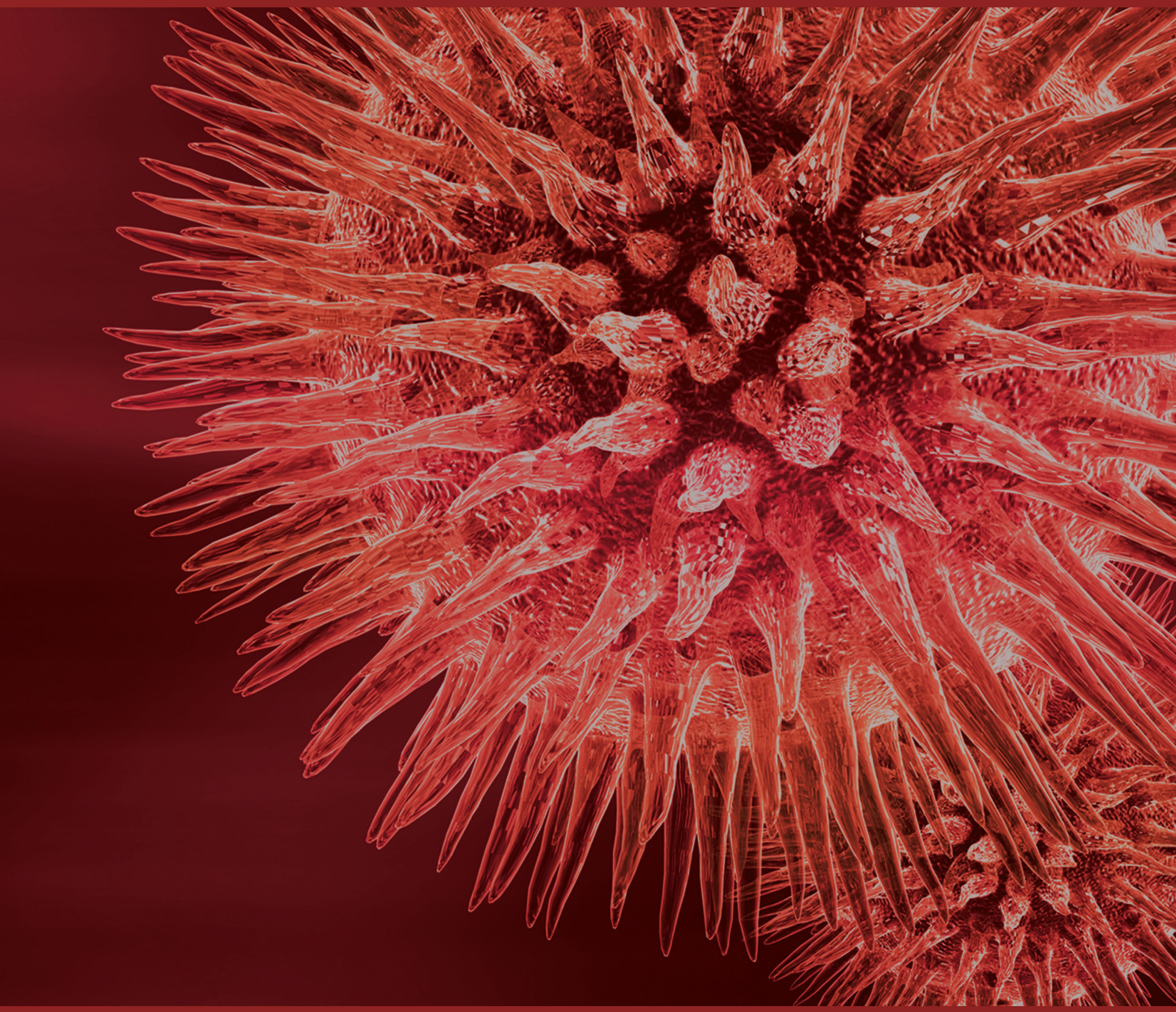
BioMed Research International

Intelligent Informatics in Translational Medicine

Guest Editors: Hao-Teng Chang, Tatsuya Akutsu, Sorin Draghici, Oliver Ray, and Tun-Wen Pai



Intelligent Informatics in Translational Medicine

BioMed Research International

Intelligent Informatics in Translational Medicine

Guest Editors: Hao-Teng Chang, Tatsuya Akutsu, Sorin Draghici, Oliver Ray, and Tun-Wen Pai

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 717210, 2 pages
http://dx.doi.org/10.1155/2015/717210

Editorial

Intelligent Informatics in Translational Medicine

Hao-Teng Chang,[1,2,3] Tatsuya Akutsu,[4] Sorin Draghici,[5] Oliver Ray,[6] and Tun-Wen Pai[7,8]

[1]*Graduate Institute of Basic Medical Science, College of Medicine, China Medical University, Taichung 40402, Taiwan*
[2]*Department of Computer Science and Information Engineering, Asia University, Taichung 41354, Taiwan*
[3]*Department of Science Education, Affiliated Dongyang Hospital of Wenzhou Medical University, Dongyang, Zhejiang 322100, China*
[4]*Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto 606-8501, Japan*
[5]*Department of Computer Science, Wayne State University, Detroit, MI 48202, USA*
[6]*Department of Computer Science, University of Bristol, Bristol BS8 1UB, UK*
[7]*Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung 20224, Taiwan*
[8]*Center of Excellence for the Oceans, National Taiwan Ocean University, Keelung 20224, Taiwan*

Correspondence should be addressed to Tun-Wen Pai; twp@mail.ntou.edu.tw

Received 14 March 2015; Accepted 14 March 2015

Starting from 2008, the organization committee members, Dr. Hui-Huang Hsu, Dr. Tun-Wen Pai, Dr. Oliver Ray, and Dr. Hao-Teng Chang, yearly held the International Workshop on Intelligent Informatics in Biology and Medicine (IIBM), which is used to be held in conjugation with International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS). The main purpose for establishing this forum is to gather scientists from multidisciplinary fields including biology, medicine, computer science, statistics, and informatics to discuss how to face the new big data era and how to employ the hundreds of thousands of biological datasets including gene expression profiles, genetic information, proteomes, metabolomes, and even molecular imaging in clinical medicine. So many genomes from various species have been sequenced in quick succession. One important reason to conduct these genome projects is to translate useful and relevant information to biomedical research as well as therapeutic strategy development. From bench to bedside, bioinformatics researches have been provided strong and powerful tools to accelerate the analyses of complicated datasets.

In the past seven years, IIBM successfully brought together computer scientists, medical doctors, biologists, statisticians, and chemists to present and discuss current topics on genomics, epigenomics, GWAS/PheWAS, imaging processing, healthcare information, big data analyses, and so on. On the platform of IIBM, the scientists share their know-how and research experiences on processing the complexity and volume of experimental data from next generation sequencing and mass spectrometry technologies. Many various sophisticated computational methodologies have been designed and developed to support the new detection techniques which can effectively improve the quality of healthcare. Intelligent information technologies indeed facilitate and accelerate researches from basic to clinical investigation in terms of translational medicine.

To record the discoveries of talents and gather more contributions to these fields, this special issue was launched and supported by this journal. This special issue focuses on addressing the questions of biomedical sciences through computational technologies. It also describes the information processes employed, with an emphasis on forthcoming high throughput technologies and biomedical systems, which provides opportunities to discuss recent hot topics and progresses in the area of biomedicine for the academic and industrial societies. This special issue received 19 submissions from which 9 papers were selected for publication. These papers address the data-analytical method design, algorithm development, mathematical modeling, and computational simulation techniques of some translational medical applications.

In "Computational Biophysical, Biochemical, and Evolutionary Signature of Human R-Spondin Family Proteins, the Member of Canonical Wnt/ß-Catenin Signaling Pathway,"

A. R. Sharma et al. applied biophysical, biochemical, and molecular evolutionary approaches to investigate human R-spondins protein family which is involved in cell growth and disease development and has been noticed as a potential therapeutic target. This bioinformatics study could be applied to the further investigation of Wnt/ß-catenin-System.

In "Detecting Epistatic Interactions in Metagenome-Wide Association Studies by metaBOOST," M. Wu and R. Jiang proposed a method called metaBOOST to detect epistatic interactions between such metagenomic biomarkers as microbial genus and high-level functional KEGG orthologs. They performed comprehensive simulations to evaluate metaBOOST and applied the method to analyze two real genome-wide datasets for pathological mechanisms of microbial communities in human complex diseases.

In "Predicting Flavin and Nicotinamide Adenine Dinucleotide-Binding Sites in Proteins Using the Fragment Transformation Method," C.-H. Lu et al. utilized fragment transformation methods to predict flavin and nicotinamide adenine dinucleotide-binding sites. The proposed method presents 68.4% and 67.1% true-positive rates for FAD and NAD binding site prediction under the false-positive rate at 5%, employing BioLiP dataset.

In "A Survey on the Computational Approaches to Identify Drug Targets in the Postgenomic Era," Y.-F. Dai and X.-M. Zhao made a survey on the recent progress being made on computational methodologies that have been developed to predict drug targets based on different kinds of omics data and drug properties. This information could be utilized to improve prediction accuracy when developing new methodologies in the future.

In "Identification of Gene and MicroRNA Signatures for Oral Cancer Developed from Oral Leukoplakia," G. Zhu et al. presented a new pipeline to identify oral cancer related genes and microRNAs by integrating both gene and miRNA expression profiles. They found some network modules as well as their miRNA regulators that play important roles in the development of oral leukoplakia to oral cancer. Among these network modules, 91.67% of genes and 37.5% of miRNAs have been previously reported to be related to oral cancer in literature.

In "A Heparan Sulfate-Binding Cell Penetrating Peptide for Tumor Targeting and Migration Inhibition," C.-J. Chen et al. analyzed a set of heparan sulfate-binding cell penetrating peptides derived from natural proteins. In addition to cellular binding and internalization, these peptides demonstrated multiple functions including strong binding activities to tumor cell surface, significant inhibitory effects on cancer cell migration, and suppression of angiogenesis *in vitro* and *in vivo*.

In "A Large-Scale Structural Classification of Antimicrobial Peptides," C.-C. Lee et al. presented a database of antimicrobial peptides (abbreviated as ADAM) which contains 7,007 unique sequences and 759 structures. ADAM systematically establishes comprehensive associations between AMP sequences and structures through structural folds and provides an easy access to view their relationships. Thirty distinct AMP structural fold clusters were detected and reported. According to ADAM, AMP structural folds are limited, only covering about 3% of the overall protein fold space.

In "Predict Metabolic Gene Biomarkers for Neurodegenerative Disease by an Integrated Network-Based Approach," W. Tian et al. utilized Met-express method to predict key enzyme-coding genes in both Parkinson's and Huntington diseases. They found that the predicted genes might be involved in some common pathogenic metabolic pathways and had significant functional association with known disease genes. The predicted genes could be used as novel biomarkers for potential therapeutic treatments.

In "The TF-miRNA Coregulation Network in Oral Lichen Planus," Y.-L. Zuo et al. employed the gene regulatory networks derived from transcriptomic and miRNA datasets to identify OLP related gene modules. In particular, they found that the gene modules were regulated by both transcription factors and miRNAs played important roles in the pathogenesis of OLP. Some of the genes in the modules have been reported to be related to the disease.

This special issue presents a broad spectrum of computational methodologies, biological investigation, and bioinformatics prediction. The papers included in this special issue provide useful messages of intelligent informatics for translational medicine applications. This issue illustrates that bioinformatics approaches can often be used by life scientists as a first step in the investigation of various disease mechanisms.

Acknowledgments

Here, we want to thank the authors and reviewers for their scientific contribution and congratulate them for the high quality of their work.

Hao-Teng Chang
Tatsuya Akutsu
Sorin Draghici
Oliver Ray
Tun-Wen Pai

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 731264, 9 pages
http://dx.doi.org/10.1155/2015/731264

Research Article

The TF-miRNA Coregulation Network in Oral Lichen Planus

Yu-Ling Zuo,[1] Di-Ping Gong,[2] Bi-Ze Li,[3] Juan Zhao,[4] Ling-Yue Zhou,[4] Fang-Yang Shao,[2] Zhao Jin,[5] and Yuan He[2]

[1] *Teaching Hospital of Chengdu University of Traditional Chinese Medicine, Shierqiao Road 39, Chengdu, Sichuan 610072, China*
[2] *Laboratory of Oral Biomedical Science and Translational Medicine, School of Stomatology, Tongji University, Middle Yanchang Road 399, Shanghai 200072, China*
[3] *Preclinical Medicine College, Chengdu University of Traditional Chinese Medicine, Shierqiao Road 37, Chengdu, Sichuan 610075, China*
[4] *Acupuncture and Tuina College, Chengdu University of Traditional Chinese Medicine, Shierqiao Road 37, Chengdu, Sichuan 610075, China*
[5] *College of Basic Medicine, Chengdu University of Traditional Chinese Medicine, Shierqiao Road 37, Chengdu, Sichuan 610075, China*

Correspondence should be addressed to Zhao Jin; dr.jinzhao@163.com and Yuan He; drheyuan@tongji.edu.cn

Received 8 September 2014; Revised 4 December 2014; Accepted 8 December 2014

Academic Editor: Tatsuya Akutsu

Oral lichen planus (OLP) is a chronic inflammatory disease that affects oral mucosa, some of which may finally develop into oral squamous cell carcinoma. Therefore, pinpointing the molecular mechanisms underlying the pathogenesis of OLP is important to develop efficient treatments for OLP. Recently, the accumulation of the large amount of omics data, especially transcriptome data, provides opportunities to investigate OLPs from a systematic perspective. In this paper, assuming that the OLP associated genes have functional relationships, we present a new approach to identify OLP related gene modules from gene regulatory networks. In particular, we find that the gene modules regulated by both transcription factors (TFs) and microRNAs (miRNAs) play important roles in the pathogenesis of OLP and many genes in the modules have been reported to be related to OLP in the literature.

1. Introduction

Oral lichen planus (OLP) is a chronic inflammatory disease that acts on mucous membranes inside the mouth and causes bilateral white lacy patches or plaques on the buccal mucosa, tongue, and gingivae [1]. It is found that OLP affects 0.5% to 2% of the adult population, especially the adults over 40 years old, where OLP tends to affect women rather than men with a ratio about 1.4 : 1 [2, 3]. Compared with cutaneous lichen planus, oral lichen planus lesions are more difficult to be treated with frequent recurrence. Furthermore, OLP may be at risk of developing into oral cancer as the result of carcinogenic exposures, where the erosive OLP lesions might be more sensitive to carcinogens than normal oral mucosa [1]. Currently, oral cavity cancer has become one of the 10 most frequently diagnosed cancers with increasing mortality in East Europe [4, 5]. However, the pathogenesis of OLP and how it is developed into oral cancer is still unclear [6, 7]. Therefore, it is extremely urgent to pinpoint the molecular mechanisms underlying the pathogenesis of OLP so that accurate diagnosis can be made and effective therapies can be developed.

Recently, the accumulation of large amount of omics data, especially transcriptome data, provides opportunities to identify the molecules related to diseases. Accordingly, many works have investigated OLPs with transcriptome data. For example, the genes CCR5, CD14, and beta-catenin have been identified to play important roles in the pathogenesis of OLP [8, 9]. Moreover, Tao et al. identified some genes that are differentially expressed in OLPs, such as FOXP3, ANGPT1, and MMP1, and these genes may be related to the development of OLP [10]. In general, the above-mentioned studies assume

those differentially expressed genes between OLPs and controls are related to OLP. However, the differentially expressed genes are usually treated independently, which is actually not the case. It has been found that complex diseases, for example, cancers, happen due to the dysregulation of functional gene sets or molecular pathways [11, 12]. In other words, the genes involved in the same disease tend to have functional relationships. Therefore, it is necessary to investigate disease related genes from a systematic perspective. Except for the above-mentioned genes, some small noncoding RNAs, that is, microRNAs (miRNAs), were found to play important roles in cancer by targeting oncogenes or tumor suppressor genes [13]. For example, mir-21 was found to be overexpressed in several tumor types [14], and let-7 inhibits lung tumorigenesis by repressing the expression of the RAS oncogene [15]. More recently, Gassling et al. found that the dysregulation of some miRNAs has important pathophysiological impacts on OLP [16]. For instance, mir-21, mir-181b, and mir-345 were found to be upregulated in OLPs and have critical roles in the malignant transformation of OLP to oral cancer.

In this work, we present a novel approach to identify gene modules that may play important roles in the pathogenesis of OLP by assuming that OLP is caused due to the dysregulation of certain gene modules. Furthermore, based on the gene modules as well as their transcription factor and miRNA regulators, we construct a TF-miRNA coregulation network. By investigating the genes and their regulators in the coregulation network, we find that some of them have already been reported to be related to OLP or oral cancer, indicating the important roles of the regulation network in OLP. In addition, we notice that the genes involved in the regulation network can serve as disease associated pattern and separate OLPs from controls very well, which is also validated by another independent real dataset, demonstrating the potential of the gene modules we identified as disease associated pattern and therapeutic targets.

2. Materials and Methods

2.1. Gene and miRNA Expression Data. The matched gene and miRNA expression data were obtained from the Gene Expression Omnibus (GEO) database [17]. Both mRNA (accession number: GSE38616) and miRNA (accession number: GSE38615) expression profiles were measured in 7 healthy individuals and 7 oral lichen planus patients [16]. To further validate the genes identified from the above datasets, another gene expression dataset (accession number: GSE52130) was retrieved from GEO, which was originally measured in 14 oral samples and 9 genital epithelium samples. Here, the gene expression profiles from the 14 oral samples consist of 7 normal oral samples and 7 oral lichen planus samples were kept for validation. All the three expression datasets have been preprocessed and normalized when we downloaded them, and the data were used in later sections without further preprocessing.

2.2. Identification of Differentially Expressed Genes and miRNAs. In general, the genes that are differentially expressed between diseases and controls are related to diseases to some extent. In this work, the genes that are differentially expressed between OLPs and controls were detected with Student's t-test. The genes with P values less than 0.05 were regarded to be differentially expressed genes (DEGs), and the same for miRNAs. Consequently, 2587 differentially expressed genes (GSE38616) and 90 differentially expressed miRNAs (GSE38615) were obtained for further analysis.

2.3. Identification of Network Modules Associated with OLP. A gene coexpression network was constructed for OLPs based on their corresponding gene expression data, where one gene was linked to another if their coexpression measured with Pearson's correlation coefficient was significantly high (P value cutoff of 0.05) and the weights accompanying the edges were their corresponding correlation coefficients. Subsequently, network modules that consist of densely connected genes were detected with ClusterONE [18], and 154 network modules were detected here. Furthermore, after investigating the network modules, we merged two modules if more than one-third of genes from the smaller module occur in the larger one. As a result, 125 modules were kept for further analysis. For each network module, it will be regarded to be related to OLP if the module is enriched with the above identified differentially expressed genes, where the enrichment analysis was performed with Fisher's exact test (P value cutoff of 0.01).

2.4. miRNA-Gene Regulations. The target genes of miRNAs were collected from both predictions and experimentally determined ones. For the predictions, several tools, including PicTar [19], miRanda [20], MicroT [21], and TargetScan [22], were employed to predict the target genes of miRNAs. Specifically, we picked up the interactions between genes and miRNAs predicted by at least two tools to avoid false positives. Moreover, the target information of miRNAs deposited in Tarbase [23] was obtained and merged with the predictions, where all the miRNA-gene interactions from Tarbase have been experimentally validated.

2.5. The TF-miRNA Coregulation Network in OLP. After obtaining the network modules, we first checked which miRNAs may regulate the network modules. Given a network module and a miRNA, the miRNA will regulate the module if its target genes are enriched in the modules with Fisher's exact test (P value cutoff of 0.01). In particular, we only considered the differentially expressed miRNAs (DemiRs) here since these DemiRs are more likely related to OLP. Furthermore, given a network module, the transcription factors (TFs) that possibly regulate the modules were identified if these TFs belong to the module and coexpress with other genes in the module. Note that here we suppose the TFs that coexpress with other genes in the module will regulate the genes within the module. Consequently, we detected 6 network modules that are coregulated by TFs and miRNAs, where the modules are enriched with DEGs. We assumed that the TF-miRNA

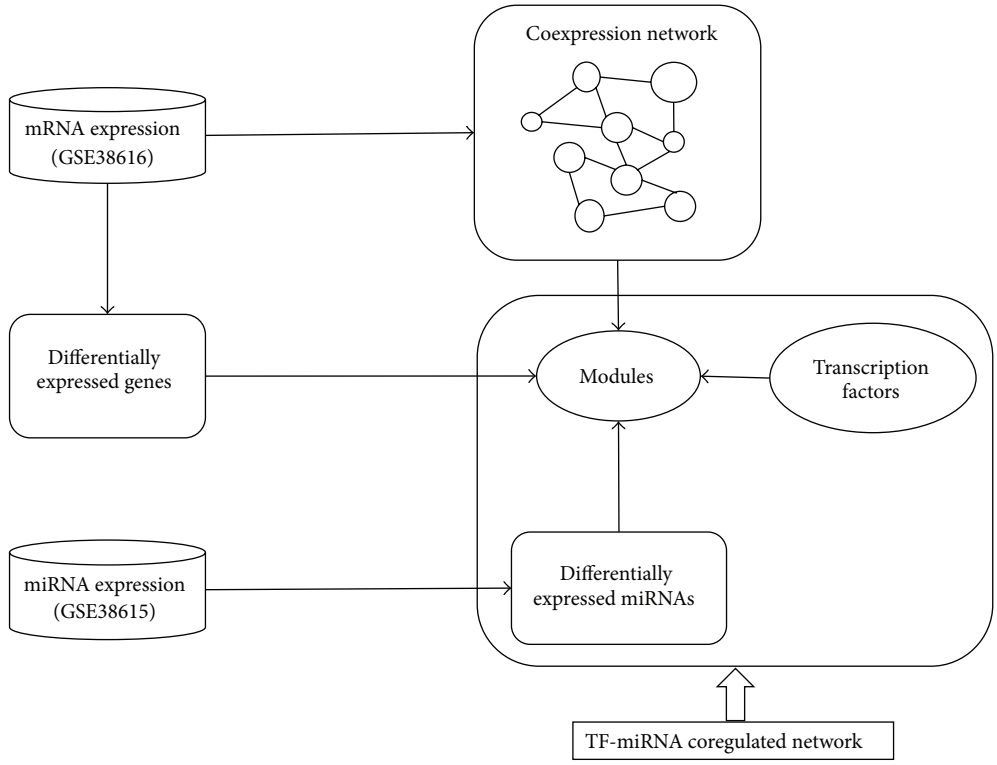


Figure 1: Schematic illustration of the pipeline to detect the TF-miRNA coregulation networks in oral lichen planus.

Table 1: The detailed information about the 6 network modules as well as their TF and miRNA regulators.

Module	#Nodes	#Edges	miRNAs	Transcription factors
Module 1	134	952	hsa-miR-628-5p	SALL2, PEG3, ZNF865, HES3, ZNF283, ZIM2, FOSB, PAX6, SIX2, ZNF616, KDM5D, SRY, RFX4, and ZFY
Module 2	62	594	hsa-miR-595	ASCL4
Module 3	160	1301	hsa-miR-34c-5p hsa-miR-34a hsa-miR-26b	ELK3, RFX8, MIXL1, HIF1A, and ZNF552
Module 4	80	405	hsa-miR-29a	LIN28A
Module 5	55	335	hsa-miR-190	THRA, NR1D1, and NR1D2
Module 6	43	241	hsa-miR-146b-5p	ZNF626

coregulation network which consists of the 6 modules plays important roles in the development of OLP.

3. Results and Discussion

Figure 1 depicts the flowchart of our proposed approach for identifying the TF-miRNA coregulation network in OLP, and we applied it to a real dataset which contains 7 healthy individuals and 7 OLP patients, where the dataset contains the matched gene (GSE38616) and miRNA (GSE38615) expression profiles. From the dataset, we detected 2587 differentially expressed genes (DEGs) and 90 differentially expressed miRNAs (DemiRs). We further constructed a gene coexpression network and identified modules that were coregulated by miRNAs and TFs (Figure 2). Table 1 lists the detailed information about the 6 network modules as well as their regulators.

3.1. The Network Modules Are Enriched with Oral Lichen Planus Related Genes. After getting the 6 network modules, we first investigated the genes in each module. By querying the PubMed, we found that 59 out of the 497 genes belonging to the 6 modules have been reported to be relevant to oral cancer. Here, for each network module, we gave examples

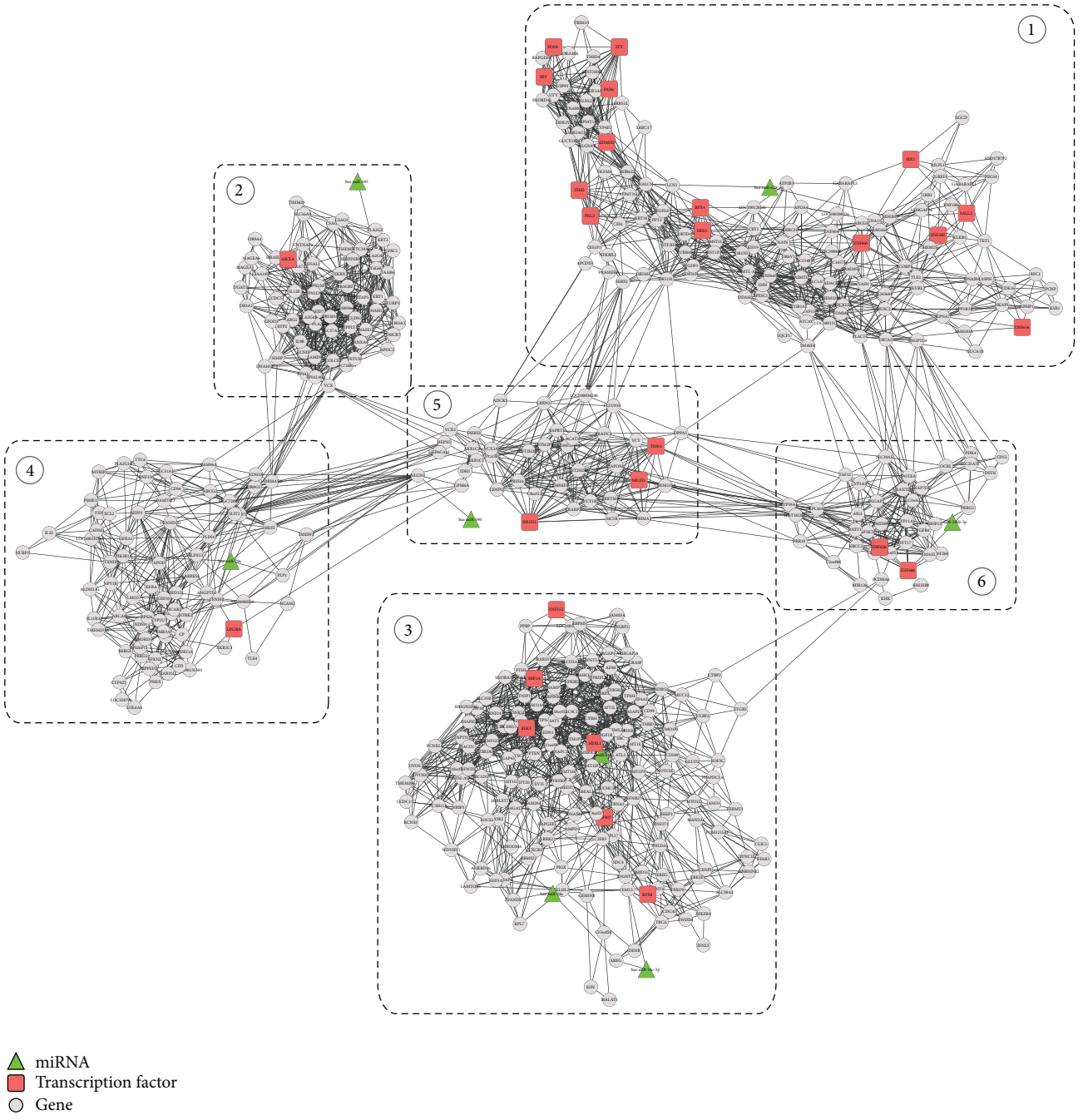


Figure 2: The TF-miRNA coregulation network with 6 modules, where the green nodes represent differentially expressed miRNAs, red nodes denote the transcription factors, and gray nodes represent genes in modules, respectively.

about genes that have been reported to be related to OLP or oral cancer in the literature (Table 2). For example, it was reported that the gene KRT18 was related to tumor differentiation and metastasis and plays important roles in the malignant transformation of OLP to oral squamous cell carcinoma [24]. Another gene PTEN was reported to be downregulated in oral squamous cells, which in turn downregulates the expression of cyclin D1 and leads to the suppression of cell growth, indicating that targeting PTEN may help treat oral cancer [25]. Moreover, IGF1R has been reported to control cell proliferation of oral cancer [26, 27]. The overlap between known oral cancer associated genes and our identified module genes indicates that the genes belonging to these modules are related to OLP as well as its development to oral cancer.

Next, we investigated the functions of the network module genes. For each module, functional and pathway enrichment analyses were performed with DAVID [28], and the

Table 2: Examples of genes from each network module that have been reported to be related to OLP or oral cancer.

Network module	Gene symbol	PubMed IDs
Module 1	KRT18	19575986; 22677743; 7527618;
	SHH	11857543; 21945071; 21496886
	FOSB	19653276; 15926923
Module 2	MAGEA3	19187853; 12855658
	KRT1	20002980; 16334838; 10896780
	MAGEA6	18197853;
Module 3	HIF1A	19717330; 19449077; 18630523
	PTEN	17067457; 15453811; 15805158
	IGF1R	17786320; 23106397; 19584075
Module 4	CP	23812204; 19884712; 17066447
	ABCG2	18429968; 15801936; 15618737
	ALDH1A1	22725270; 22782852; 21441790
Module 5	IDH1	22385606; 21383741; 19378339
	CRABP2	19197536; 16568407; 11437413
Module 6	CD55	21545652; 17234541; 15668483
	FGFR4	18487077; 20127014; 23481570
	CYP3A5	16338276; 18628519; 1808564

detailed results can be found in Table 3, where only the processes we thought related to OLP were listed for clarification. From the analysis, we can see the biological processes in which the modules involved are related to the initiation and development of OLP. For example, it was reported recently that the pathogenesis of OLP was associated with some systemic diseases that can cause midbrain injuries [29]. The inhibition of phospholipase A2 activity that is associated with numerous inflammatory processes was found to be related to the mechanism of OLP [30]. The sensory perception, such as anxiety and tension, has been reported to be an important factor in the development of OLP [31]. It was reported that oral lichen planus can be caused by a variety of stimuli and the preservation of keratin in oral mucosa was an efficient way for the treatment of the disease [32]. Compared with normal controls, the OLPs tend to have increased angiogenesis, indicating OLP is associated with the induction of aberrant angiogenesis [7]. In addition, the symptoms of OLP are always accompanied by compromised wound healing [33], and the epidermal growth factor receptors were found to be significantly higher in OLPs [34].

Except for biological processes, functional enrichments analysis implies that some molecular functions, such as cytosol, steroid hydroxylase activity, and oxidoreductase activity, also have important impacts on OLPs. Considering that OLP is often treated with steroids and vitamin A analogues [35], it is not surprising that steroid hormone receptor activity, retinoid metabolic process, and vitamin A metabolic process are enriched in our identified modules. Moreover, the metabolism of xenobiotics by cytochrome P450 has been reported to result in the oral and pharyngeal cancers [36]. In addition, it was found that the metallic ion content can increase the damage to the oral mucosa cells [37], which is consistent with the enrichment of the iron ion homeostasis and binding.

From the analysis of the genes belonging to our identified modules, we can see that these modules are indeed related to the development of OLP. In addition, we identified some important biological processes that have important roles in the development of OLP, such as the metabolism of xenobiotics by cytochrome P450 and vitamin A metabolic process. The detailed information about the biological processes in which the TFs and miRNAs are involved can be found in Supplementary Table I available online at http://dx.doi.org/10.1155/2015/731264.

3.2. miRNAs Regulators of Network Modules Are Associated with OLP. In the TF-miRNA regulation network, there are in total 8 miRNAs, which were picked up from the 90 differentially expressed miRNAs. Among the 8 miRNAs, some of them have been reported to be related to OLP or oral cancer in the literature. For example, miR-26b was found to be significantly low expressed in OLP lesions compared with controls [38], miR-29a was remarkably differentially expressed in the oral squamous cell carcinoma metastasis [39], and miR-628 was able to discriminate hand-foot-mouth diseases from healthy controls [40]. According to the Human microRNA Disease Database (HMDD) [41], a manually curated disease-miRNA association database, mir-146b was reported to be associated with diverse neoplasms including oral cancer. In addition, two of the 8 DemiRs, that is, hsa-miR-146b-5p and hsa-miR-26b, have been reported previously to be related to OLP [16].

Furthermore, we derived the interactions between miRNAs and target genes from the 6 modules in TF-miRNA coregulation network. Figure 3 shows the regulation network composed of miRNAs and target genes. By investigating the expressions of miRNAs and their target genes, we noticed that the expressions of 4 miRNAs, hsa-miR-190, hsa-miR-146b-5p, hsa-miR-29a, and hsa-miR-595, were negatively correlated with that of their target genes, which is consistent with the observation that miRNAs generally repress the expression of their target genes. Interestingly, these four miRNAs were highly expressed in OLP while their target genes were downregulated. The detailed information about the expression levels of miRNAs and their target genes in modules can be found in Supplementary Table II.

3.3. Transcription Factors Regulating the Network Modules Are Associated with OLP. We also investigated the 27 transcription factors involved in the TF-miRNA coregulation network. By querying the PubMed, some TFs were found already reported to be related to OLP or oral cancer. For example, in module 3, the transcription factor HIF1A is a master transcriptional regulator of the adaptive response to hypoxia. It was found that RTP801 and VEGF, the target genes of HIF1A, were significantly low expressed in OLPs [42]. The transcription factor LIN28A from module 4 has been reported to regulate cancer stem cell-like properties and can act as an appropriate target for oral squamous cell carcinoma treatment [43]. In module 1, AX6 was found to

Table 3: Functional enrichment analysis of genes in network modules.

Network module	Enriched functions related to OLP	P value (<0.05)
Module 1	GO:0030901~midbrain development	0.00337086
	GO:0016702~oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen	0.00560592
	GO:0008285~negative regulation of cell proliferation	0.01400315
	GO:0007600~sensory perception	0.032006
	GO:0007435~salivary gland morphogenesis	0.0376765
	GO:0002052~positive regulation of neuroblast proliferation	0.04294326
Module 2	GO:0004623~phospholipase A2 activity	0.00233
	GO:0050877~neurological system process	0.00764739
	GO:0051606~detection of stimulus	0.04745557
Module 3	GO:0001525~angiogenesis	0.00589117
	GO:0005829~cytosol	0.00679876
	GO:0042060~wound healing	0.01651834
	GO:0045740~positive regulation of DNA replication	0.01978034
	GO:0007173~epidermal growth factor receptor signaling pathway	0.01978034
	GO:0015629~actin cytoskeleton	0.04636455
Module 4	GO:0006879~cellular iron ion homeostasis	0.00604241
	GO:0000041~transition metal ion transport	0.029349054
	GO:0055114~oxidation reduction	0.03217
	hsa00980:Metabolism of xenobiotics by cytochrome P450	0.03978611
Module 5	GO:0003707~steroid hormone receptor activity	0.00495459
	GO:0006766~vitamin metabolic process	0.01044366
	GO:0001523~retinoid metabolic process	0.04657368
	GO:0006776~vitamin A metabolic process	0.04657368
Module 6	GO:0008395~steroid hydroxylase activity	0.02851299
	GO:0005887~integral to plasma membrane	0.02900047
	GO:0008202~steroid metabolic process	0.04570667

regulate the proliferation and apoptosis processes in human retinoblastoma cells [44]. In module 2, ASCL4 was found to be essential for the determination of cell fate as well as the development and differentiation of numerous tissues [45].

In addition, we investigated the top 25 biological processes regulated by these TFs as shown in Supplementary Figure 1, where the percentage denotes the fraction of all TFs from the TF-miRNA coregulation network that were involved in the corresponding process. Consistent with the above observations, the TFs identified here are involved in a lot of OLP related processes, such as cell differentiation, Notch signaling pathway, steroid hormone mediated signaling pathway, and wound healing.

The analysis of TFs involved in TF-miRNA coregulation network indicates that these TFs regulate OLP related biological processes and play important roles in promoting the progression and development of OLP.

4. Conclusion

The potential malignant transformation of oral lichen planus (OLP) to oral cancer makes it demanding to understand the pathogenesis of this disease. In this paper, we introduced a novel approach to identify the TF-miRNA coregulation network that plays important roles in OLP. Unlike traditional approaches, the regulatory circuit we detected here provides new insights into observing disease associated patterns. The overlap between known OLP associated genes and our identified module genes implies that these gene modules are significantly related to OLP. The discriminative capacity of these modules in separating OLPs from controls confirms again the important roles of these modules in OLP and their potential as disease associated pattern. In addition, the regulators of these gene modules, including transcription factors and miRNAs, were also found to play important roles

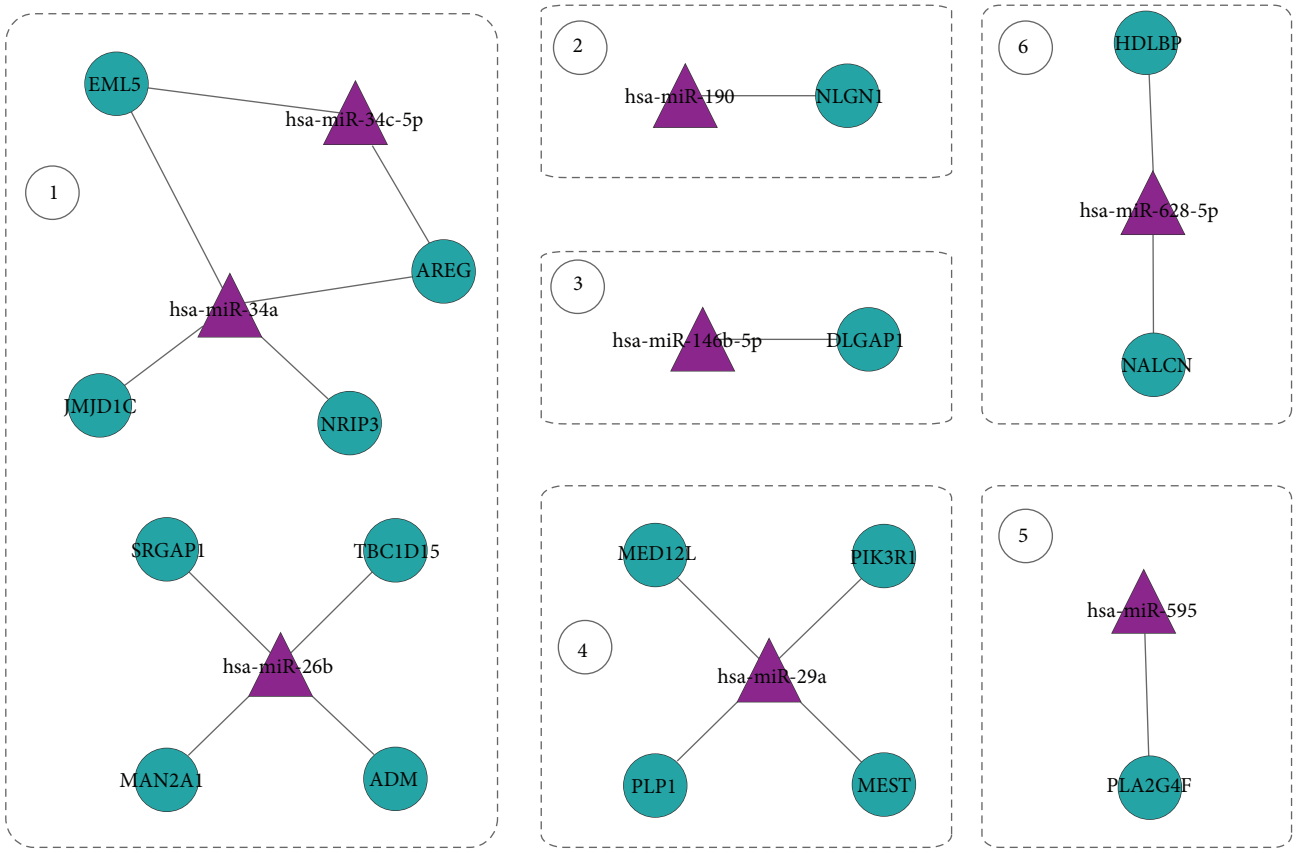


Figure 3: The regulatory relationship between miRNA and their targets in our regulatory network.

in the manifestation and progression of OLP, indicating their potential as new therapeutic targets when treating OLPs.

Conflict of Interests

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

Acknowledgments

This work was partly supported by the guided project from Shanghai Science and Technology Commission (09411965900), the Natural Science Foundation of Shanghai Science and Technology Commission (14ZR1443600), the guided project from Shanghai Science and Technology Commission (14411971800), the Science Foundation of Shanghai Health Bureau (201440274), and the guided project from Science and Technology Department of Sichuan Province (2011SZ0300).

References

[1] S. Silverman Jr., "Oral lichen planus: a potentially premalignant lesion," *Journal of Oral and Maxillofacial Surgery*, vol. 58, no. 11, pp. 1286–1288, 2000.

[2] F. A. de Sousa, T. C. Paradella, A. A. H. Brandão, and L. E. B. Rosa, "Oral lichen planus versus epithelial dysplasia: difficulties in diagnosis," *Brazilian Journal of Otorhinolaryngology*, vol. 75, no. 5, pp. 716–720, 2009.

[3] P. B. Sugerman and N. W. Savage, "Oral lichen planus: causes, diagnosis and management," *Australian Dental Journal*, vol. 47, no. 4, pp. 290–297, 2002.

[4] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.

[5] T. Zheng, P. Boyle, H. Hu et al., "Dentition, oral hygiene, and risk of oral cancer: a case-control study in Beijing, People's Republic of China," *Cancer Causes and Control*, vol. 1, no. 3, pp. 235–241, 1990.

[6] M. Gorsky, J. B. Epstein, H. Hasson-Kanfi, and E. Kaufman, "Smoking habits among patients diagnosed with oral lichen planus," *Tobacco Induced Diseases*, vol. 2, no. 2, pp. 103–108, 2004.

[7] N. Mittal, G. S. Madhu Shankari, and S. Palaskar, "Role of angiogenesis in the pathogenesis of oral lichen planus," *Journal of Oral and Maxillofacial Pathology*, vol. 16, no. 1, pp. 45–48, 2012.

[8] M. Srinivasan, K. N. Kodumudi, and S. L. Zunt, "Soluble CD14 and toll-like receptor-2 are potential salivary biomarkers for oral lichen planus and burning mouth syndrome," *Clinical Immunology*, vol. 126, no. 1, pp. 31–37, 2008.

[9] M. Ebrahimi, L. Boldrup, Y.-B. Wahlin, P. J. Coates, and K. Nylander, "Decreased expression of the p63 related proteins β-catenin, E-cadherin and EGFR in oral lichen planus," *Oral Oncology*, vol. 44, no. 7, pp. 634–638, 2008.

[10] X. A. Tao, C. Y. Li, J. Xia et al., "Differential gene expression profiles of whole lesions from patients with oral lichen planus,"

Journal of Oral Pathology and Medicine, vol. 38, no. 5, pp. 427–433, 2009.

[11] X. Liu, Z.-P. Liu, X.-M. Zhao, and L. Chen, "Identifying disease genes and module biomarkers by differential interactions," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 241–248, 2012.

[12] K.-Q. Liu, Z.-P. Liu, J.-K. Hao, L. Chen, and X.-M. Zhao, "Identifying dysregulated pathways in cancers from pathway interaction networks," *BMC Bioinformatics*, vol. 13, no. 1, article 126, 2012.

[13] G. A. Calin, C. Sevignani, C. D. Dumitru et al., "Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2999–3004, 2004.

[14] J. A. Chan, A. M. Krichevsky, and K. S. Kosik, "MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells," *Cancer Research*, vol. 65, no. 14, pp. 6029–6033, 2005.

[15] S. M. Johnson, H. Grosshans, J. Shingara et al., "RAS is regulated by the let-7 microRNA family," *Cell*, vol. 120, no. 5, pp. 635–647, 2005.

[16] V. Gassling, J. Hampe, Y. Açil, J. H. Braesen, J. Wiltfang, and R. Häsler, "Disease-associated miRNA-mRNA networks in oral lichen planus," *PLoS ONE*, vol. 8, no. 5, Article ID e63015, 2013.

[17] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[18] T. Nepusz, H. Yu, and A. Paccanaro, "Detecting overlapping protein complexes in protein-protein interaction networks," *Nature Methods*, vol. 9, no. 5, pp. 471–472, 2012.

[19] A. Krek, D. Grün, M. N. Poy et al., "Combinatorial microRNA target predictions," *Nature Genetics*, vol. 37, no. 5, pp. 495–500, 2005.

[20] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, "Human microRNA targets," *PLoS Biology*, vol. 2, no. 11, article e363, 2004.

[21] M. Maragkakis, P. Alexiou, G. L. Papadopoulos et al., "Accurate microRNA target prediction correlates with protein repression levels," *BMC Bioinformatics*, vol. 10, article 295, 2009.

[22] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[23] P. Sethupathy, B. Corda, and A. G. Hatzigeorgiou, "TarBase: a comprehensive database of experimentally supported animal microRNA targets," *RNA*, vol. 12, no. 2, pp. 192–197, 2006.

[24] L. Su, P. R. Morgan, and E. B. Lane, "Protein and mRNA expression of simple epithelial keratins in normal, dysplastic, and malignant oral epithelia," *The American Journal of Pathology*, vol. 145, no. 6, pp. 1349–1357, 1994.

[25] S. M. Xie, L. J. Shen, C. Yin, P. Ruan, and X. Yao, "Expression of tumor suppressor gene PTEN, PIP3 and cyclin D1 in oral squamous cell carcinoma and their correlations," *Zhonghua Kou Qiang Yi Xue Za Zhi*, vol. 41, no. 7, pp. 407–410, 2006.

[26] G. Brady, S. Crean, P. Naik, and S. Kapas, "Upregulation of IGF-2 and IGF-1 receptor expression in oral cancer cell lines," *International Journal of Oncology*, vol. 31, no. 4, pp. 875–881, 2007.

[27] M. J. Jameson, L. E. Taniguchi, K. K. Vankoevering et al., "Activation of the insulin-like growth factor-1 receptor alters p27 regulation by the epidermal growth factor receptor in oral squamous carcinoma cells," *Journal of Oral Pathology & Medicine*, vol. 42, no. 4, pp. 332–338, 2013.

[28] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.

[29] E. H. van der Meij, H. Mast, and I. van der Waal, "The possible premalignant character of oral lichen planus and oral lichenoid lesions: a prospective five-year follow-up study of 192 patients," *Oral Oncology*, vol. 43, no. 8, pp. 742–748, 2007.

[30] J. Vibha, K. Choudhary, M. Singh, M. Rathore, and N. Shekhawat, "A study on pharmacokinetics and therapeutic efficacy of Glycyrrhiza glabra a miracle medicinal herb," *Botany Research International*, vol. 2, pp. 157–163, 2009.

[31] C. M. Allen, F. M. Beck, K. M. Rossie, and T. J. Kaul, "Relation of stress and anxiety to oral lichen planus," *Oral Surgery, Oral Medicine, Oral Pathology*, vol. 61, no. 1, pp. 44–46, 1986.

[32] G. R. Ogden, A. Nairn, A. Carmichael et al., "Preservation of keratin expression in oral mucosa using a novel transport medium," *Journal of Oral Pathology and Medicine*, vol. 21, no. 1, pp. 17–20, 1992.

[33] S. Guo and L. A. Dipietro, "Factors affecting wound healing," *Journal of Dental Research*, vol. 89, no. 3, pp. 219–229, 2010.

[34] H. Kobayashi, K. Kumagai, A. Gotoh et al., "Upregulation of epidermal growth factor receptor 4 in oral leukoplakia," *International Journal of Oral Science*, vol. 5, no. 1, pp. 14–20, 2013.

[35] S. H. Günther, "Vitamin A acid in treatment of oral lichen planus," *Archives of Dermatology*, vol. 107, no. 2, p. 277, 1973.

[36] C.-Y. Lin, T.-S. Pan, C.-C. Ting et al., "Cytochrome P450 metabolism of betel quid-derived compounds: implications for the development of prevention strategies for oral and pharyngeal cancers," *The Scientific World Journal*, vol. 2013, Article ID 618032, 11 pages, 2013.

[37] E. Fernández-Miñano, C. Ortiz, A. Vicente, J. L. Calvo, and A. J. Ortiz, "Metallic ion content and damage to the DNA in oral mucosa cells of children with fixed orthodontic appliances," *BioMetals*, vol. 24, no. 5, pp. 935–941, 2011.

[38] K. Danielsson, M. Ebrahimi, Y. B. Wahlin, K. Nylander, and L. Boldrup, "Increased levels of COX-2 in oral lichen planus supports an autoimmune cause of the disease," *Journal of the European Academy of Dermatology and Venereology*, vol. 26, no. 11, pp. 1415–1419, 2012.

[39] N. A. Serrano, C. Xu, Y. Liu et al., "Integrative analysis in oral squamous cell carcinoma reveals DNA copy number-associated miRNAs dysregulating target genes," *Otolaryngology—Head and Neck Surgery*, vol. 147, no. 3, pp. 501–508, 2012.

[40] L. Cui, Y. Qi, H. Li et al., "Serum microRNA expression profile distinguishes enterovirus 71 and coxsackievirus 16 infections in patients with hand-foot-and-mouth disease," *PLoS ONE*, vol. 6, no. 11, Article ID e27071, 2011.

[41] Y. Li, C. Qiu, J. Tu et al., "HMDD v2.0: a database for experimentally supported human microRNA and disease associations," *Nucleic Acids Research*, vol. 42, no. 1, pp. D1070–D1074, 2014.

[42] M. Ding, J. Y. Xu, and Y. Fan, "Altered expression of mRNA for HIF-1α and its target genes RTP801and VEGF in patients with oral lichen planus," *Oral Diseases*, vol. 16, no. 3, pp. 299–304, 2010.

[43] T. Wu, J. Jia, X. Xiong et al., "Increased expression of Lin28B associates with poor prognosis in patients with oral squamous cell carcinoma," *PLoS ONE*, vol. 8, no. 12, Article ID e83869, 2013.

[44] S. W. Bai, B. Li, H. Zhang et al., "Pax6 regulates proliferation and apoptosis of human retinoblastoma cells," *Investigative Ophthalmology and Visual Science*, vol. 52, no. 7, pp. 4560–4570, 2011.

[45] M. Jonsson, E. Björntorp Mark, C. Brantsing, J. M. Brandner, A. Lindahl, and J. Asp, "*Hash4*, a novel human achaete-scute homologue found in fetal skin," *Genomics*, vol. 84, no. 5, pp. 859–866, 2004.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 432012, 9 pages
http://dx.doi.org/10.1155/2015/432012

Research Article

Prediction of Metabolic Gene Biomarkers for Neurodegenerative Disease by an Integrated Network-Based Approach

Qi Ni,[1] Xianming Su,[2] Jingqi Chen,[1] and Weidong Tian[1]

[1]*State Key Laboratory of Genetic Engineering, Department of Biostatistics and Computational Biology, School of Life Sciences, Fudan University, Shanghai 200438, China*
[2]*Department of Geriatric Cardiology, The First Affiliated Hospital of Xi'an Jiao Tong University, Xi'an, Shaanxi 710000, China*

Correspondence should be addressed to Weidong Tian; weidong.tian@fudan.edu.cn

Received 7 September 2014; Accepted 27 November 2014

Academic Editor: Hao-Teng Chang

Neurodegenerative diseases (NDs), such as Parkinson's disease (PD) and Huntington's disease (HD), have become more and more common among aged people worldwide. One hallmark of NDs is the presence of intracellular accumulation of specific pathogenic proteins that may result from abnormal function of metabolic processes. Previously, we have developed a computational method named Met-express that predicted key enzyme-coding genes in cancer development by integrating cancer gene coexpression network with the metabolic network. Here, we applied Met-express to predict key enzyme-coding genes in both PD and HD. Functional enrichment analysis and literature review of predicted genes suggested that there might be some common pathogenic metabolic pathways for PD and HD. We further found that the predicted genes had significant functional association with known disease genes, with some of them already documented as biomarkers or therapeutic targets for NDs. As such, the predicted metabolic genes may be of use as novel biomarkers not only for ND diagnosis but also for potential therapeutic treatments.

1. Introduction

Neurodegenerative diseases (NDs), such as Parkinson's disease (PD) and Huntington's disease (HD), occur as a result of progressive loss of structure or function of neurons [1]. Onset of NDs is generally for people at old age, but symptoms can appear at any age. ND patients often have difficulties with motor or cognition. Currently, PD affects about 0.16% of the general population and up to 2% among people above 80 years old. Symptoms of PD conventionally include slowness of movement, unrest shaking, rigidity, and postural instability [2]. PD can also cause ruddy and oily skin, depression mood, and dysautonomia problems [3]. The prevalence of HD is about 5 affected individuals per 100 000 population, with a higher incidence rate in white population compared with Asia or African people [4]. Individuals with HD usually show distinct chorea and motor impersistence and may have psychiatric disturbances [4, 5]. Most patients start presenting cognitive decline at early stage resulting in dementia in late stage [5]. Both of these disorders bring serious disability to the patients and cause significant burden to their families. Many researchers have worked on therapies for NDs for a long time. However, the precise pathologies of NDs are still poorly understood and effective therapies are still needed to be explored.

The pathogenesis of NDs is usually associated with several interacting pathways, which finally coincide into a common pathway of apoptosis [6]. Changes in membrane phospholipid molecules were implicated in cell death [7]. Mitochondrial dysfunction, one of the most studied pathways, can result in oxidative stress and energy supply abnormality for neuronal cells [8]. Immune pathways, such as inflammation progress, lead to oxidative stress and cause apoptosis through more direct mechanisms [9]. In addition, accumulation of abnormal proteins is a typical symptom of many NDs and has been thought to be a key factor in the pathogenesis for a long time. Each disorder has specific protein aggregates as its hallmark: α-synuclein in the substantia nigra of PD and huntingtin in the striatum of HD [1]. Abnormal protein accumulation can result from misfolded proteins of the synthesis

pathways or toxic metabolites associating with the protein synthesis pathways. Genetic mutations explained some familial form cases, but many sporadic cases were caused by unclear reasons [10]. Mutations of protein coding genes *PARK2* and *HTT* are culprits of PD and HD [1], respectively. The expansion of trinucleotide CAG repeats coding for polyglutamine in the N terminus of the huntingtin protein can cause HD. When the CAG repeats reach 36, the disease is penetrant [4]. Oxidation products of proteins, such as advanced glycation end-products (AGEs) and the deposition of AGEs-cross-linked proteins, exist in many NDs [11]. Oxidative modifications, nitration, and phosphorylation of proteins were also suggested to be involved in the aggregation of pathogenic proteins [1]. These discoveries lead to the common hypothesis that the misfolding and accumulation of specific proteins are pivotal to ND pathology.

The accumulation of misfolded proteins in ND patients is associated with abnormal metabolic conditions [11–13]. Detection of enzyme-coding genes that are significantly differentially expressed and that may drive the metabolic changes in ND patients may therefore help understand the mechanisms of ND development. In a previous study, we have developed a computational method named Met-express that integrated cancer gene coexpression network and metabolic network to predict key enzyme-coding genes in cancer development [14]. In Met-express, a coexpression network of cancer gene was first constructed, from which cancer-specific gene coexpression modules were determined. Then, within a cancer-specific gene coexpression module, Met-express was aimed at identifying the enzyme-coding genes that were coexpressed with significantly more metabolite-sharing enzyme-coding genes. The hypothesis was that these genes likely bear a higher potential to alter the metabolism of cancer cells and therefore might serve as anticancer drug targets. We demonstrated with both literature search and real experiments to show that the predicted enzyme-coding genes by Met-express in cancer were critical for cancer development. As a general method, Met-express can be extended for predicting key enzyme-coding genes in other diseases, such as NDs. Thus, we applied Met-express to both PD and HD and have predicted a number of key enzyme-coding genes for the diseases. Functional enrichment analysis and literature validation of predicted genes suggested that there might be some common pathogenic metabolic pathways for PD and HD. In addition, predicted genes were found to have significant functional associations with known ND genes. As such, these genes may not only help to understand the mechanisms of ND pathogenesis but also be used as biomarkers for ND diagnosis and for developing novel therapeutic strategies.

2. Materials and Methods

2.1. Microarray Data Preparation. All microarray data were downloaded from GEO database (http://www.ncbi.nlm.nih.gov/geo/). For PD, GDS files were downloaded only if the dataset (i) contained samples from brain tissues of both normal people and patients, (ii) had more than 10 samples, and (iii) had not been treated with any stimulus. As for HD, GDS files were also following the three standards except that they were from blood tissues. These resulted in four GDS files of PD and three of HD. As two files of PD were from the same samples and tested on similar microarrays, we combined the two files to one file to extend the data size (Supplementary Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/432012). For each GDS file, we selected only genes with a median ratio of null values less than 80% and used R (http://www.r-project.org/) to normalize gene expression values as previously described [14].

2.2. Met-Express Procedure. We followed the procedures as Chen et al. described [14] to detect the key enzyme-coding genes of PD and HD. Here, we briefly describe the workflow of Met-express. (1) We calculated Pearson correlation coefficients (PCC) based on expression levels for all pairs of genes and used a network partition algorithm named Qcut to divide gene coexpression network into gene coexpression modules, retaining modules containing more than 10 genes. Then, we evaluated the specificity of each coexpression module to the disease conditions by computing the AUC of a ROC curve plotted using the median expression value of the genes inside the module. (2) We extracted the information of reactions, corresponding enzyme-coding genes, and associated metabolites from the KEGG Markup Language files (http://www.genome.jp/kegg/). Two enzymes were linked if the substrate metabolite of one enzyme was the product metabolite of the other. The direction was defined as starting from the enzyme generating the metabolite to the enzyme consuming it. Following this rule, we constructed the metabolic network by linking enzymes, when the corresponding metabolite participated in fewer than 18 reactions and not in "Xenobiotics Biodegradation and Metabolism." In the construction process, a reversible reaction would be split into two reactions. The above procedures resulted in 860 enzyme-coding genes with outward edges, which were considered to have large impacts on the metabolic network. (3) Finally, we computed the importance score of each gene by considering both specificity of the gene coexpression module to the disease and enrichment of the degree of metabolic links of the gene inside the module. For each dataset, we used the median importance score as the cutoff to select the candidate genes. Then, those genes predicted in at least 2 datasets were taken as the key enzyme-coding genes for the disease.

2.3. Functional Enrichment Analysis. Gene ontology (GO) enrichment analysis and KEGG pathway enrichment analysis were both conducted by Fisher test using fisher.test in R. P value was adjusted by false discovery rate. All the enzyme-coding genes in the constructed metabolic network were used as the background.

2.4. Functional Association Test. To test the connection between our predicted key enzyme-coding genes and known disease genes in database, we calculated the shortest path lengths between each pair of genes in a functional association network, specifically the FunCoup [15] network. Here, the shortest paths were calculated by Dijkstra's algorithm [16]. For each identified enzyme-coding gene of each disease, we

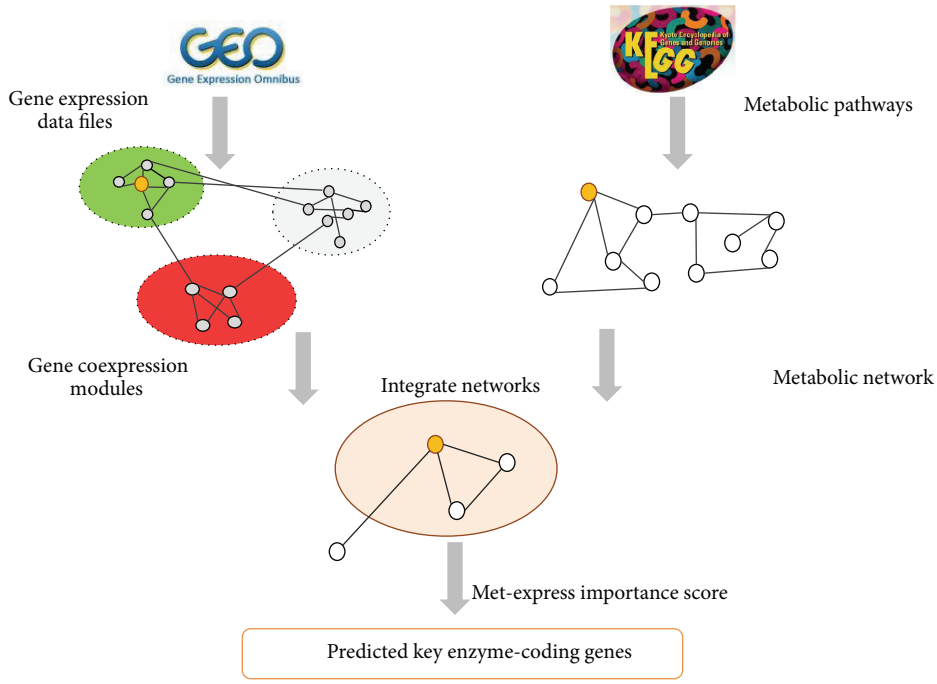


FIGURE 1: The procedure of Met-express. In the gene coexpression modules part, red represents upregulated modules, green represents downregulated modules, and grey means that genes in the module do not change much in expression. For detailed procedures please refer to Section 2, and the methods in Chen et al. [14].

calculated the mean shortest path between the key enzyme-coding gene and all the known disease genes in DrugBank or HGMD. The mean distance for all key enzyme-coding genes was adopted as the distance between our identified genes and known disease genes. As a comparison, we randomly selected the same number of genes as that of the predicted enzyme-coding genes and repeated the above procedures 1000 times. The probability of the case that the mean shortest path length for random selected gene was less than or equal to that of our identified genes was used as the P value.

3. Results and Discussion

3.1. A Brief Introduction of Met-Express. Met-express is a published method originally applied to cancer expression data. It has been successfully used to predict key enzyme-coding genes as potential candidates for therapeutic uses [14]. As in Figure 1, it integrated disease specific gene coexpression network and the metabolic network, to identify those enzyme-coding genes that have a potential to influence downstream genes and thus may play an important role in the disease-related pathways.

3.2. Application of Met-Express to Predict Key Enzyme-Coding Genes in Neurodegenerative Diseases. We constructed a metabolic network with 860 enzyme-coding genes and then applied Met-express to predict key enzyme-coding genes in each of the selected ND datasets. For both PD and HD diseases we combined those genes predicted in at least two of the disease-related datasets and considered them candidate key enzyme-coding genes to the respective diseases.

We applied Met-express to three datasets (GDS2821, GDS3128_GDS3129, and GDS4154) related to PD and predicted 118, 164, and 114 key enzyme-coding genes, respectively. The predicted genes from GDS2821 and GDS3128_GDS3129 shared 35 genes (P value = 0.007), while the number of shared genes for GDS2821 and GDS4154 and GDS3128_GDS3129 and GDS4154 is 15 and 33 (P value = 0.012), respectively. In total, the number of genes predicted in at least two datasets was 69, and we considered them key enzyme-coding genes for PD (P value = 0.013, Supplementary Table 2). Seven out of these 69 genes were predicted in all the three datasets (P value = 0.040). Three microarray data of HD, GDS1331, GDS1332, and GDS4541, were analyzed and resulted in 114, 155, and 79 predicted genes, respectively. There were 50 genes predicted in at least two datasets, and they were considered key enzyme-coding genes for HD (Supplementary Table 2). Four of the 50 genes were predicted in all the three datasets. Among the predicted key enzyme-coding genes in PD and HD, 11 genes were detected in both diseases (Figure 2(e)). However, the 11 genes showed different expression patterns. Among them, only ALDOA and PFKM had consistent expression pattern and were downregulated in both diseases. CYP2B6 and GALC were upregulated in PD and had different expression pattern in different HD datasets. The remaining 7 genes were expressed differently in different datasets. These showed the complex mechanism involved in the pathogenesis of NDs.

Functional enrichment analysis of predicted genes resulted in 94 GO terms and 26 KEGG pathways enriched for PD genes and 45 GO terms and 33 KEGG pathways enriched for HD genes (Figures 2(a), 2(b), 2(c), and 2(d)).

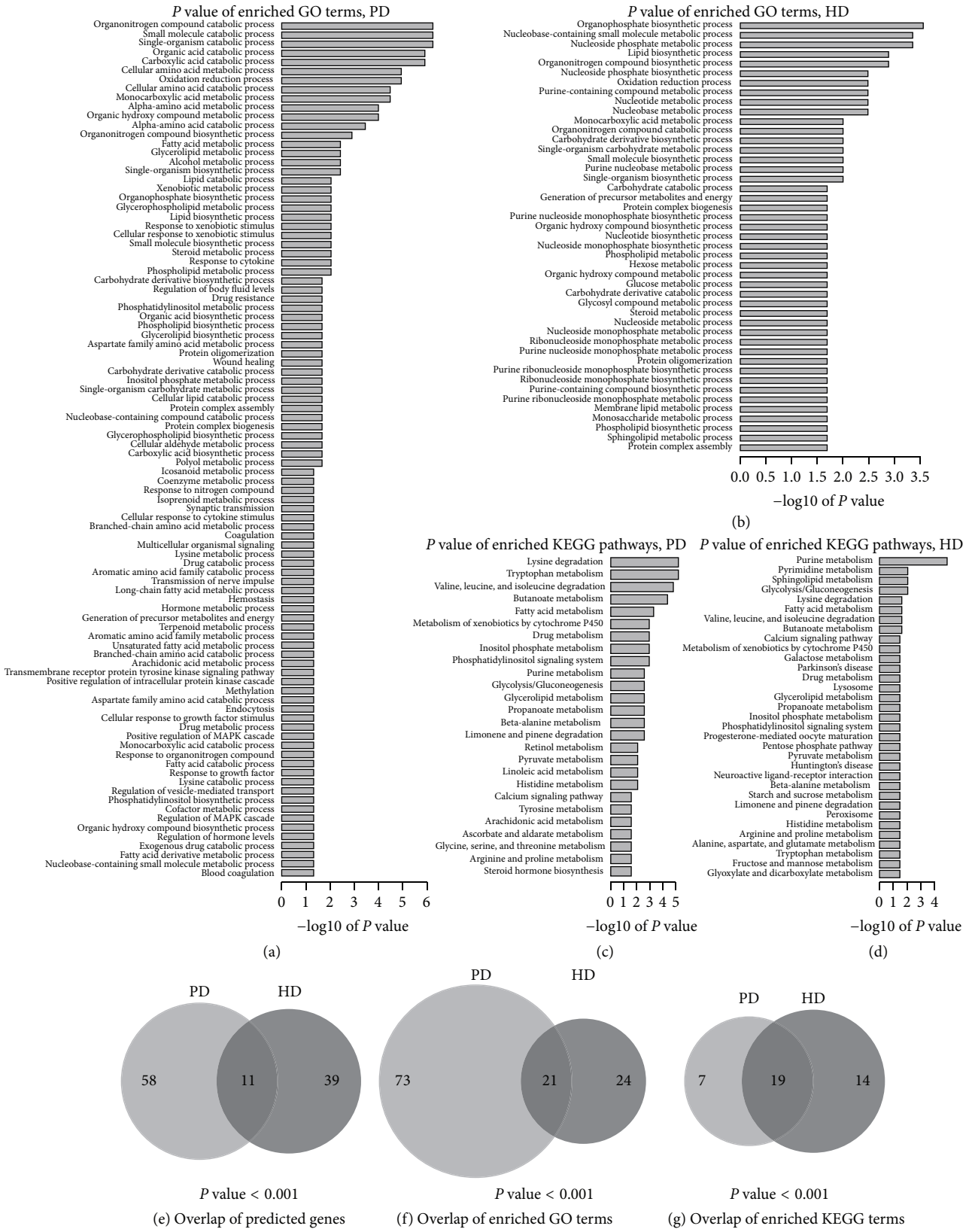


FIGURE 2: (a) and (b) The −log 10 based adjusted P value of significantly enriched GO terms (FDR adjusted P values ≤ 0.05) for predicted genes in PD (a) and HD (b). (c) and (d) The −log 10 based adjusted P value of significantly enriched KEGG terms (FDR adjusted P values ≤ 0.05) for predicted genes in PD (c) and HD (d). (e) The number of overlapped genes between the predicted key enzyme-coding genes of PD and HD. (f) and (g) The number of overlapped significantly enriched GO terms (f) and KEGG terms (g) for predicted key enzyme-coding genes of PD and HD.

The enriched GO terms and KEGG pathways covered various biological processes including "oxidation reduction process," "lipid biosynthetic process," "purine metabolism," and "tryptophan metabolism." Among the enriched terms of these two diseases, 21 GO (P value < 0.001) terms and 19 (P value < 0.001) KEGG pathways were detected in both diseases (Figures 2(f) and 2(g)). The significant functional overlaps between PD and HD genes indicated that there might be common mechanisms in the pathogenesis of NDs. The common pathways included "protein oligomerization," "lipid biosynthetic process," and "phosphatidylinositol signaling system," which have already been suggested to play critical roles in the development of NDs. For example, the accumulation of α-synuclein oligomers was found in PD patients [17], while, in cell line and mouse HD models, the accumulation of intracellular polyglutamine oligomers was suggested to be directly related to HD phenotypes [18]. Cholesterol biosynthesis, one of the most important lipid biosynthetic processes in brain, has been found to be significantly altered with decreased activity in both Parkinson cell lines [19] and HD patients' blood [20]. Alterations in the phosphatidylinositol signaling pathway have also been observed in both cell line models and ND patients and have been linked to autophagy disruption in the pathology of PD and HD [21–23]. The above evidence supported our predictions that alteration and disruption of these pathways may be common features of NDs. Hence, enzymes and metabolites predicted in these pathways may serve as biomarkers for NDs, and targeting on the common metabolic pathways may effectively help with treatment of these diseases.

3.3. Predicted Genes Were Significantly Associated with Known Disease Genes from DrugBank and HGMD. A few of our predicted enzyme-coding genes have already been recorded as biomarkers or drug targets in current known databases: four of the key enzyme-coding genes of PD have been recorded in DrugBank [24] database (http://www.drugbank.ca/), which function in catalyzing PD drug-related reactions (P value = 0.0308), while no key enzyme-coding genes of HD have been recorded in this database (Figure 3(a)). The low overlap between our predicted genes and DrugBank database may be due to the relatively small number of known disease-related enzyme genes. To further investigate the functional association of the predicted genes with known disease genes, we calculated the gene-gene shortest path using the FunCoup [15] network.

The idea is that if a predicted gene is strongly functionally associated with a known disease gene, we would expect to see that they are close to each other in the network; that is, they have significant shorter shortest-path length in the network than random genes. Here, for each disease, the mean shortest path lengths between our predicted genes and the known disease genes in DrugBank or HGMD [25] databases were computed as an evaluation of functional association (Figure 3(b)) and compared to the mean shortest path lengths between random sets of genes and known disease genes. As illustrated in Figures 3(c) and 3(d), the mean shortest path lengths of PD key enzyme-coding genes with two lists of disease genes were both significantly shorter than random (both P value < 0.01). The mean shortest path lengths of HD key enzyme-coding genes with two database disease genes were also both significantly shorter than random (P value < 0.05). Besides DrugBank and HGMD, we have compared our predicted genes to OMIM [26] (data not shown because of too few genes documented for HD) and GAD [27] (Supplementary Figure 1) databases with functional association test and validated that our predicted genes were also significantly functionally associated with the known disease genes in these two databases. These results indicated that our identified key enzyme-coding genes were strongly functionally associated with known disease genes and might participate in the related processes of known disease genes.

3.4. Some Predicted Enzyme-Coding Genes Have Been Implicated to Function in NDs. Besides validation with known databases, literature data mining was also performed for some of the predicted genes. Here we present several examples.

ALDH2 is one of the predicted genes for PD. It encodes mitochondrial aldehyde dehydrogenase 2, which participates in "Glycolysis/Gluconeogenesis," "fatty acid degradation," and many other amino acid metabolic pathways [28]. The activity of mitochondrial aldehyde dehydrogenase 2 was found significantly to be increased in the putamen of PD patients compared to controls [29], and this gene was recently recommended as a new therapeutic target for PD [30]. *MTHFD1*, another predicted key enzyme-coding gene for PD which encodes a protein that possesses three distinct enzymatic activities, has been suggested to be involved in PD mechanisms through homocysteine catabolic process [31].

There are also a number of predicted genes that have been reported in ND-related pathways. For example, *ALDOA* has the most significant association with known disease genes of both PD and HD based on the mean shortest path calculation (Table 1). This gene is downregulated in both PD and HD and was found to be associated with PD-pathogenic proteins α-synuclein and DJ1 in mouse cells [32] and significantly downregulated in the mitochondrial fraction of PC12 cells after PD-linked compound dopamine treatments [33]. *TDO2*, another predicted gene for PD, encodes a critical enzyme, tryptophan 2,3-dioxygenase, in the kynurenine pathway of tryptophan metabolism [34]. This enzyme initiates the metabolism of tryptophan by catalyzing tryptophan to kynurenine, which then is metabolized in two alternative ways either to kynurenic acid or to 3-hydroxykynurenine, both neurotoxic. In PD patients, kynurenine/tryptophan ratio was reported to be higher than in healthy people [35], and the level of kynurenic acid was lower in postmortem analyses of HD patients' brains than control [36]. These indicated that the kynurenine pathway of tryptophan metabolism may be associated with the pathogeneses of both PD and HD and that *TDO2* can be used as a biomarker or therapeutic target for these two diseases.

We have also predicted some novel key genes for NDs, which have little or no previous related reports. Some of these genes function in important pathways of NDs. For example, *AACS*, one of the 11 predicted genes in both two diseases, encodes an enzyme catalyzing cholesterol synthesis and fatty acids synthesis with high expression in human brain [37].

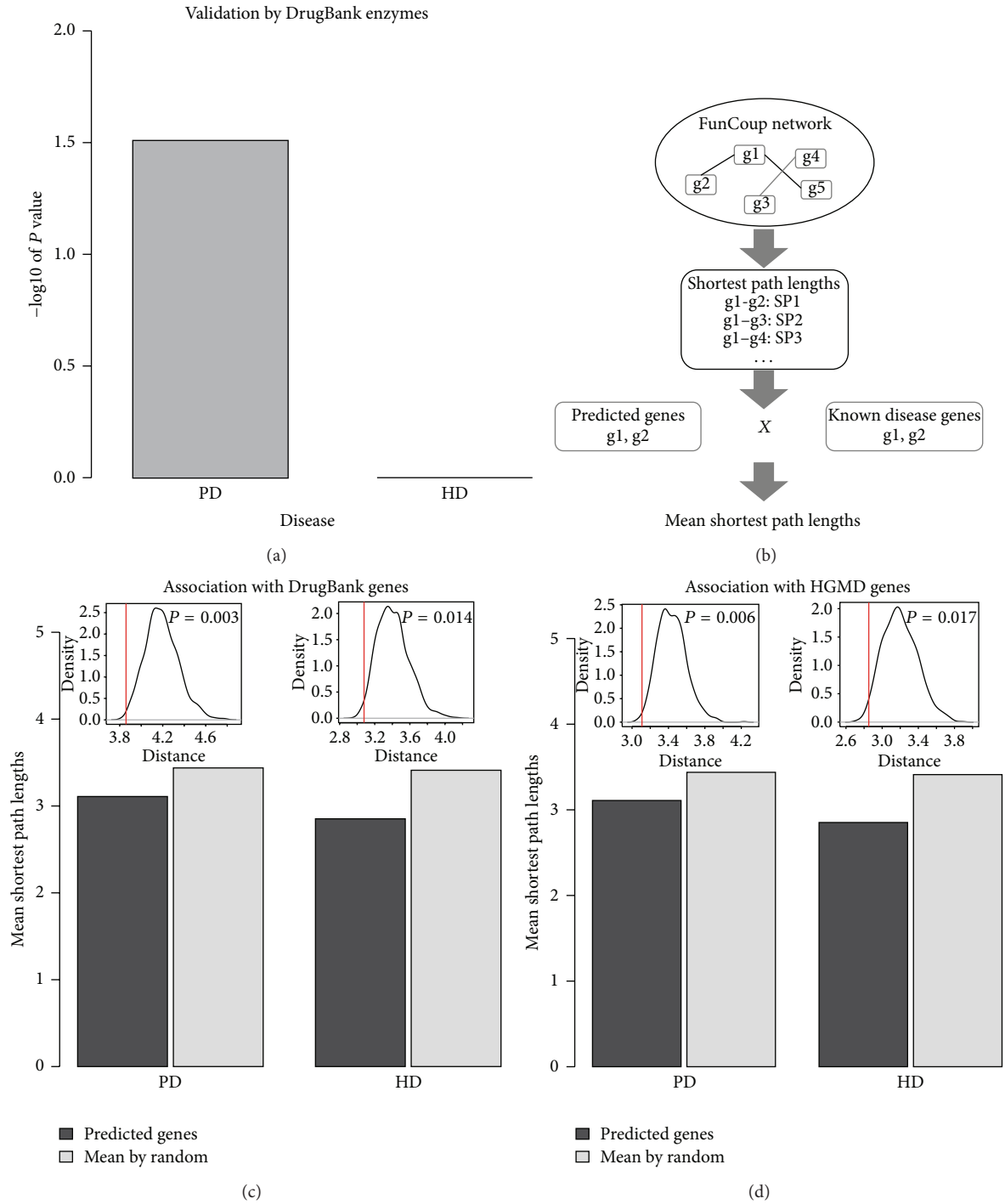


FIGURE 3: (a) The −log 10 based P value of enrichment of predicted key enzyme-coding genes of PD and HD in DrugBank enzymes for the two diseases, respectively. (b) The flowchart of how to calculate the mean shortest path lengths between predicted enzyme-coding genes and known disease genes from public databases. (c) and (d) The mean shortest path lengths of predicted enzyme-coding genes for PD and HD with disease genes from DrugBank (c) and HGMD (d) (the bars, compared with random), and the corresponding distributions of the mean shortest path lengths of random selected genes (the line plots above the bars). The red lines marked the actual mean shortest path lengths for predicted enzyme-coding genes with known disease genes. The black lines represented the distributions of the mean shortest path lengths of random selected genes with known disease genes.

Table 1: Predicted enzyme-coding genes with top 10 closest mean shortest paths to known disease genes.

Disease	Gene symbol	Protein annotation
PD	ALDOA	Fructose-bisphosphate aldolase A
	PFKM	ATP-dependent 6-phosphofructokinase, muscle type
	GOT2	Aspartate aminotransferase, mitochondrial
	ALDH2	Aldehyde dehydrogenase, mitochondrial
	HADHB	Trifunctional enzyme subunit beta, mitochondrial
	MTHFD1	C-1-Tetrahydrofolate synthase, cytoplasmic
	ALDH9A1	4-Trimethylaminobutyraldehyde dehydrogenase
	INPPL1	Phosphatidylinositol 3,4,5-trisphosphate 5-phosphatase 2
	ECHS1	Enoyl-CoA hydratase, mitochondrial
	OGDH	2-Oxoglutarate dehydrogenase, mitochondrial
HD	ALDOA	Fructose-bisphosphate aldolase A
	GPI	Glucose-6-phosphate isomerase
	PFKM	ATP-dependent 6-phosphofructokinase, muscle type
	IMPDH2	IMP (inosine 5′-monophosphate) dehydrogenase 2
	GLA	Alpha-galactosidase A
	SPTLC1	Serine palmitoyltransferase 1
	AK1	Adenylate kinase isoenzyme 1
	EHHADH	Peroxisomal bifunctional enzyme
	POLE3	DNA polymerase epsilon subunit 3
	DCK	Deoxycytidine kinase

It is important for neurite outgrowth, and knockdown of *AACS* has been reported to influence the expression of some neural marker, such as synaptopodin and MAP2 [38]. *SGMS1* encodes the Golgi isoform of sphingomyelin synthase. Its overexpression was associated with suppression of ceramide production and apoptosis after photodamage [39]. *SPTLC1* encodes a subunit of serine palmitoyltransferase, which catalyzes the rate-limited step of sphingolipid synthesis. Mutation in *SPTLC1* would cause neuronal dysfunction [40]. These results implied that *AACS*, *SGMS1*, and *SPTLC1* might contribute to the developments of NDs via lipid metabolic process and oxidation reduction process. Therefore, our predicted key enzyme-coding genes may serve as possible new candidates for further researches of ND mechanisms and therapies.

3.5. Discussion. Met-express integrates gene coexpression network with metabolic network to predict the primary enzyme-coding genes that may be critical for the development of NDs. The predicted genes and enzymes may influence multiple pathways and can bring important influences for ND-related mechanisms, some of which may help explore the common mechanisms of NDs. For example, as previously discussed, *ALDH9A1* was predicted as key gene for both PD and HD by Met-express. It functions in many important pathways, including Glycolysis/Gluconeogenesis, pyruvate metabolism, arginine and proline metabolism, histidine metabolism, tryptophan metabolism, beta-alanine metabolism, and glycerolipid metabolism [28], all of which were enriched in both PD and HD in our analysis. Another example is *EHHADH*, which was also predicted for both diseases. It functions in pathways such as tryptophan metabolism, beta-alanine metabolism, butanoate metabolism, and fatty acid metabolism, which were enriched by both PD key enzyme-coding genes and HD enzyme-coding genes. Although there have been few published evidences relating *ALDH9A1* and *EHHADH* to PD or HD presently, their roles in the many common pathways between PD and HD make them possible targets for further mechanistic investigations.

Many of the enzyme-coding genes predicted by Met-express encode upstream enzymes in their respective pathways. The abnormal expression of these genes in ND patients may have profound impacts on downstream genes and compounds. As a result, altering the abnormal expression of these genes in ND patients bares the potential to help reverse diseases. For example, as previously analyzed, *ALDH2*, a gene predicted for PD, has been raised by others as a potential therapeutic target for PD [30]. Here, Met-express has provided a pool of candidate genes that are of potential value for further therapeutic investigations.

4. Conclusions

Using Met-express, which integrates gene coexpression network and the enzyme network to find key enzyme-coding genes for diseases, we identified 69 and 50 key enzyme-coding genes for Parkinson's disease and Huntington's disease, respectively. Comparison between the functional analyses of the predicted genes indicated that there might be some common pathogenic metabolic pathways for NDs. The predicted genes for PD had a significant overlap with the annotated enzymes in DrugBank. Moreover, predicted genes for PD and HD both showed significantly closer association with known disease genes from DrugBank, HGMD, and other databases than random as evaluated by the mean shortest path lengths. Some of the identified key genes have been reported to be important in the disease-related processes by previous publications. Some novel findings showed potential in influencing pathways that are important in NDs. Thus, application of Met-express to NDs can provide candidates for disease biomarkers and may help with the further research of the etiology, pathology, and therapy of these diseases.

Conflict of Interests

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

Authors' Contribution

Weidong Tian conceived the idea. Qi Ni and Xianming Su conducted the data analysis and drafted the paper. Jingqi Chen helped with the data analysis. Weidong Tian and Jingqi Chen helped in revising the paper. Qi Ni and Xianming Su contributed equally to the work.

Acknowledgments

The authors thank Xinran Dong, Wei Ning, Yulan Lu, and Jifeng Zhang for the technical assistance during the project. This work was supported by the National Natural Science Foundation of China (91231116, 31071113, 30971643, and 31471245), the Specialized Research Fund for the Doctoral Program of Higher Education of China (20120071110018), the Innovation Program of Shanghai Municipal Education Commission (13ZZ006), and the Shanghai Shuguang Program (13SG05).

References

[1] C. A. Ross and M. A. Poirier, "Protein aggregation and neurodegenerative disease," *Nature Medicine*, vol. 10, pp. S10–S17, 2004.

[2] National Collaborating Centre for Chronic Conditions, *Parkinson's Disease: National Clinical Guideline for Diagnosis and Management in Primary and Secondary Care*, Royal College of Physicians, London, UK, 2006.

[3] M. J. Baptista, M. R. Cookson, and K. Gwinn-Hardy, "Parkinson's Disease: Insights from Genetic Causes," in *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*, pp. 1359–1360, Springer, Berlin, Germany, 2006.

[4] F. O. Walker, "Huntington's disease," *The Lancet*, vol. 369, no. 9557, pp. 218–228, 2007.

[5] S. Wieczorek and J. T. Epplen, "Huntington's disease," in *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*, pp. 832–835, Springer, Berlin, Germany, 2006.

[6] Z. T. Kincses and L. Vecsei, "Pharmacological therapy in Parkinson's disease: focus on neuroprotection," *CNS Neuroscience and Therapeutics*, vol. 17, no. 5, pp. 345–367, 2011.

[7] E.-J. Bae, H.-J. Lee, Y.-H. Jang et al., "Phospholipase D1 regulates autophagic flux and clearance of α-synuclein aggregates," *Cell Death & Differentiation*, vol. 21, no. 7, pp. 1132–1141, 2014.

[8] N. Exner, A. K. Lutz, C. Haass, and K. F. Winklhofer, "Mitochondrial dysfunction in Parkinson/s disease: molecular mechanisms and pathophysiological consequences," *The EMBO Journal*, vol. 31, no. 14, pp. 3038–3062, 2012.

[9] A. Campbell, "Inflammation, neurodegenerative diseases, and environmental exposures," *Annals of the New York Academy of Sciences*, vol. 1035, pp. 117–132, 2004.

[10] H. Ischiropoulos and J. S. Beckman, "Oxidative stress and nitration in neurodegeneration: cause, effect, or association?" *The Journal of Clinical Investigation*, vol. 111, no. 2, pp. 163–169, 2003.

[11] S. Grimm, A. Hoehn, K. J. Davies, and T. Grune, "Protein oxidative modifications in the ageing brain: consequence for the onset of neurodegenerative disease," *Free Radical Research*, vol. 45, no. 1, pp. 73–88, 2011.

[12] A. H. V. Schapira, C. W. Olanow, J. T. Greenamyre, and E. Bezard, "Slowing of neurodegeneration in Parkinson's disease and Huntington's disease: future therapeutic perspectives," *The Lancet*, vol. 9942, no. 384, pp. 545–555, 2014.

[13] K. Jomova, D. Vondrakova, M. Lawson, and M. Valko, "Metals, oxidative stress and neurodegenerative disorders," *Molecular and Cellular Biochemistry*, vol. 345, no. 1-2, pp. 91–104, 2010.

[14] J. Chen, M. Ma, N. Shen, J. J. Xi, and W. Tian, "Integration of cancer gene co-expression network and metabolic network to uncover potential cancer drug targets," *Journal of Proteome Research*, vol. 12, no. 6, pp. 2354–2364, 2013.

[15] A. Alexeyenko, T. Schmitt, A. Tjärnberg, D. Guala, O. Frings, and E. L. L. Sonnhammer, "Comparative interactomics with Funcoup 2.0," *Nucleic Acids Research*, vol. 40, no. 1, pp. D821–D828, 2012.

[16] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.

[17] H. A. Lashuel, C. R. Overk, A. Oueslati, and E. Masliah, "The many faces of α-synuclein: from structure and toxicity to therapeutic target," *Nature Reviews Neuroscience*, vol. 14, no. 1, pp. 38–48, 2013.

[18] S. Hands, M. U. Sajjad, M. J. Newton, and A. Wyttenbach, "In vitro and in vivo aggregation of a fragment of huntingtin protein directly causes free radical production," *The Journal of Biological Chemistry*, vol. 286, no. 52, pp. 44512–44520, 2011.

[19] R. Musanti, E. Parati, E. Lamperi, and G. Ghiselli, "Decreased cholesterol biosynthesis in fibroblasts from patients with Parkinson disease," *Biochemical Medicine and Metabolic Biology*, vol. 49, no. 2, pp. 133–142, 1993.

[20] M. Valenza and E. Cattaneo, "Emerging roles for cholesterol in Huntington's disease," *Trends in Neurosciences*, vol. 34, no. 9, pp. 474–486, 2011.

[21] D. Heras-Sandoval, J. M. Pérez-Rojas, J. Hernández-Damián, and J. Pedraza-Chaverri, "The role of PI3K/AKT/mTOR pathway in the modulation of autophagy and the clearance of protein aggregates in neurodegeneration," *Cellular Signalling*, vol. 26, no. 12, pp. 2694–2701, 2014.

[22] T. F. Franke, "PI3K/Akt: getting it right matters," *Oncogene*, vol. 27, no. 50, pp. 6473–6488, 2008.

[23] S. Humbert, E. A. Bryson, F. P. Cordelières et al., "The IGF-1/Akt pathway is neuroprotective in Huntington's disease and involves huntingtin phosphorylation by Akt," *Developmental Cell*, vol. 2, no. 6, pp. 831–837, 2002.

[24] V. Law, C. Knox, Y. Djoumbou et al., "DrugBank 4.0: shedding new light on drug metabolism," *Nucleic Acids Research*, vol. 42, no. 1, pp. D1091–D1097, 2014.

[25] P. D. Stenson, "The human gene mutation database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution," in *Current Protocols in Bioinformatics*, unit 1.13, John Wiley & Sons, New York, NY, USA, 2012.

[26] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 33, pp. D514–D517, 2005.

[27] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, "The Genetic Association Database [1]," *Nature Genetics*, vol. 36, no. 5, pp. 431–432, 2004.

[28] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[29] T. M. Michel, L. Käsbauer, W. Gsell et al., "Aldehyde dehydrogenase 2 in sporadic Parkinson's disease," *Parkinsonism & Related Disorders*, vol. 20, no. 1, pp. S68–S72, 2014.

[30] C. H. Chen, J. C. B. Ferreira, E. R. Gross, and D. Mochly-Rosen, "Targeting aldehyde dehydrogenase 2: new therapeutic opportunities," *Physiological Reviews*, vol. 94, no. 1, pp. 1–34, 2014.

[31] J. Dorszewska, J. Florczak, A. Rozycka et al., "Oxidative DNA damage and level of thiols as related to polymorphisms of MTHFR, MTR, MTHFD1 in Alzheimer's and Parkinson's diseases," *Acta Neurobiologiae Experimentalis*, vol. 67, no. 2, pp. 113–129, 2007.

[32] J. Jin, G. J. Li, J. Davis et al., "Identification of novel proteins associated with both α-synuclein and DJ-1," *Molecular & Cellular Proteomics*, vol. 6, no. 5, pp. 845–859, 2007.

[33] A. A. Dukes, V. S. van Laar, M. Cascio, and T. G. Hastings, "Changes in endoplasmic reticulum stress proteins and aldolase A in cells exposed to dopamine," *Journal of Neurochemistry*, vol. 106, no. 1, pp. 333–346, 2008.

[34] T. Tatusova, S. Ciufo, B. Fedorov, K. O'Neill, and I. Tolstoy, "RefSeq microbial genomes database: new representation and annotation strategy," *Nucleic Acids Research*, vol. 42, no. 1, pp. D553–D559, 2014.

[35] N. Szabó, Z. T. Kincses, J. Toldi, and L. Vécsei, "Altered tryptophan metabolism in Parkinson's disease: a possible novel therapeutic approach," *Journal of the Neurological Sciences*, vol. 310, no. 1-2, pp. 256–260, 2011.

[36] T. W. Stone, C. M. Forrest, N. Stoy, and L. G. Darlington, "Involvement of kynurenines in Huntington's disease and stroke-induced brain damage," *Journal of Neural Transmission*, vol. 119, no. 2, pp. 261–274, 2012.

[37] M. Ohgami, N. Takahashi, M. Yamasaki, and T. Fukui, "Expression of acetoacetyl-CoA synthetase, a novel cytosolic ketone body-utilizing enzyme, in human brain," *Biochemical Pharmacology*, vol. 65, no. 6, pp. 989–994, 2003.

[38] S. Hasegawa, H. Kume, S. Iinuma, M. Yamasaki, N. Takahashi, and T. Fukui, "Acetoacetyl-CoA synthetase is essential for normal neuronal development," *Biochemical and Biophysical Research Communications*, vol. 427, no. 2, pp. 398–403, 2012.

[39] D. Separovic, K. Hanada, M. Y. Awad Maitah et al., "Sphingomyelin synthase 1 suppresses ceramide production and apoptosis post-photodamage," *Biochemical and Biophysical Research Communications*, vol. 358, no. 1, pp. 196–202, 2007.

[40] A. McCampbell, D. Truong, D. C. Broom et al., "Mutant SPTLC1 dominantly inhibits serine palmitoyltransferase activity in vivo and confers an age-dependent neuropathy," *Human Molecular Genetics*, vol. 14, no. 22, pp. 3507–3521, 2005.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 841956, 10 pages
http://dx.doi.org/10.1155/2015/841956

Research Article

Identification of Gene and MicroRNA Signatures for Oral Cancer Developed from Oral Leukoplakia

Guanghui Zhu,[1] Yuan He,[2] Shaofang Yang,[2] Beimin Chen,[3] Min Zhou,[2] and Xin-Jian Xu[1]

[1]*Department of Mathematics, Shanghai University, Shanghai 200444, China*
[2]*Laboratory of Oral Biomedical Science and Translational Medicine, School of Stomatology, Tongji University, Middle Yanchang Road 399, Shanghai 200072, China*
[3]*Shanghai Tenth People's Hospital, Middle Yanchang Road 30, Shanghai 200072, China*

Correspondence should be addressed to Xin-Jian Xu; xinjxu@shu.edu.cn

Received 26 July 2014; Accepted 14 October 2014

Academic Editor: Hao-Teng Chang

In clinic, oral leukoplakia (OLK) may develop into oral cancer. However, the mechanism underlying this transformation is still unclear. In this work, we present a new pipeline to identify oral cancer related genes and microRNAs (miRNAs) by integrating both gene and miRNA expression profiles. In particular, we find some network modules as well as their miRNA regulators that play important roles in the development of OLK to oral cancer. Among these network modules, 91.67% of genes and 37.5% of miRNAs have been previously reported to be related to oral cancer in literature. The promising results demonstrate the effectiveness and efficiency of our proposed approach.

1. Introduction

Oral leukoplakia (OLK) is one of the most common malignant disorders of the oral mucosa. There are about 2% to 3% of OLK cases that develop into oral cancers annually [1]. Therefore, the early diagnosis of the risk of OLKs developing into oral cancers can help prevent the disease process with timely and effective intervention. Unfortunately, it is still unclear how the OLKs develop into oral cancer until now. Recently, with the emergency of microarray technology that can monitor thousands of genes simultaneously, gene biomarkers are being detected for oral cancers. For example, Saintigny et al. [2] defined a 29-transcripts signature while Kondoh et al. [3] defined another 11-genes signature that can help separate oral cancers developed from OLKs from normal OLKs. Despite the good discrimination capacity of the transcript signature, few of the genes in the signature have functional relationships which make it difficult to understand the malignant transformation of oral leukoplakia.

In this work, we developed a novel pipeline to detect genes that play important roles in the development of oral cancer from a systematic perspective. Based on the protein-protein interaction (PPI) network and gene expression profiles, we detected the network modules that are associated with oral cancer development. In particular, we identified microRNAs (miRNAs) that regulate the oral cancer associated modules. Both the genes from our identified network modules and those miRNA regulators are found to be indeed related to the development of oral cancers, indicating the important roles of these modules and their miRNA regulators in the pathogenesis of oral cancer.

2. Materials and Methods

Figure 1 depicts our proposed pipeline to identify networks modules as well as their miRNA regulators that play important roles in the development of oral cancer.

2.1. Gene Expression. Two independent datasets (GSE33299, GSE26549) were downloaded from the NCBI Gene Expression Omnibus (GEO) [4]. The GSE33299 [5] dataset consists of miRNA expression profiles measured in three different tissues, including normal mucousal tissue (5 samples), OLK

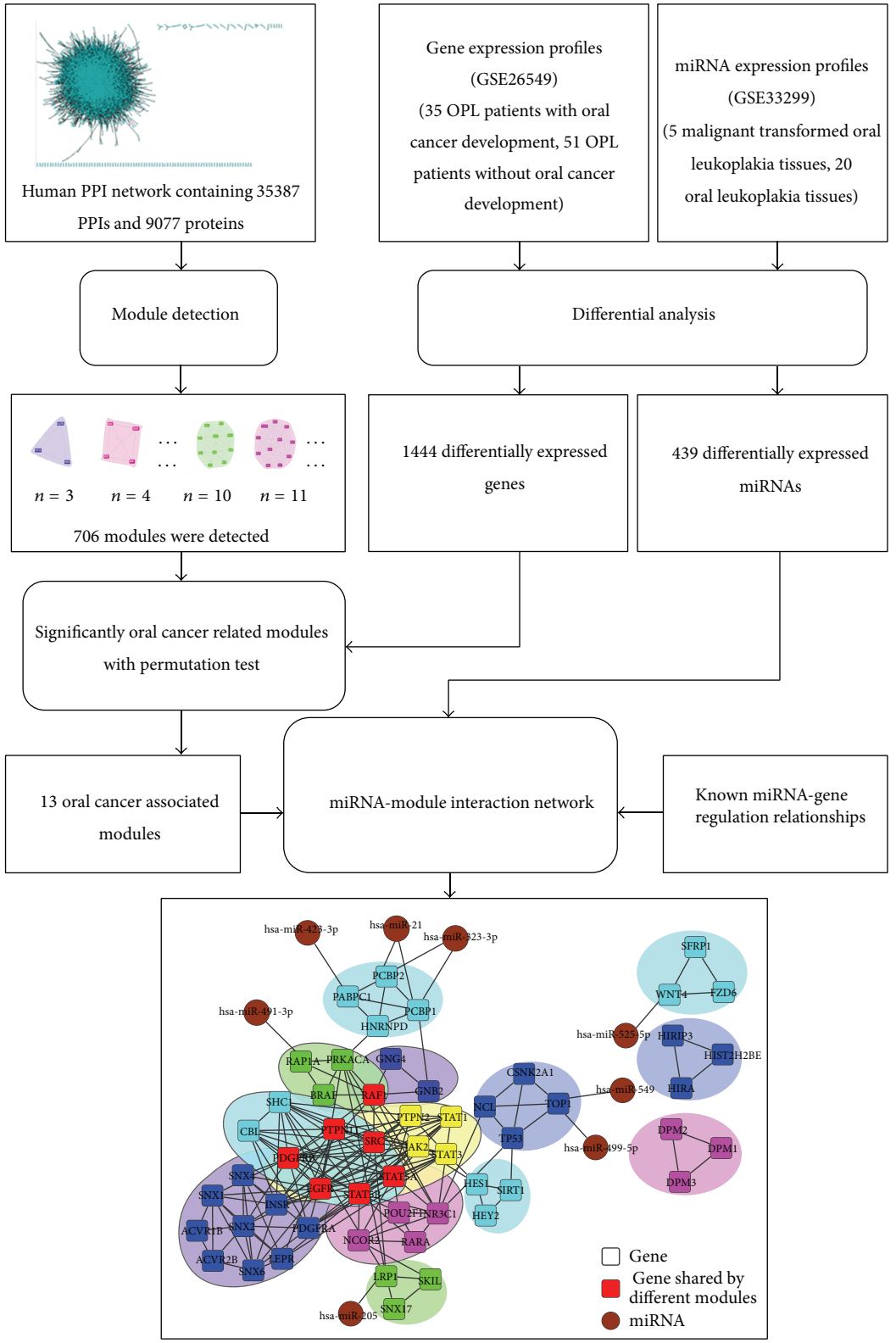


Figure 1: Schematic illustration of the pipeline to construct oral cancer associated network and detect key genes.

tissue (20 samples) and malignant transformed OLK tissue (5). The gene expression data were preprocessed with background subtraction and normalization using the global lowess regression model, and the missing expression values in the data were generated with Impute package in R [6]. The GSE26549 [2] dataset contains gene expression profiles measured in 35 OLK patients who developed oral cancer in follow-up time and 51 OLK patients that develop to oral cancers. All the expression data were normalized with quantile normalization and the robust multi-array average (RMA) approach. Specifically, the expression value of the gene associated with multiple probes was calculated as the average expression value of all related probes.

2.2. Identification of Differentially Expressed Genes and miRNAs. The miRNAs that are differentially expressed between OLKs and malignant transformed OLKs were detected by student's t-test with P-value cutoff of 0.01. As a result, 439 differentially expressed miRNAs (DEmiRs) were detected. Similarly, 1444 differentially expressed genes (DEGs) were detected by comparing their expression in OLK patients with or without oral cancer consequence with the help of student's t-test (P-value cutoff of 0.01).

2.3. Identification of Modules Associated with the Development of Oral Cancer. The human PPIs consist of 35387 PPIs among 9077 proteins (self-interactions excluded) were collected from the HPRD database [7]. The network modules were detected from the PPI network by utilizing CFinder [8] with default parameters. As a result, there are in total 706 modules that were identified. For each module, we defined a score S to measure its relevance to the development of oral cancer as follows:

$$S = \frac{\sum_{i=1}^{m} t_i}{m}, \tag{1}$$

where m denotes the number of genes in the module, t_i denotes the t-score of gene i obtained with student's t-test analysis on the gene expression data between OLKs with or without oral cancer. Moreover, to control the false discovery rate, the P-value for each module was defined as the probability that this module was observed by chance. In particular, the same number of genes as that in the module was randomly picked up and the score for this gene set was calculated as described in (1). This procedure was repeated for 10000 times, and the frequency of observing a score larger than S was defined as the P-value for the corresponding module. Consequently, 13 modules were selected with P-value threshold of 0.01, and these modules were regarded as modules associated with oral cancer (MAOCs) and play important roles in the development of oral cancer.

2.4. Network Analysis. In complex networks, compared with average nodes, the hub nodes that have larger degree and link to more nodes generally play more important roles in the system. Therefore, we detected the hub genes from the 13 MAOCs. In the detail, we first defined the scaled connectivity K_i for the ith gene as follows:

$$K_i = \frac{k_i}{k_{\max}}, \tag{2}$$

where the k_i is the degree of gene i, $k_{\max}$ denotes the maximum degree of the genes in the MAOCs. Subsequently, the genes that have K_i larger than 0.9 were regarded as hub genes hereinafter.

Except for the hub genes, there are some genes that link distinct modules, which were also assumed to be important since they bridge the signal flows between distinct modules. In particular, we grouped these genes into two categories: (1) genes located in one module and interact with genes from other modules; (2) genes outside of MAOCs but interact with genes from other modules. Especially, for each of the above identified genes, the connecting score CS_i for the ith gene was defined as the number of MAOCs that this gene links to.

2.5. Regulations between miRNAs and Genes. Target genes of miRNAs were predicted through utilizing tools including PicTar [9], miRanda [10], MicroT [11] and TargetScan [12]. Target genes of each miRNA were kept only if they were predicted by at least two tools. Target genes of each miRNA were extended with experimentally determined miRNA target genes deposited in Tarbase [13]. A set of 23336 miRNA-gene interactions involving 548 miRNAs and 6797 genes was obtained.

2.6. Identification of miRNAs Regulating MAOCs. Based on miRNA-gene regulations, given one miRNA, a module will be regarded to be regulated by this miRNA if its target genes are significantly enriched in this module, where the enrichment analysis was conducted with Fisher exact test with P-value cutoff of 0.01. As a result, for each DEmiR, we identified the modules that can be regulated by this miRNA.

3. Results

3.1. Network Modules Associated with Oral Cancer. In the 13 MAOCs we identified, there are in total 54 genes involved, among which 38.89% are DEGs (more details see Section 2). Figure 2 shows the PPI network consists of proteins involved in the 13 modules, where the interactions between proteins were obtained from the HPRD database, where distinct modules were marked with different colors while those genes occurring in more than two modules were marked in red. Table 1 summarizes the genes from the 13 modules. Furthermore, the functional enrichment analysis was performed for these genes with the Database for Annotation, Visualization and Integrated Discovery (DAVID) service [14, 15]. From the results (supplementary Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/841956) we can see that the well known processes associated with cancer development, such as cell proliferation, apoptosis, cycle, migration and differentiation, are significantly in the module genes. Moreover, the epidermal growth factor receptor signaling pathway that has been previously reported to be

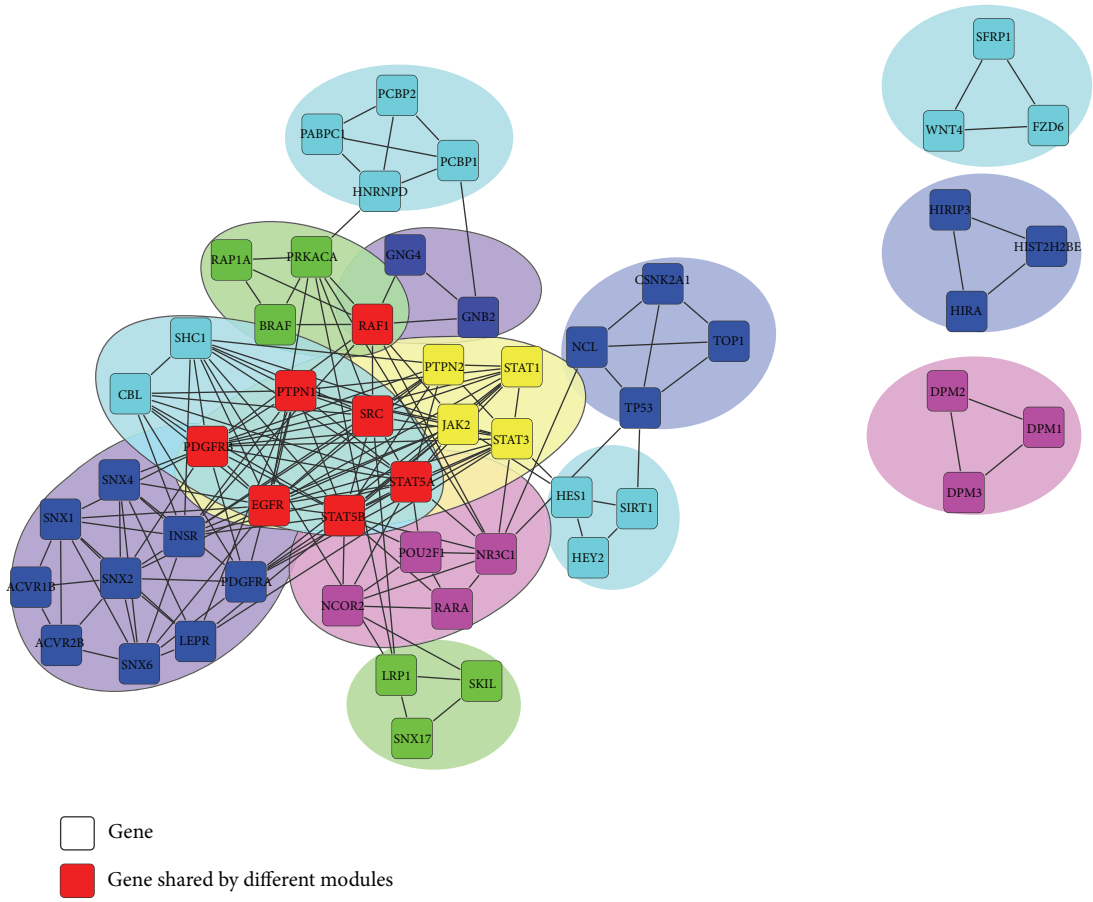


Figure 2: The protein-protein interaction network consists of proteins as well as their interactions from the 13 modules, where distinct modules were marked in different colors and the genes in red denote those occurring in more than two modules.

related to squamous cell carcinoma (SCC) [16] is significantly enriched in the module genes. In addition, the pathways in cancer and Wnt signaling pathway, which are known to be related to cancer, were also found to be significantly enriched [17]. The functional enrichment analysis results indicate that our identified modules are related to the development of oral cancer.

Since there are some genes that have already been known to be related to oral cancer, to validate the associations between our identified modules and oral cancer, we obtained a gene list of 2990 genes from the GeneCards [18] by querying "oral cancer". Among the 54 genes in the 13 modules, 30 (55.56%) genes can be found in the 2990 genes (Table 2). That is, these 30 genes have already been annotated to be related to oral cancer. In particular, 15 of the 30 genes (50%) were found to be ranked in the top 32% of the gene list according to the relevance scores provided by GeneCards, where the relevance score was calculated by Lucene scoring to determine how relevant a given document is to a query. Considering that the SCC accounts for the majority of oral cancer cases, a list of genes was retrieved from GeneCards by querying "squamous carcinoma". As a result, 33 (61.11%) of our module genes can be found in the 2648 genes annotated to be related to SCC, while 24 (72.73%) of the these genes were ranked in the top 54% of the gene list according to the relevance scores provided by GeneCards (supplementary Table 2). The overlap between our identified module genes and those known oral cancer genes indicates that our identified modules are indeed related to the development of oral cancer.

3.2. Topological Important Genes in the MAOCs. In general, the hub nodes and those linking modules play more important roles in the system underlying the network. In our identified MAOCs, we picked up 4 hub genes, 5 intra-MAOC genes and 5 inter-MAOC genes, where the intra-MAOC genes are those inside one module and have interactions with genes belonging to other modules while the inter-MAOC genes are those outside of MAOCs but have interactions with genes belonging to MAOCs. Table 3 lists these topological important genes. Note that it is possible that some genes are both hub genes and intra-MAOC genes, for example, PDGFRB gene.

(1) The hub genes we identified have more interactions than other genes, that is, higher degree, and therefore have more important roles. Among the 4 hub genes, the three genes EGFR, PDGFRB, STAT5A have already been annotated to be related to oral cancer according to GeneCards (see Table 2). In the MAOCs, the gene STAT5B has high degree

Table 1: The modules we identified to be associated with oral cancer, and the enriched functions and pathways of the proteins from these modules.

	Genes in module (DEGs were in bold)	Top 10 enriched biological processes	Top 10 enriched KEGG pathways
Module 1	DPM3 DPM1 **DPM2**	GO: 0032870 cellular response to hormone stimulus	hsa05220: Chronic myeloid leukemia
Module 2	**HIRA HIST2H2BE HIRIP3**		
Module 3	EGFR PDGFRB PTPN11 SHC1 **SRC STAT5A STAT5B** CBL	GO: 0007169 transmembrane receptor protein tyrosine kinase signaling pathway	hsa05200: Pathways in cancer
Module 4	CSNK2A1 NCL **TOP1 TP53**		
Module 5	NR3C1 **POU2F1** RARA **STAT5A STAT5B** NCOR2	GO: 0009725 response to hormone stimulus	hsa04062: Chemokine signaling pathway
Module 6	**LRP1** SKIL SNX17	GO: 0007167 enzyme linked receptor protein signaling pathway	hsa05210: Colorectal cancer
Module 7	GNB2 GNG4 **RAF1**		
Module 8	EGFR INSR LEPR **PDGFRA** PDGFRB SNX6 **SNX1** SNX2 SNX4 ACVR1B ACVR2B	GO: 0009719 response to endogenous stimulus	hsa04012: ErbB signaling pathway
Module 9	EGFR JAK2 PDGFRB PTPN2 PTPN11 **SRC** STAT1 STAT3 **STAT5A STAT5B**	GO: 0010033 response to organic substance	hsa05214: Glioma
Module 10	**SIRT1** HEY2 **HES1**	GO: 0007166 cell surface receptor linked signal transduction	hsa05212: Pancreatic cancer
Module 11	WNT4 SFRP1 **FZD6**		
Module 12	PABPC1 **HNRNPD PCBP1 PCBP2**	GO: 0006468 protein amino acid phosphorylation	hsa05221: Acute myeloid leukemia
Module 13	PRKACA **RAF1 RAP1A** BRAF		
		GO: 0018108 peptidyl-tyrosine phosphorylation	hsa04630: Jak-STAT signaling pathway
		GO: 0018212 peptidyl-tyrosine modification	hsa05218: Melanoma

Table 2: The 30 genes that have been annotated to be related to oral cancer and their relevance scores according to GeneCards.

Gene symbol	Relevance score
TP53	10.19
EGFR	10.17
RARA	9.54
SRC	9.48
BRAF	9.37
TOP1	9.37
STAT3	9.32
SFRP1	9.26
RAF1	9.18
STAT1	9.1
LEPR	9.08
STAT5A	8.89
PCBP2	8.43
HNRNPD	8.32
PDGFRB	0.18
PDGFRA	−0.21
NCL	−0.29
PRKACA	−0.39
NR3C1	−0.44
CSNK2A1	−0.52
FZD6	−0.59
INSR	−0.59
PTPN11	−0.64
JAK2	−0.77
WNT4	−0.87
RAP1A	−0.95
SNX1	−1.01
PTPN2	−1.13
PABPC1	−1.16
POU2F1	−1.25

of 17. Despite this gene has not been annotated to be related to oral cancer in GeneCards, both STAT5A and STAT5B belong to the STAT5 family the aberrant activity of which has been found to be related to various cancers [19]. Especially for SCC, blockade of STAT5B in a xenograft model in head and neck squamous cell carcinoma (HNSCC) resulted in tumor growth inhibition [20], and the constitutive activation of STAT5A was one of the early events in tobacco mediated-oral squamous cell carcinoma (OSCC) in the eastern Indian population [21]. Among the other two hub genes both with high degree of 16, EGFR has been among the most important prognostic factors for HNSCC [16], while the kinase PDGFRB was found to be up regulated in tumor indicating the effectiveness of tyrosine and serine-threonine kinase inhibitors in the treatment of HNSCCs [22].

(2) Intra-MAOC genes interact with genes in distinct MAOCs which effected oral cancer development in the common pathological system. The 5 intra-MAOC genes with highest connecting scores were identified as key genes, all of which are found to be related to oral cancer according to GeneCards (see Table 2). For example, the gene RAF1 (Figure 3) has a medium degree of 10 and the highest connecting score of 8, implying that RAF1 is important for oral cancer development although it is not a hub gene. In fact, RAF1/Rok-α interaction has been found to play a critical role in the pathogenesis of SCCs [23]. Except for the hub genes PDGFRB and EGFR, SRC has also been previously reported to be related to SCC with increased expression in HNSCC [24].

(3) Inter-MAOC genes are those outside of MAOCs and interact with genes in more than 2 distinct MAOCs. Similarly, the top 5 inter-MAOC genes with the highest connecting scores were identified as key genes. For example, the PRKCA gene with the top connecting score of 8 (Figure 3) interacts with 6 genes (INSR, EGFR, RARA, SRC, RAF1, TP53) from MAOCs, where these 6 genes have been reported to be oral cancer related genes according to GeneCards (see Table 2). Therefore, it is reasonable to conclude that PRKCA plays important roles in the development of oral cancer. Actually, PRKCA was found to be highly expressed in HNSCC [25]. Moreover, it was found that the phosphorylation of MAPK1/MAPK3 was correlated with tumor growth in OSCC [26], the methylation of ESR1 promoter was associated with squamous cell cervical cancer [27], and the activation of FYN kinase was related to OSCC [28].

From above results, we can see that most of our identified key genes (11/12) have been reported previously to be associated with oral cancer, which demonstrate that our identified network modules are indeed related to the pathogenesis of oral cancer.

3.3. miRNAs Regulators of Network Modules. Among our identified 439 differentially expressed miRNAs (DEmiRs), 8 DEmiRs were found to regulate 5 modules as shown in Figure 4. By investigating the expression profiles of these DEmiRs as well as those of their target genes, we noticed that the expression of 4 miRNAs is negatively correlated with that of their target genes (Table 4). For example, the expression of hsa-miR-491-3p is down-regulated in oral cancers samples while the expression of its corresponding target gene RAP1A is up-regulated in oral cancers, which is consistent with the general miRNA regulation principle that miRNAs repress the expression of their target genes. In literature, RAP1A has been reported as a critical mediator of HNSCC [29]. Moreover, in module 13, RAP1A interacts with BRAF and RAF1 that have also been reported to be associated with the progression of SCC [23, 30–32]. The evidence from literature makes it clear that module 13 is indeed related to the oral cancer, and it is reasonable that the regulator, that is, hsa-miR-491-3p, of this module also play important roles in the development of oral cancer.

Another example is hsa-miR-21 that regulates the module 12 by directly regulating its target genes PCBP1 and PCBP2. In the oral cancers, hsa-miR-21 was found to be up-regulated while its target genes PCBP1 and PCBP2 were found to be down-regulated. The two genes PCBP1 and PCBP2 have been reported to be related to SCC [33, 34] in literature. In module 12, the four member genes interact with each other and form

Table 3: The details of topological important genes, including hub genes, intra-MAOC genes and inter-MAOC genes.

Hub gene	Degree	Intra-MAOC gene	Connecting score	Inter-MAOC gene	Connecting scores
STAT5B	17	RAF1	8	PRKCA	8
EGFR	16	PDGFRB	7	MAPK1	7
PDGFRB	16	SRC	6	MAPK3	7
STAT5A	16	PRKACA	6	ESR1	6
		EGFR	4	FYN	6

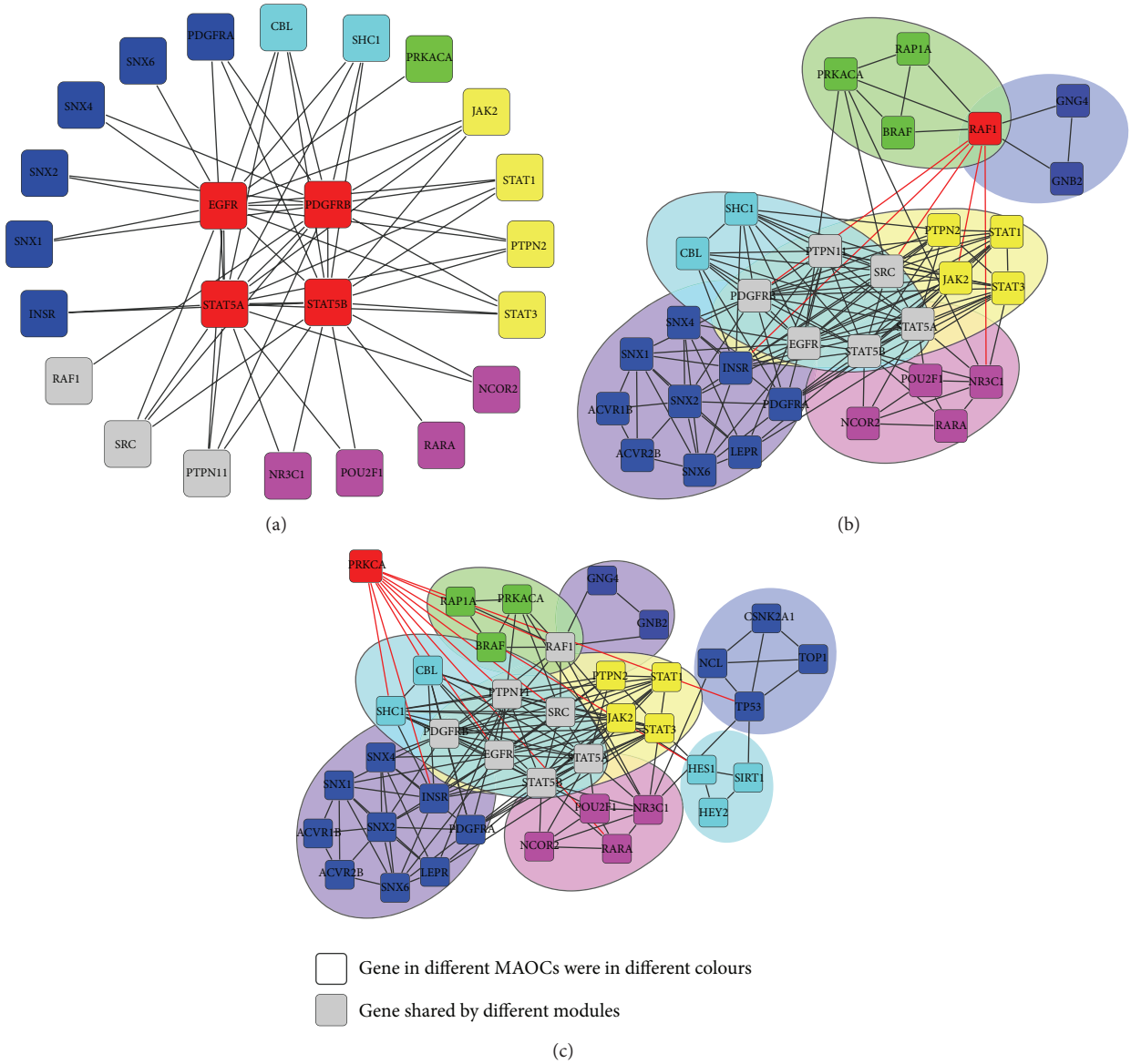


Figure 3: As representative for the three kinds of genes which were detected as key genes associated with oral cancer development, (a) 4 hub genes, (b) RAF1 as an intra-MAOC gene had 5 interactions between modules getting an connecting score of 8, and (c) PRKCA as an inter-MAOC gene connected 8 modules getting an connecting score of 8 were shown in this figure.

a clique structure. Among the 4 genes, the knockdown of HNRNPD has been found to reduce the hTERT promoter activity in OSCC while PABPC was shown as critical in hTERT regulation by human papillomavirus 16 E6 which had been associated with HNSCC [35–37]. It is obvious that module 12 plays important roles in SCC. Except for miR-21, another two miRNAs miR-323-3p and miR-423-3p were also found to regulate module 12. As the regulators of module 12, and we believe its miRNA regulators should also be related to SCC. In fact, the deregulation of miR-21 has found

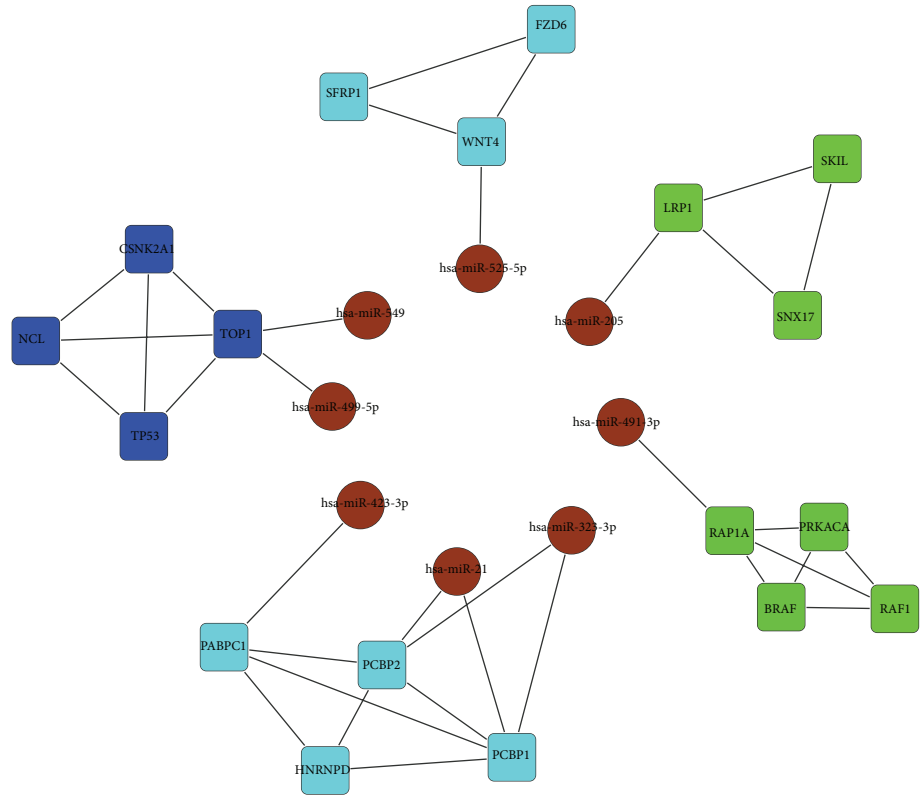


Figure 4: The miRNA-module interaction network, where the nodes in circle are miRNAs and the rest are genes and distinct modules were marked in different color.

Table 4: The miRNA regulators of modules and their target genes.

miRNA (in bold if the expression of miRNA and that of its target gene is negatively correlated)	miRNA Expression change direction	miRNA target gene	Gene Expression change direction
hsa-miR-499-5p	Up	TOP1	up
hsa-miR-549	Up	TOP1	up
hsa-miR-205	Up	LRP1	down
hsa-miR-525-5p	down	WNT4	down
hsa-miR-21	Up	PCBP1 PCBP2	down down
hsa-miR-323-3p	down	PCBP1 PCBP2	down down
hsa-miR-423-3p	down	PABPC1	up
hsa-miR-491-3p	Up	RAP1A	down

to be related to HNSCC in literature [38], and the higher expression of miR-21 was associated with shortened survival time in squamous cell lung carcinoma [39].

Except for the above two miRNAs, another two hsa-mir-499 and miR-205 have also been reported to play important roles in OSCC in literature. It was found that hsa-mir-499 was associated with the reduced risk of HNSCC [40] and miR-205 can be used as a biomarker in discriminating SCC from adenocarcinoma and small cell lung carcinoma with high accuracy [41]. The evidence supports from literature demonstrate that the miRNA regulators of our identified modules are indeed related to the development of oral cancer.

4. Conclusions

The oral leukoplakia frequently develops into oral cancers, however, the mechanism underlying which is still unclear. In this work, with the assumption that the oral cancers were caused due to the dysfunction of multiple functionally related genes [42], we proposed a novel pipeline to identify gene modules that play important roles in the development of oral cancers. The genes in our identified modules were found to be indeed related to oral cancer and the miRNA regulators of these modules were also reported in literature to be related to the cancer, indicating the effectiveness of

our proposed approach. In particular, we identified some key genes that play crucial roles in the malignant transformation of oral leukoplakia to oral cancer, including 4 hub genes (STAT5B, EGFR, PDGFRB, STAT5A), 5 intra-MAOC genes (RAF1, PDGFRB, SRC, PRKACA, EGFR), 5 inter-MAOC genes (PRKCA, MAPK1, MAPK3, ESR1, FYN) and 8 miRNAs (hsa-miR-499-5p, hsa-miR-549, hsa-miR-205, hsa-miR-525-5p, hsa-miR-21, hsa-miR-323-3p, hsa-miR-423-3p, hsa-miR-491-3p). These findings may provide new clues for future research.

This work also supplies a feasible work frame to detect key genes and miRNAs in development of other special type of caner based on gene and miRNA expression profiles. However the process of identification is largely based on the acknowledged PPI network that is not yet complete, so our method is supposed to achieve more realistic detection while the PPI network is more complete and reliable. On the other hand, considering that the process of identification is depend on the algorithm to detect the modules in PPI network and the parameters set to the algorithm and both the algorithm and the special parameters are not designed based on evolving nature of relations between molecules in cancer development but mathematical logic, selecting module detecting algorithm that can exploit the significance in biological process is supposed to improve the identification capacity of our method.

Conflict of Interests

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

Authors' Contribution

Guanghui Zhu and Yuan He contributed equally to this paper.

Acknowledgment

This work was partly supported by the Natural Science Foundation of China (11331009), the Science and Technology Commission of Shanghai Municipality (13ZR1416800), the guided project from Shanghai Science and Technology Commission (09411965900) and the Natural Science Foundation of Shanghai Science and Technology Commission (14ZR1443600).

References

[1] D. Arnaoutakis, J. Bishop, W. Westra, and J. A. Califano, "Recurrence patterns and management of oral cavity premalignant lesions," *Oral Oncology*, vol. 49, no. 8, pp. 814–817, 2013.

[2] P. Saintigny, L. Zhang, Y.-H. Fan et al., "Gene expression profiling predicts the development of oral cancer," *Cancer Prevention Research*, vol. 4, no. 2, pp. 218–229, 2011.

[3] N. Kondoh, S. Ohkura, M. Arai et al., "Gene expression signatures that can discriminate oral leukoplakia subtypes and squamous cell carcinoma," *Oral Oncology*, vol. 43, no. 5, pp. 455–462, 2007.

[4] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[5] W. Xiao, Z.-X. Bao, C.-Y. Zhang et al., "Upregulation of miR-31∗ is negatively associated with recurrent/newly formed oral leukoplakia," *PLoS ONE*, vol. 7, no. 6, Article ID e38648, 2012.

[6] T. Hastie, R. Tibshirani, B. Narasimhan, and G. Chu, "Impute: imputation for microarray data," *Bioinformatics*, vol. 17, pp. 520–525, 2001.

[7] T. S. Keshava Prasad, R. Goel, K. Kandasamy et al., "Human protein reference database—2009 update," *Nucleic Acids Research*, vol. 37, no. 1, pp. D767–D772, 2009.

[8] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[9] A. Krek, D. Grün, M. N. Poy et al., "Combinatorial microRNA target predictions," *Nature Genetics*, vol. 37, no. 5, pp. 495–500, 2005.

[10] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, "Human microRNA targets," *PLoS Biology*, vol. 2, no. 11, 2004.

[11] M. Maragkakis, P. Alexiou, G. L. Papadopoulos et al., "Accurate microRNA target prediction correlates with protein repression levels," *BMC Bioinformatics*, vol. 10, article 1471, p. 295, 2009.

[12] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[13] P. Sethupathy, B. Corda, and A. G. Hatzigeorgiou, "TarBase: a comprehensive database of experimentally supported animal microRNA targets," *RNA*, vol. 12, no. 2, pp. 192–197, 2006.

[14] 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.

[15] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.

[16] J. R. Grandis, M. F. Melhem, W. E. Gooding et al., "Levels of TGF-α and EGFR protein in head and neck squamous cell carcinoma and patient survival," *Journal of the National Cancer Institute*, vol. 90, no. 11, pp. 824–832, 1998.

[17] C. Y. Logan and R. Nusse, "The Wnt signaling pathway in development and disease," *Annual Review of Cell and Developmental Biology*, vol. 20, pp. 781–810, 2004.

[18] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, "GeneCards: integrating information about genes, proteins and diseases," *Trends in Genetics*, vol. 13, no. 4, p. 163, 1997.

[19] S.-H. Tan and M. T. Nevalainen, "Signal transducer and activator of transcription 5A/B in prostate and breast cancers," *Endocrine-Related Cancer*, vol. 15, no. 2, pp. 367–390, 2008.

[20] S. Xi, Q. Zhang, W. E. Gooding, T. E. Smithgall, and J. R. Grandis, "Constitutive activation of Stat5b contributes to carcinogenesis in vivo," *Cancer Research*, vol. 63, no. 20, pp. 6763–6771, 2003.

[21] P. Kar and P. C. Supakar, "Expression of Stat5A in tobacco chewing-mediated oral squamous cell carcinoma," *Cancer Letters*, vol. 240, no. 2, pp. 306–311, 2006.

[22] W. M. Ongkeko, X. Altuna, R. A. Weisman, and J. Wang-Rodriguez, "Expression of protein tyrosine kinases in head and neck squamous cell carcinomas," *American Journal of Clinical Pathology*, vol. 124, no. 1, pp. 71–76, 2005.

[23] E. Grabocka and D. Bar-Sagi, "Raf-1 and Squamous Cell Carcinoma: rok-ing the Boat," *Cancer Cell*, vol. 16, no. 2, pp. 85–86, 2009.

[24] P. Koppikar, S.-H. Choi, A. M. Egloff et al., "Combined inhibition of c-Src and epidermal growth factor receptor abrogates growth and invasion of head and neck squamous cell carcinoma," *Clinical Cancer Research*, vol. 14, no. 13, pp. 4284–4291, 2008.

[25] Z.-J. Zhao, L. Peng, F.-Y. Liu, L. Sun, and C.-F. Sun, "PKCα take part in CCR7/NF-κB autocrine signaling loop in CCR7-positive squamous cell carcinoma of head and neck," *Molecular and Cellular Biochemistry*, vol. 357, no. 1-2, pp. 181–187, 2011.

[26] A. Aguzzi, D. Maggioni, G. Nicolini, G. Tredici, R. M. Gaini, and W. Garavello, "MAP kinase modulation in squamous cell carcinoma of the oral cavity," *Anticancer Research*, vol. 29, no. 1, pp. 303–308, 2009.

[27] V. Kirn, I. Zaharieva, S. Heublein et al., "ESR1 promoter methylation in squamous cell cervical cancer," *Anticancer Research*, vol. 34, no. 2, pp. 723–727, 2014.

[28] B. Lewin, A. Siu, C. Baker et al., "Expression of Fyn kinase modulates EMT in oral cancer cells," *Anticancer Research*, vol. 30, no. 7, pp. 2591–2596, 2010.

[29] R. Banerjee, B. S. Henson, N. Russo, A. Tsodikov, and N. J. D'Silva, "Rap1 mediates galanin receptor 2-induced proliferation and survival in squamous cell carcinoma," *Cellular Signalling*, vol. 23, no. 7, pp. 1110–1118, 2011.

[30] C. D. Hu, K. I. Kariya, T. Okada, X. Qi, C. Song, and T. Kataoka, "Effect of phosphorylation on activities of Rap1A to interact with Raf-1 and to suppress Ras-dependent Raf-1 activation," *Journal of Biological Chemistry*, vol. 274, no. 1, pp. 48–51, 1999.

[31] F. Su, A. Viros, C. Milagre et al., "RAS mutations in cutaneous squamous-cell carcinomas in patients treated with BRAF inhibitors," *The New England Journal of Medicine*, vol. 366, no. 3, pp. 207–215, 2012.

[32] R. D. York, H. Yao, T. Dillon et al., "Rap1 mediates sustained MAP kinase activation induced by nerve growth factor," *Nature*, vol. 392, no. 6676, pp. 622–626, 1998.

[33] W. Ren, X. Wang, L. Gao et al., "miR-21 modulates chemosensitivity of tongue squamous cell carcinoma cells to cisplatin by targeting PDCD4," *Molecular and Cellular Biochemistry*, vol. 390, no. 1-2, pp. 253–262, 2014.

[34] P. Roychoudhury, R. R. Paul, R. Chowdhury, and K. Chaudhuri, "HnRNP E2 is downregulated in human oral cancer cells and the overexpression of *hnRNP E2* induces apoptosis," *Molecular Carcinogenesis*, vol. 46, no. 3, pp. 198–207, 2007.

[35] X. Kang, W. Chen, R. H. Kim, M. K. Kang, and N.-H. Park, "Regulation of the hTERT promoter activity by MSH2, the hnRNPs K and D, and GRHL2 in human oral squamous cell carcinoma cells," *Oncogene*, vol. 28, no. 4, pp. 565–574, 2009.

[36] R. A. Katzenellenbogen, P. Vliet-Gregg, M. Xu, and D. A. Galloway, "Cytoplasmic poly(A) binding proteins regulate telomerase activity and cell growth in human papillomavirus type 16 E6-expressing keratinocytes," *Journal of Virology*, vol. 84, no. 24, pp. 12934–12944, 2010.

[37] J. Mork, A. K. Lie, E. Glattre et al., "Human papillomavirus infection as a risk factor for squamous-cell carcinoma of the head and neck," *The New England Journal of Medicine*, vol. 344, no. 15, pp. 1125–1131, 2001.

[38] D. Chen, R. J. Cabay, Y. Jin et al., "MicroRNA deregulations in head and neck squamous cell carcinomas," *Journal of Oral & Maxillofacial Research*, vol. 4, no. 1, article e2, 2013.

[39] W. Gao, H. Shen, L. Liu, J. Xu, and Y. Shu, "MiR-21 overexpression in human primary squamous cell lung carcinoma is associated with poor patient prognosis," *Journal of Cancer Research and Clinical Oncology*, vol. 137, no. 4, pp. 557–566, 2011.

[40] Z. Liu, G. Li, S. Wei et al., "Genetic variants in selected pre-microrna genes and the risk of squamous cell carcinoma of the head and neck," *Cancer*, vol. 116, no. 20, pp. 4753–4760, 2010.

[41] W. Huang, Y. Jin, Y. F. Yuan et al., "Validation and target gene screening of hsa-miR-205 in lung squamous cell carcinoma," *Chinese Medical Journal*, vol. 127, no. 2, pp. 272–278, 2014.

[42] X. Liu, Z.-P. Liu, X.-M. Zhao, and L. Chen, "Identifying disease genes and module biomarkers by differential interactions," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 241–248, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 237969, 15 pages
http://dx.doi.org/10.1155/2015/237969

Research Article

A Heparan Sulfate-Binding Cell Penetrating Peptide for Tumor Targeting and Migration Inhibition

Chien-Jung Chen,[1] Kang-Chiao Tsai,[1] Ping-Hsueh Kuo,[1] Pei-Lin Chang,[1] Wen-Ching Wang,[1,2] Yung-Jen Chuang,[3,4] and Margaret Dah-Tsyr Chang[1,4]

[1] *Institute of Molecular and Cellular Biology, National Tsing Hua University, Hsinchu 30013, Taiwan*
[2] *Biomedical Science and Engineering Center, National Tsing Hua University, Hsinchu 30013, Taiwan*
[3] *Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu 30013, Taiwan*
[4] *Department of Medical Science, National Tsing Hua University, Hsinchu 30013, Taiwan*

Correspondence should be addressed to Margaret Dah-Tsyr Chang; dtchang@life.nthu.edu.tw

Received 18 August 2014; Revised 31 October 2014; Accepted 14 November 2014

Academic Editor: Hao-Teng Chang

As heparan sulfate proteoglycans (HSPGs) are known as co-receptors to interact with numerous growth factors and then modulate downstream biological activities, overexpression of HS/HSPG on cell surface acts as an increasingly reliable prognostic factor in tumor progression. Cell penetrating peptides (CPPs) are short-chain peptides developed as functionalized vectors for delivery approaches of impermeable agents. On cell surface negatively charged HS provides the initial attachment of basic CPPs by electrostatic interaction, leading to multiple cellular effects. Here a functional peptide (CPP*ecp*) has been identified from critical HS binding region in hRNase3, a unique RNase family member with *in vitro* antitumor activity. In this study we analyze a set of HS-binding CPPs derived from natural proteins including CPP*ecp*. In addition to cellular binding and internalization, CPP*ecp* demonstrated multiple functions including strong binding activity to tumor cell surface with higher HS expression, significant inhibitory effects on cancer cell migration, and suppression of angiogenesis *in vitro* and *in vivo*. Moreover, different from conventional highly basic CPPs, CPP*ecp* facilitated magnetic nanoparticle to selectively target tumor site *in vivo*. Therefore, CPP*ecp* could engage its capacity to be developed as biomaterials for diagnostic imaging agent, therapeutic supplement, or functionalized vector for drug delivery.

1. Introduction

Carcinoma is a malignant cancer originating in the ectodermal and endodermal epithelial cells. Interaction between cell surface and microenvironment plays a crucial role in malignant tumor progression. Alterations of cell surface receptor, coreceptor, and adhesive protein expression are reported in various cancer types *in vitro* and *in vivo* [1–3]. Abnormal expression of cell surface molecules notably contributes to enhance tumor cell growth, survival, migration, and invasiveness [4]. Characterization of such alterations and development of novel agent for specific targeting are unmet medical need for early cancer diagnosis.

Glycosaminoglycans (GAGs) including heparan sulfate (HS), chondroitin sulfate (CS), keratan sulfate (KS), or dermatan sulfate (DS) are covalently attached to their core proteins to form proteoglycans. HS proteoglycan (HSPG) present in the extracellular matrix (ECM) provides structural frameworks to mediate cell-cell communication and function in growth factor-receptor binding [5, 6]. HSPGs are key players in modulating tumor progression processes including metastasis, angiogenesis, proliferation, and malignant transformation [4]. Thus, upregulation of cell surface HS may play an active and crucial role in directing malignant phenotype of cancer during different developmental stages.

Cell penetrating peptides (CPPs) are short-chain cationic and/or amphipathic peptides which may be internalized into living cells [7]. CPPs are able to mediate translocation of a conjugated cargo (e.g., anticancer therapeutics) across plasma membrane, providing an effective and nontoxic mechanism

for drug delivery [8]. Most CPPs are rich in positively charged Arg and Lys residues and are internalized after initially interacting with cell surface negatively charged GAGs which cluster CPPs on outer membrane surfaces [9, 10].

CPPs might be potentially used in clinical procedures such as gene therapy and cancer therapy [8, 11]. However, most CPPs are unfeasible for *in vivo* researches due to nonspecificity of their highly cationic characteristics. Cell surface negatively charged HS initializes the contact of CPPs, so particular HS binding CPPs might own mysterious sequence to exert multiple functions including HS binding, cellular binding, lipid binding, and *in vivo* tissue targeting activities. CPP*ecp* is a recently identified CPP not only binding to negatively charged molecules including GAGs and lipids on cell surface *in vitro* but also targeting mucosal tissues *in vivo* [12–14]. In this study, we aim to collect and analyze the characteristics of HS-binding cell penetrating peptides derived from natural proteins. Besides, CPP*ecp* itself falling in this classification has demonstrated multiple functions including *in vitro* tumor binding, tumor migration inhibition and angiogenesis inhibition activities, and *in vivo* cargo delivery to tumor site. Here, we provide more clues for the design of peptide therapeutics or intratumor delivery strategy by linking of a tumor targeting CPP. Furthermore, CPP*ecp* might be a unique HS probe for cancer diagnosis to facilitate the quality of therapeutic index and molecular imaging in translational medicine.

2. Materials and Methods

2.1. Synthetic Peptides. Peptides CPP*ecp* (NYRWRCKNQN) and EDN^{32-41} (NYQRRCKNQN) or CPP*ecp* with *N*-terminally conjugated fluorescein isothiocyanate (FITC) or tetramethylrhodamine (TMR) were synthesized by Genemed Synthesis Inc. and their purities (>90%) were assessed by analytical high-performance liquid chromatography. Peptide sequences were confirmed by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry in Genemed Synthesis Inc.

2.2. Flow Cytometry. Cells (3.0×10^5/well) were added into six-well plates and cultured in the indicated medium. After 24 h, 5 μM FITC-CPP*ecp* dissolved in medium was added into a well and the samples were incubated for 1 h. Cells were then harvested, washed, and suspended in PBS. The fluorescent intensities of the cell samples were measured using a FACSCalibur flow cytometer (BD Biosciences, Franklin Lakes, NJ) and excitation and emission wavelengths of 488 nm and 515–545 nm, respectively. The relative internalization of FITC-CPP*ecp* was reported as the mean fluorescent signal for 10,000 cells.

2.3. Fluorescence Microscopy. CT-26 cells were cultured on coverslips (5.0×10^3/coverslip) in RPMI-1640. After 24 h, cell samples were incubated with FITC or FITC-CPP*ecp* at 37°C for 10 min. Alternatively, CT-26 cells were pretreated with heparinase II (2.5 mU/mL) (Sigma-Aldrich, Missouri, USA) at 37°C for 2 h followed by treatment with 5 μM TMR-CPP*ecp* at 37°C for 10 min. The cells were then washed twice with PBS, fixed with 4% (w/v) paraformaldehyde, and rinsed twice with PBS. The coverslips were mounted in a Vectashield antifade mounting medium with DAPI (Vector Labs). Inverted fluorescent microscopy was performed using Axiovert 135 (Carl Zeiss, Göttingen, Germany) to assess the distribution of the FITC-CPP*ecp* or TMR-ECP*ecp* in the cells.

2.4. In Vitro Cell Migration Assay. Effect of CPP*ecp* on cell migration was assessed using a 24-well transwell plate inserted with incorporating polyethylene terephthalate filter membrane with 8 μm pores (BD FalconTM Cell Culture Insert System).

Approximately 4×10^4 CT-26 cells (obtained from ATCC, number: CRL-2638) were suspended in 200 μL of serum-free RPMI-1640 medium (Sigma-Aldrich, Missouri, USA) and pretreated with 1.25, 2.5, 5, and 12.5 μM CPP*ecp* or EDN^{32-41} at RT for 30 min, and then seeded on the upper compartment of transwell insert membrane. The lower compartment of membrane containing 300 μL 1% FBS (Gibco/Invitrogen) RPMI-1640 medium was used as chemoattractant. After incubating at 37°C and 5% CO_2 for 18 h, the migrated cells on the lower surface of membrane were fixed with 4% formaldehyde for 15 min and stained with 0.05% crystal violet for 20 min. The nonmigrated cells on the upper surface of membrane were removed by cotton swab. Numbers of migrated cells were counted in a randomly selected microscopic field (100x) using inverted microscopy (Olympus CK40, Artisan Technology Group, Mercury Drive Champaign, USA).

Approximately 5×10^4 human umbilical vein endothelial cells (HUVECs) (obtained from BCRC, number: H-UV001) were suspended in 200 μL complete EC medium (Gibco) containing 0, 5, or 12.5 μM CPP*ecp* and then seeded on the upper compartment of filter. The lower compartment of filter contains 500 μL complete EC medium with 20 ng/mL VEGF (R&D) as stimulator. After incubating at 37°C and 5% CO_2 for 4 h, the migrated cells on the lower surface of filter were fixed with 4% formaldehyde at RT for 15 min and stained with Hocechst at RT for 15 min. The nonmigrated cells on the upper surface of filter were removed by cotton swab. Filter membrane of transwell insert was cut down and mounted with Fluoromount mounting medium (Sigma Aldrich, Missouri, USA). Numbers of migrated cells were counted in five randomly selected microscopic fields at magnification 100x using inverted fluorescent microscope (TE2000E, Nikon, Kanagawa, Japan) with a cooled CCD (Evolution VF, MediaCybernetics, Bethesda, MD).

The result was represented as mean ± SD (standard deviation) of three independent experiments. Statistically significant differences were analyzed using unpaired Student's *t*-test. Asterisks showed level of statistical significance: $^*P < 0.05$; $P < 0.01$; $^{***}P < 0.001$ compared with control.

2.5. Zebrafish Angiogenesis Model. Tg(kdr:EGFP) zebrafish, a well-studied model for vascular embryogenesis [15], was performed to assess the effects of CPP*ecp* on angiogenesis. The Tg(kdr:EGFP) (kindly provided by Dr. Yung-Jen

Chuang's lab at NTHU) is a transgenic zebrafish line that expresses eGFP driven by the kdr promoter in vasculature endothelial cells during zebrafish embryogenesis, which can serve as an *in vivo* angiogenesis model for drug screening [16]. Fertilized eggs were generated from adult mating pairs and incubated at 28.5°C in a recirculating aquaculture system. The zebrafish embryos were separately injected with 6.3 or 31.5 ng CPP*ecp* (4.6 nL; 4.56 or 22.8 pmol) into yolk sac at 60 h postfertilization (hpf), and PBS injection was set as control (16–20 zebrafish were used for each treatment condition). After incubating for 24 h, development of subintestinal vessels (SIV) pattern in the zebrafish yolk sac was observed and imaged by inverted fluorescent microscope (TE2000E, Nikon, Kanagawa, Japan) with a cooled CCD (Evolution VF, MediaCybernetics, Bethesda, MD).

2.6. Animal Model. All work performed with animals was approved by the Institutional Animal Care and Use Committee at the National Tsing Hua University. Five-week-old female Balb/c mice (supplied by National Laboratory Animal Center, Taiwan) were housed in laboratory animal room at National Tsing Hua University and allowed to adapt to new surrounding for about seven to fourteen days. Animal rooms had a twelve-to-twelve-hour light-dark/day-night cycle and were maintained at constant temperature and humidity. For establishment of tumor-bearing mouse model, CT-26, a mouse colon carcinoma cell was suspended at a density of 1×10^6 cells in 100 μL PBS containing 50% Matrigel (BD Biosciences, San Jose, CA) and subcutaneously injected into the right back of each mouse. Once subcutaneous tumor volumes grew up to 100 mm^3, all mice were subjected to various treatments. At the end of the experiment, the mice were sacrificed by CO_2 narcosis. All of the organs including kidney, liver, spleen, trachea, lung, intestine, heart, pancreas, stomach, and tumor of these mice were taken, fixed with paraformaldehyde, embedded in paraffin, and sliced into 5 μm tissue slides for Prussian blue staining.

2.7. Magnetic Nanoparticle Conjugated CPPecp and Prussian Blue Staining. To analyze *in vivo* tissue targeting of CPP*ecp*, we have conjugated CPP*ecp* onto a dextran-coated Fe_3O_4 type of magnetic nanoparticle (MNP) to form MNP-conjugated CPP*ecp* (MNP-CPP*ecp*) with a mean diameter of 59.3 nm (kindly provided by MagQu. Co., Ltd.) [17]. CT-26 tumor-bearing mouse was utilized to investigate biodistribution of MNP-CPP*ecp* and Prussian blue staining was employed to demonstrate ferric iron in mouse tissues. The CT-26 tumor-bearing mouse was intravenously injected with 150 μL MNP-CPP*ecp* (0.06 emu/g) and sacrificed by CO_2 narcosis at a time point of 3, 6, 12, and 24 h after administration. The kidney, heart, liver, spleen, stomach, pancreas, small intestine, large intestine, trachea, lung, and tumor of mice were taken, fixed with paraformaldehyde, embedded in paraffin, and sectioned into 5 μm thick tissue slides, following by deparaffinizing in xylene solution (J. T. Baker Phillipsburg, NJ, USA) and serially rehydrating with 100%, 95%, 85%, 75%, and 50% alcohol. The slides were continuously immersed in working solution (20% hydrochloric acid and 10% potassium ferrocyanide (Sigma, MO, USA) solution mixture, 1 : 1 volume ratio) at room temperature for 30 min and then counterstained with fast nuclear red (Sigma, MO, USA) at RT for 5 min. After dehydration through 95% and 100% alcohol and clearing with xylene, each slide was finally covered with coverslip. Tissue images were digitized using light microscope (Eclipse E400, Nikon) with digital microscopy camera (AxioCam ICc 5, ZEISS).

3. Results and Discussion

3.1. Heparan Sulfate Binding Cell Penetrating Peptides Derived from Natural Proteins. Heparan sulfate (HS) serves as the initial anchoring site for many CPPs through electrostatic interactions between negatively charged sulfates or carboxyl groups and basic amino acids Arg as well as Lys [18]. Till now 27 CPPs from natural proteins including 14 viral protein-derived peptides, 7 animal homeostatic modulator-derived peptides, 3 antimicrobial peptides, and 3 toxin-derived peptides have been demonstrated or predicted to be able to interact with cell surface HS and penetrate cross the plasma membrane. *In silico* secondary structures of all 27 HS-binding CPPs were predicted by Network Protein Sequence Analysis [19]. As shown in Table 1, 17 peptides including CPPs 2–6, 8–12, 15, and 18–23 exist as α helix (H). Seven peptides including CPPs 1, 7, 13, 14, 16, 17, and 24 form random coil (C). CPP 23 exists as β sheet (E), and CPPs 26 and 27 exist as mixed α helix (H) with β sheet (E) structures. Among 27 CPPs seventeen structures have been validated by *in vitro* 3D structures deposited in Protein Data Bank (Table 1, underline) [20]. All 14 viral protein-derived CPPs are highly cationic (high pI values) with 10 peptides forming α helix and 4 existing as random coil, penetrating cells through direct translocation [21–24] and lipid raft-mediated endocytosis [25–29]. Most of the 7 animal homeostatic modulator-derived CPPs may be internalized into cytosol through HS-mediated and energy-dependent endocytosis, among which 5 animal protein-derived peptides are demonstrated to possess either α helix or β sheet to interact with the plasma membrane, while our CPP*ecp* and apolipoprotein B binding domain are unique such that they hold random coil structures in this category. As for 3 antimicrobial peptides, all of them are suggested to interact with cell surface HS and penetrate membrane barrier via energy-dependent endocytosis. LL-37 holds high level of α helix, SynB1 possesses β sheet, and SynB3 retains random coil structures [30–32]. For the last category toxin-derived CPPs, bovine prion-derived bPrPp forming α helix and mixed α helix with β strand are distributed in the internal region of venom-derived crotamine, and scorpion toxin-derived maurocalcine [33–36].

Previous researches have shown that the interactions between the positively charged peptide and highly negatively charged membrane components, such as the GAG moieties of cell surface proteoglycans, play a crucial role in the overall process of cellular permeability of highly basic or amphipathic CPPs [37]. Although this investigation may also reflect nonspecific electrostatic interactions between these

Table 1: pI, sequence, and structures of HS-binding cell penetrating peptides.

	Peptide name pI	Sequence and predicted secondary structure*	Heparan sulfate binding region	Internalization mechanism	Ref.
		Viral protein-derived CPP			
1	TAT peptide (49–57) pI: 12.70	**RKKRRQRRR** CCCCCCCCC	RKKRRQRR	Lipid raft-mediated macropinocytosis	[25, 26]
2	Nucleoplasmin NLS (155–170) pI: 11.47	**KRPAAIKKAGQAKKKK** CcHHHHHHHhHHHhCC	Not reported	Not reported	[58]
3	HTLV-II Rex (4–16) pI: 12.85	**TRRQRTRRARRNR** CCCCHHHHCCCCC	TRRQRT	Direct translocation	[21, 22]
4	Lambda-N (48–62) pI: 11.83	**QTRRRERRAEKQAQW** CCHHHHHHHHHHCCC	RRRERR	Not reported	[22]
5	Phi21 N (12–29) pI: 11.45	**TAKTRYKARRAELIAERR** CCCCCCCHHHHHHHHHHH	KTRYKARRA	Not reported	[22]
6	Delta N (1–22) pI: 11.44	**MDAQTRRRERRAEKQAQWKAAN** CCCCHHHHHHHHHHHHHHHHHH	TRRRERRA	Not reported	[22]
7	FHV coat (35–49) pI: 13.00	**RRRRNRTRRNRRRVR** CCCCCCCCCCCCCCC	RRRRNRTRRNRRRVR	Not reported	
8	BMV coat (8–26) pI: 12.78	**KMTRAQRRAAARRNRWTAR** CcCHHHHHHHHHHhccccC	ARRNRW	Not reported	
9	HIV-1 Rev (35–46) pI: 12.85	**RQARRNRRRRWR** CCCCCCCHHHHH	RQARRNRRRRWR	Not reported	[22]
10	Rev (26–42) pI: 12.54	**TRQARRNRRRRWRERQF** CCCCCCCCHHHHHHHHH	TRQARRNRRRRWRERQF	Energy dependent lipid raft-mediated macropinocytosis	[27, 28]
11	CPP from pestivirus envelope glycoprotein (Erns) (194–220) pI: 11.72	**ENARQGAARVTSWLGRQLRIAGKRLEGRSKTWFGAYA** CCCcchHHHHHHHHHHHHHHHHHhhhCCCCccccccC	Basic residues	Direct translocation	[23]
12	gp41 fusion sequence pI: 11.33	**GALFLGWLGAAGSTMGAWSQPKKKRKV** HHHHHHHHHHHHHHHHHHHCCCCCCCCCC	WSQPKKKRKV	Direct translocation	[24]
13	VP22 pI: 12.10	**DAATATRGRSAASRPTERPRAPARSASRPRRPVD** CCCccCCCCCCCCCCCCCCCCCCCCCCCCCCCC	SRPRRP	Energy dependent lipid raft-mediated macropinocytosis	[27, 29]
14	SV40 NLS pI: 11.33	**PKKKRKV** CCCCCCC	PKKKRKV	Not reported	[59, 60]
		Animal homeostatic modulator-derived CPP			
15	Penetratin pI: 12.31	**RQIKIWFQNRRMKWKK** CCCHHHHHHHCCCCCC	NRRMKW	Direct translocation Endocytosis	[61]
16	CPP*ecp* pI: 10.05	**NYRWRCKNQN** CCCCCCCCCC	RWRCK	Macropinocytosis	[12, 62]

Table 1: Continued.

	Peptide name pI	Sequence and predicted secondary structure*	Heparan sulfate binding region	Internalization mechanism	Ref.
17	Apolipoprotein B binding domain pI: 9.82	**SVKAQYKKNSDKHRLMRKRGLK** CCccccCCCCCCCCCCcCCCC	Basic residues	Endocytosis	[63, 64]
18	hCT (9~32) pI: 6.74	**LGTYTQDFNKFHTFPQTAIGVGAP** HHHHHHHHHHHHHHHCHHHHHCCCC	Not reported	Endocytosis	[63, 65]
19	pVEC (615–632) pI: 12.48	**LLIILRRRIRKQAHAHSK** ChhHHHHHHHHHHHhcCC	LRRRIRK	Macropinocytosis and clathrin mediated endocytosis	[66–68]
20	hLF peptide pI: 10.93	**KCFQWQRNMRKVRGPPVSCIKR** CCCchhHHHHhCCCCCcececC	MRKVRG	Lipid raft-mediated endocytosis	[69]
21	PDX-1-PTD pI: 12.31	**RHIKIWFQNRRMKWKK** ChhhhHhhhhhhhhcC	NRRMKWKK	Caveolae-dependent endocytosis and lipid raft-mediated macropinocytosis	[70]
		Antimicrobial peptide			
22	LL-37 (1–37) pI: 10.61	**LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES** HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCC	FRKSKEKI	Endocytosis	[30, 31, 71]
23	SynB1 (1–18) pI: 12.30	**RGGRLSYSRRRFSTSTGR** CCCCEEEEECCEEEEECC	Basic residues	Endocytosis	[32]
24	SynB3 pI: 12.18	**RRLSYSRRRF** CCCCcccCCC	Basic residues	Endocytosis	[32]
		Toxin-derived CPP			
25	bPrPp (1–28) pI: 10.03	**MVKSKIGSWILVLFVAMWSDVGLCKKRP** CCCCCCCCHHHHHHHHHHHHHHHHCCCC	Basic residues	Macropinocytosis	[33]
26	Crotamine (1–42) pI: 9.51	**YKQCHKKGGHCFPKEKICLPPSSDFGKMDCRWRWKCCKKGSG** CCHHHHHCEEEECCCCCCCCCCCCEECCCCCCCCCCEEEECCCC	RWRWK	Endocytosis	[34]
27	Maurocalcine (MCa) (1–33) pI: 9.46	**GDCLPHLKLCKENKDCCSKKCKRRGTNIEKRCR** CCCCCCCCCCCCHHHCCCCEEECCCCCCCCEEE	SKKCKR and EKRCR	Macropinocytosis	[35, 36]

*The confidence of the prediction is denoted by scaling the predictions from week (lower-case letter) to strong (upper-case letter). "H," "E," and "C" refer to α-helical, β-strand, and random coil propensities, respectively.

basic peptides and HS, it has been characterized that negatively charged heparin more effectively blocks uptake of CPPs than other soluble GAGs such as CS and hyaluronic acid [38], likely suggesting that there might be some structural requirements involved in the strong interaction between CPP and HS. In Table 1, 19 of these 27 HS-binding CPPs generally possess conventional heparin binding sequences such as XBBXB and XBBBXXBX where B is a basic amino acid and X represents a random amino acid, and they can also be divided into cationic and amphipathic groups. Most viral factor-derived peptides are basic amino acid-rich. For example, cationic TAT is an extensively used CPP rich in Arg and can interact with sulfated proteoglycans and negatively charged phospholipids on the cell membrane [25]. It should be noted that although 10-amino acid CPP*ecp* is almost equal size to 9-residue TAT and 10-residue SynB3, the features of TAT and SynB3 are quite different from CPP*ecp*. Both TAT derived from viral protein and SynB3 belonging to antimicrobial peptide are highly cationic peptides with high pI values above 12, while our newly identified CPP*ecp* containing only 2 Arg and 1 Lys in a total of 10 amino acids is amphipathic with a pI value of 10.05. Interestingly, the proportion of basic residues in amphipathic crotamine (26%) is close to CPP*ecp* (30%). "RWRCK" motif of CPP*ecp* was previously predicted as a unique functional pattern in all 13 hRNaseA family members employing Reinforced Merging for Unique Segments system (ReMUS) [39]. Another peptide CyLoP-1 (CRWRWKCCKK) derived from crotamine also exhibited efficient intracellular delivery activity. In both cases positively charged residues conducting electrostatic interaction and aromatic Trp exerting transient membrane destabilization were essential to maintain CPP functionality [40, 41]. Taken together, a similar motif "RWRXK" shown on the loop, where X might be a random amino acid, is present in both CPP*ecp* and crotamine, suggesting that combination of positively charged residues and nonpolar aromatic residues, especially Trp, might provide a design rationale for novel amphipathic cell penetrating peptides.

3.2. Cellular Binding of CPPecp to Tumor Cell with Higher HS Expression Level . Heparan sulfate (HS) is reported to be overexpressed in several tumors [42, 43], while HSPG profiles on different tumor cell surface are largely unclear. Here a mouse colon cancer CT-26 cell line was used for *in vitro* and *in vivo* analyses. Cellular binding activity of CPP*ecp* and HS expression level on cell surface of CT-26 cells were accessed for quantitative analysis employing flow cytometry and fluorescent microscopy with fluorescence-labeled CPP*ecp* FITC-CPP*ecp* and an anti-HS monoclonal antibody recognizing an epitope of *N*-sulfated glucosamine on membrane HS (US Biological, Swampscott, MA, USA). Figure 1(a) showed significant FITC-CPP*ecp* binding activity to CT-26 cells, which correlated well with significantly higher HS expression (Figure 1(b)). In addition, 5 μM FITC-CPP*ecp* rapidly and efficiently internalized into CT-26 cells within 10 min as analyzed by fluorescent microscopy (Figure 1(c)). To further address the importance of HS for CPP*ecp* anchor in the absence of autofluorescence background, removal of cell surface HS by heparinase was carried out along with CPP*ecp* labeled with tetramethylrhodamine (TMR). CT-26 cells were incubated in medium with (+) or without (−) heparinase II for 2 h and then treated with 5 μM TMR-CPP*ecp* for 10 min. TMR-CPP*ecp* rapidly and efficiently bound to CT-26 cell surface (Figure 1(d), upper panel), while removal of cellular HS led to significant reduction in CPP*ecp* attachment (Figure 1(d), lower panel). Taken together, our HS-binding CPP*ecp* possessed strong binding activity to tumor cell surface with higher HS expression, while depletion of cell surface HS abolished such highly selective binding activity of CPP*ecp* to tumor cells.

3.3. Effect of CPPecp on Migration of Mouse Colon Carcinoma Cell. It has been shown that HSPGs may modulate cell migration by interacting with growth factors or chemokines and drives cell migrate toward specific stimuli [44]. Since CPP*ecp* with a novel heparin-binding motif in ECP has already been identified to possess high recognition activity to cellular surface HSPG and penetration activity into cells [12], here whether CPP*ecp* might modulate cancer cell migration through interaction with HSPG was further investigated using *in vitro* transwell migration assay, while EDN^{32-41}, a 10-amino acid peptide derived from comparable sequence motif of human RNase2 (EDN), possessing a conventional heparin-binding motif was also analyzed as a control. Figure 2 (black bar) showed that migration activity of CT-26 cell was significantly inhibited by CPP*ecp* in a dose-dependent manner such that it decreased to 83%, 71%, 56%, and 54% upon treatment with 1.25, 2.5, 5, and 12.5 μM CPP*ecp*, respectively. Yet treatment with 1.25, 2.5 and 5 μM EDN^{32-41} could not inhibit migration activity of CT-26 cells, and presence of higher concentration of EDN^{32-41} (12.5 μM) decreased 33% tumor migration (Figure 2, gray bar). These results indicated that CPP*ecp* containing core RWRCK motif, rather than containing known heparin-binding motif, inhibited CT-26 cell migration across the membrane *in vitro*. It has been reported that cancer migration was inhibited by antagonism of HS side chains. For example, A5G27 peptide derived from laminin α5 globular domain recognizes HS side-chains of CD44 variant 3 and blocks bioactivity of fibroblast growth factor-2 (FGF-2). It significantly inhibits FGF-2-induced WiDr colon cancer cell migration and invasion [45]. Collectively, inhibitory effect of CPP*ecp* on cancer cell migration is possibly arisen from interaction with cell surface HS.

3.4. Effects of CPPecp on Migration of Vascular Endothelial Cell. Cell surface HS proteoglycan (HSPG) serves as a coreceptor to coordinate binding of vascular endothelial growth factor (VEGF) toward its receptor. It has been reported to be associated with angiogenesis [46, 47]. However, vascular endothelial cell migration is a crucial step in formation of new blood vessel and tumor angiogenesis [48]. To test the hypothesis that CPP*ecp* interacting with cell surface HSPGs also affected angiogenesis, a common model cell line human umbilical vein endothelial cell (HUVEC) was used for *in vitro* transwell migration assay. Figure 3 indicated that VEGF-induced HUVEC migration was restored by cotreatment

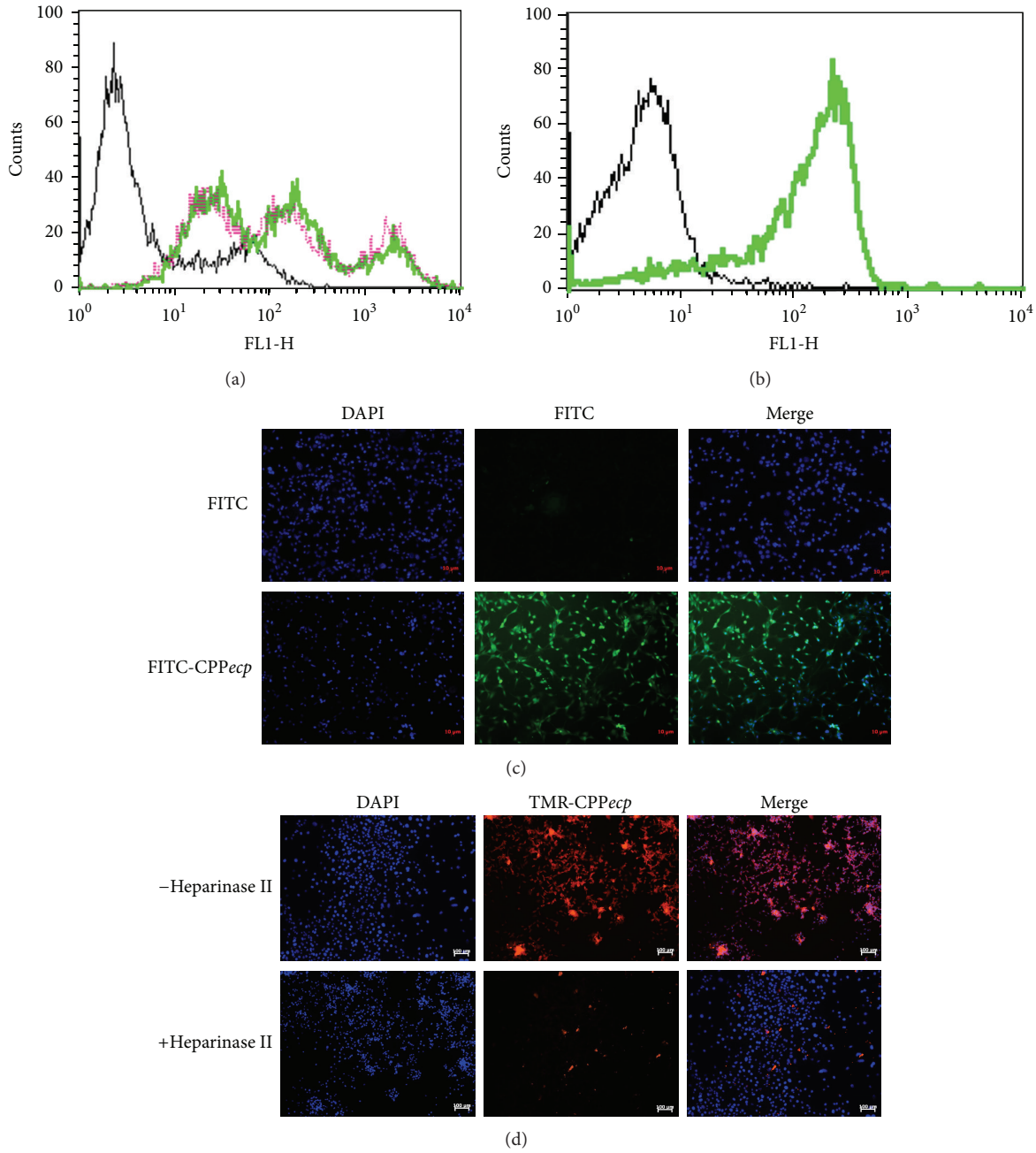


Figure 1: Effect of surface HS level on CPP*ecp* binding to CT-26 cells. (a) CT-26 cells were preincubated at 4°C for 30 min and then incubated with 5 μM FITC-CPP*ecp* for 1 h. The cells were washed twice with 500 μL PBS, trypsinized at 37°C for 15 min, suspended in 500 μL PBS, and subjected to flow cytometry. (b) CT-26 cells were stained with anti-HS monoclonal antibody (10E4) at 4°C for 1 h, washed twice with 500 μL PBS, and hybridized with FITC-conjugated anti-mouse secondary antibody at 4°C for 1 h. After being washed twice with 500 μL PBS, cells were suspended in 500 μL PBS and subjected to flow cytometry. (c) CT-26 cells were treated with 5 μM FITC-CPP*ecp* at 37°C for 10 min. Uptake of FITC-CPP*ecp* by CT-26 cells was examined by fluorescent microscopy. FITC was set as a negative control. DAPI staining of cells indicated intact nucleus. Scale bars in panel represented 10 μm. Green, FITC-labeled CPP*ecp*; blue, DAPI (nucleus). (d) CT-26 cells were pretreated with or without heparinase II (2.5 milliunit/mL) at 37°C for 2 h followed by treatment with 5 μM TMR-CPP*ecp* at 37°C for 10 min. Uptake of TMR-CPP*ecp* by CT-26 cells was examined by fluorescence microscopy. TMR-CPP*ecp* bound on CT-26 tumor cell DAPI staining of cells indicated intact nucleus. Scale bars in panel represented 10 μm. Red, TMR-labeled CPP*ecp*; blue, DAPI (nucleus).

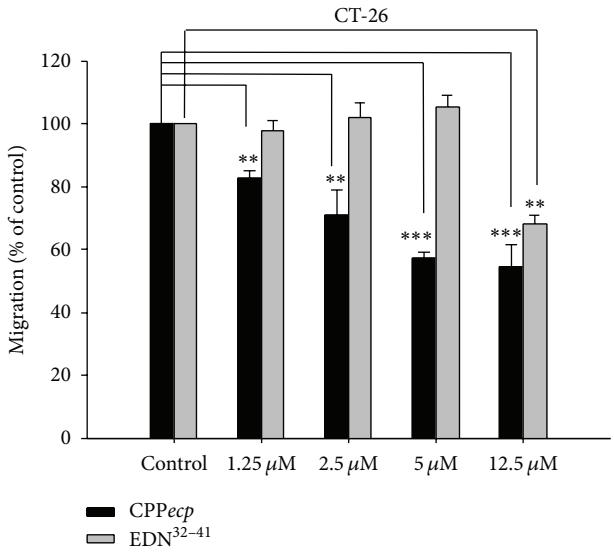


Figure 2: Inhibitory effect of CPP*ecp* on CT-26 cell migration. CT-26 cells were pretreated with CPP*ecp* or EDN^{32-41} at indicated concentration in serum-free RPMI-1640 medium at room temperature for 30 min and then seeded onto the upper side of transwell insert membrane at 37°C for 18 h. Number of migrated cells without CPP*ecp* or EDN^{32-41} treatment was set as 100%. The data represents means ± SD (standard deviation) of three independent experiments. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$ compared with control.

with 5 or 12.5 μM CPP*ecp*, leading to, respectively, 77% and 64% migration activity. This result indicated that CPP*ecp* could inhibit VEGF-induced HUVEC migration. Likewise, the CD44-binding peptide A5G27 derived from laminin α5 globular domain inhibits FGF-induced angiogenesis in Chick CAM Assay [49]. Moreover, an HS-binding peptide 6a-P, corresponding to the HSPG binding domain of VEGF, binds to HSPG and affects interaction between VEGF and HSPG [50]. It interferes with angiogenesis by inhibiting VEGF-induced HUVEC migration and binding of VEGF to HUVEC. As a result, involvement of our CPP*ecp* in angiogenesis may be attributed to interaction with cell surface HSPG.

3.5. Effects of CPPecp on Angiogenesis during Embryonic Development of Zebrafish. Cell surface HSPGs serve as a coreceptor to coordinate binding of VEGF toward its receptor and have been reported to be associated with angiogenesis [46, 47]. Tg(kdr:EGFP) zebrafish, a well-studied model for vascular embryogenesis, has been used as model for drug screening and angiogenesis studies [51, 52]. It was thus utilized to investigate CPP*ecp* effects on *in vivo* embryonic angiogenesis by injecting 4.6 nL of 4.56 or 22.8 pmol CPP*ecp* or PBS (control) into yolk sac of zebrafish at 60 h postfertilization (hpf), and the development of subintestinal vessel (SIV) pattern (Figure 4(a), SIV networks are indicated with red rectangle) at 24 h postinjection (hpi) was monitored with images by inverted fluorescent microscope. Here 16–20 zebrafish were tested for each treatment group. The observed SIV patterns of zebrafish were divided into three groups according to growth level of SIV: normal, mildly inhibited, and severely inhibited phenotypes (Figure 4(b)). In the normal phenotype SIV developed as smooth basket-like pattern with 5-6 arcades. Both mild and severe inhibition phenotypes could be further classified as ectopic SIV pattern, in which SIV exhibited tortuous network and was unable to demonstrate complete basket-like pattern that normal phenotype developed. However, severe inhibition phenotype displayed more incomplete SIV network than mild inhibition phenotype did. In contrast, the zebrafish injected with CPP*ecp* appeared to be tortuous, in which SIV pattern shrank significantly as compared with that of PBS control (Figure 4(c)). Figure 4(d) illustrated quantitative analysis data in which percentage of ectopic SIV phenotype (mildly inhibited phenotype plus severely inhibited phenotype) rose from 39.6% up to 49.2% and 52.6% upon injection with 4.56 and 22.8 pmol CPP*ecp*, respectively. Moreover, severely inhibited SIV phenotype increased from 11.1% up to 26.2% and 32.4% upon injection with 4.56 and 22.8 pmol, respectively. In other words, percentage of severely inhibited phenotype in ectopic phenotype of zebrafish increased from 27.3% (control) up to 52.3% and 60.7% upon injection with 4.56 and 22.8 pmol CPP*ecp*, respectively (Figure 4(e)). These data revealed that our CPP*ecp* possessed antiangiogenesis activity in inhibiting SIV growth of zebrafish. As a result, involvement of CPP*ecp* in angiogenesis may be attributed to interaction with cell surface HS. CPP*ecp* is the first antiangiogenic peptide deciphered in embryonic development of zebrafish.

3.6. Time-Dependent Biodistribution of MNP-CPPecp in CT-26 Tumor-Bearing Mouse. To better understand biodistribution of our HS-binding CPP*ecp* *in vivo*, CPP*ecp* was conjugated with well-dispersed Fe_3O_4 magnetic beads (59.3 nm for diameter) to form magnetic nanoparticle-conjugated CPP*ecp* (MNP-CPP*ecp*). CT-26 tumor-bearing mice were intravenously injected with MNP-CPP*ecp* (0.06 emu/g) and sacrificed at different time point after administration (Figure 5(a)). MNP-CPP*ecp* signal was detected using Prussian blue staining to indicate ferric iron in tissue section (blue color). Figure 5(b) indicated that stainable ferric iron (blue color as indicated by yellow arrow) was barely detectable in trachea, heart, and large intestine at all indicated time points, and so did other tissues including stomach, pancreas and kidney (data not shown). The MNP-CPP*ecp* mainly accumulated in liver tissues from 3 h up to 24 h owing to uptake and removal by macrophages of reticuloendothelial system, which played a role in clearance of external substance in liver [53, 54]. Interestingly, Prussian blue staining signals in CT-26 tumor section suggested MNP-CPP*ecp* accumulation from 12 h to 24 h, whereas MNP signal was only detected in liver at 24 h. One recent report showed that exendin-4 peptide-conjugated superparamagnetic iron oxide nanoparticles were inevitably accumulated in liver tissue, suggesting that a nanoparticle might unavoidably be captured by this metabolic organ [55]. However, it is worth noting that CPP*ecp* has potential to target colon carcinoma *in vivo*, suggesting

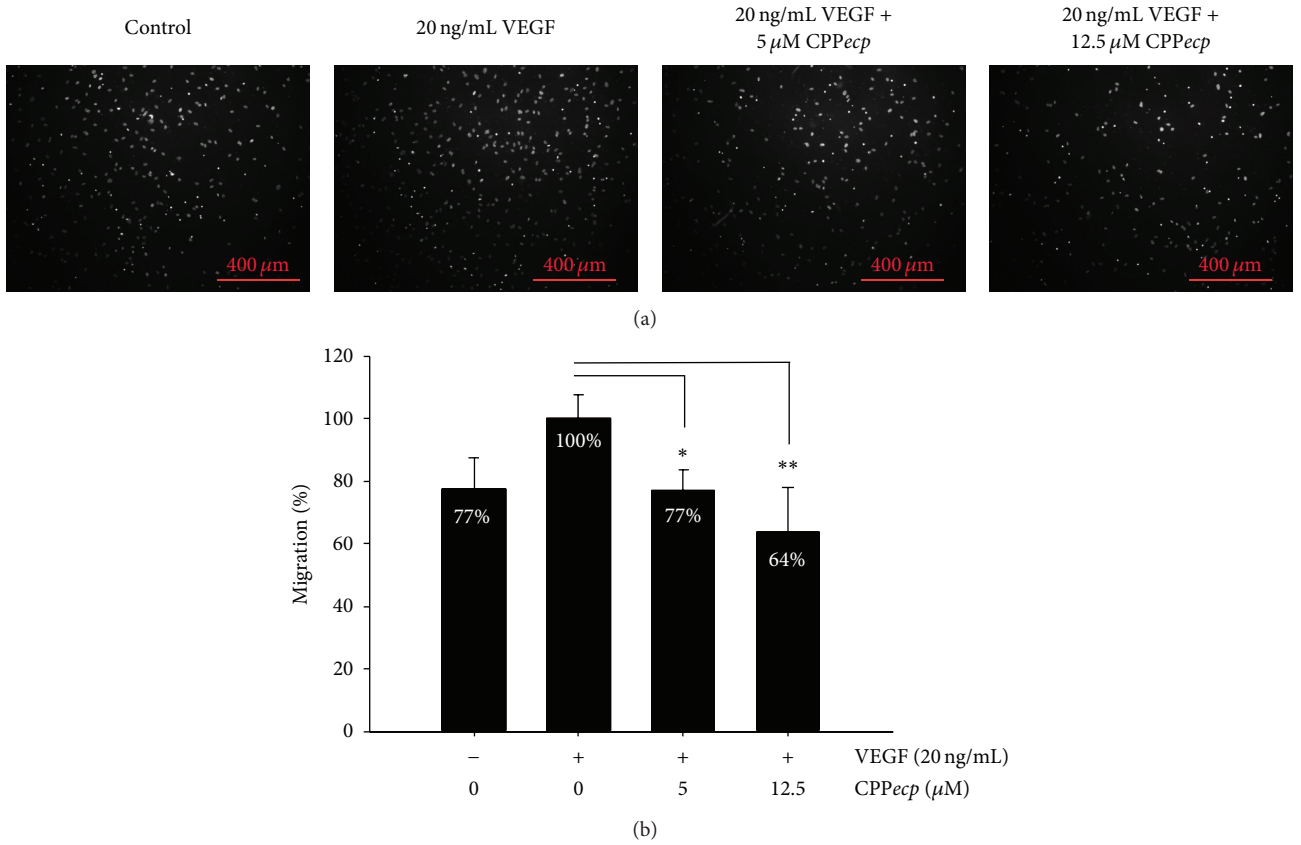


Figure 3: Inhibitory effect of CPP*ecp* on HUVEC migration. HUVECs were seeded onto the upper side of transwell insert membrane containing CPP*ecp* at indicated concentration at 37°C for 4 h. The lower side of transwell was filled with complete EC medium supplementing with 20 ng/mL VEGF. Migrated cells on the lower surface of transwell insert membrane were stained with Hoechst (a). Percentage of migrated cells in the presence of VEGF and set as 100% (positive control). Alternation of HUVEC migration activity in the presence of VEGF and various concentrations of CPP*ecp* were quantified as compared with positive control (b). The data represents means ± SD (standard deviation) of three independent experiments. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$ compared with control. Magnification: 100x. Scale bar: 400 μm.

that CPP*ecp* might be applied for a potent carrier for drug delivery.

3.7. Heparan Sulfate-Binding Cell Penetrating Peptide for Tumor Targeted Strategy. Although CPPs as noninvasive agents have promising biomedical potential for molecular delivery, they are mostly unfeasible for *in vivo* researches due to nonspecificity of their highly cationic characteristic such as TAT peptide. Due to high uptake rates *in vitro* and relatively low specificity *in vivo* of most CPPs, conventional CPPs would be designed for topical applications in CPP-based delivery (Table 1). Further analysis of natural protein-derived CPPs revealed that 5 CPPs exerted *in vitro* tumor suppression as well as cell internalization activities (Table 2). Although TAT peptide (46–57) demonstrated antiangiogenesis and apoptosis-inducing activities, TAT peptide was proved to show low target specificity *in vivo* [56]. Distinct from conventional highly cationic CPPs, 4 amphipathic CPPs including CPP*ecp*, crotamine, NFL-TBS (40–63), and p28 peptides demonstrated unique tumor targeting activity *in vivo*. Even though specific protein receptors for CPP*ecp* and crotamine remain to be investigated, HSPG acting as coreceptor is indispensable for the translocation of CPP*ecp* and crotamine [12, 34]. In addition, both CPP*ecp* and crotamine targeted highly proliferating cells such as tumor tissues [14, 57]. Interestingly, a motif decorating a hydrophilic aromatic amino acid participating in membrane permeation between two arginines (RWR) appeared to be conserved in both CPP*ecp* and crotamine, leading to similar characteristics of these 2 multifunctional HS-binding CPPs. Therefore, amphipathic CPPs might own promising potential to be designed as peptide-based drugs. In particular, HS-binding CPPs are suitable drug carriers for *in vivo* application in delivery of functional therapeutics.

4. Conclusions

CT-26 colon tumor cells revealed high CPP*ecp* binding activity due to high HSPG expression on cell surface. CPP*ecp* displays not only significantly inhibitory effects on CT-26 cancer migration and angiogenesis *in vitro* but also antiangiogenesis activity during zebrafish embryogenesis *in vivo*. Moreover, covalent linkage of CPP*ecp* to magnetic nanoparticle shows potential for *in vivo* targeting to a subcutaneous CT-26

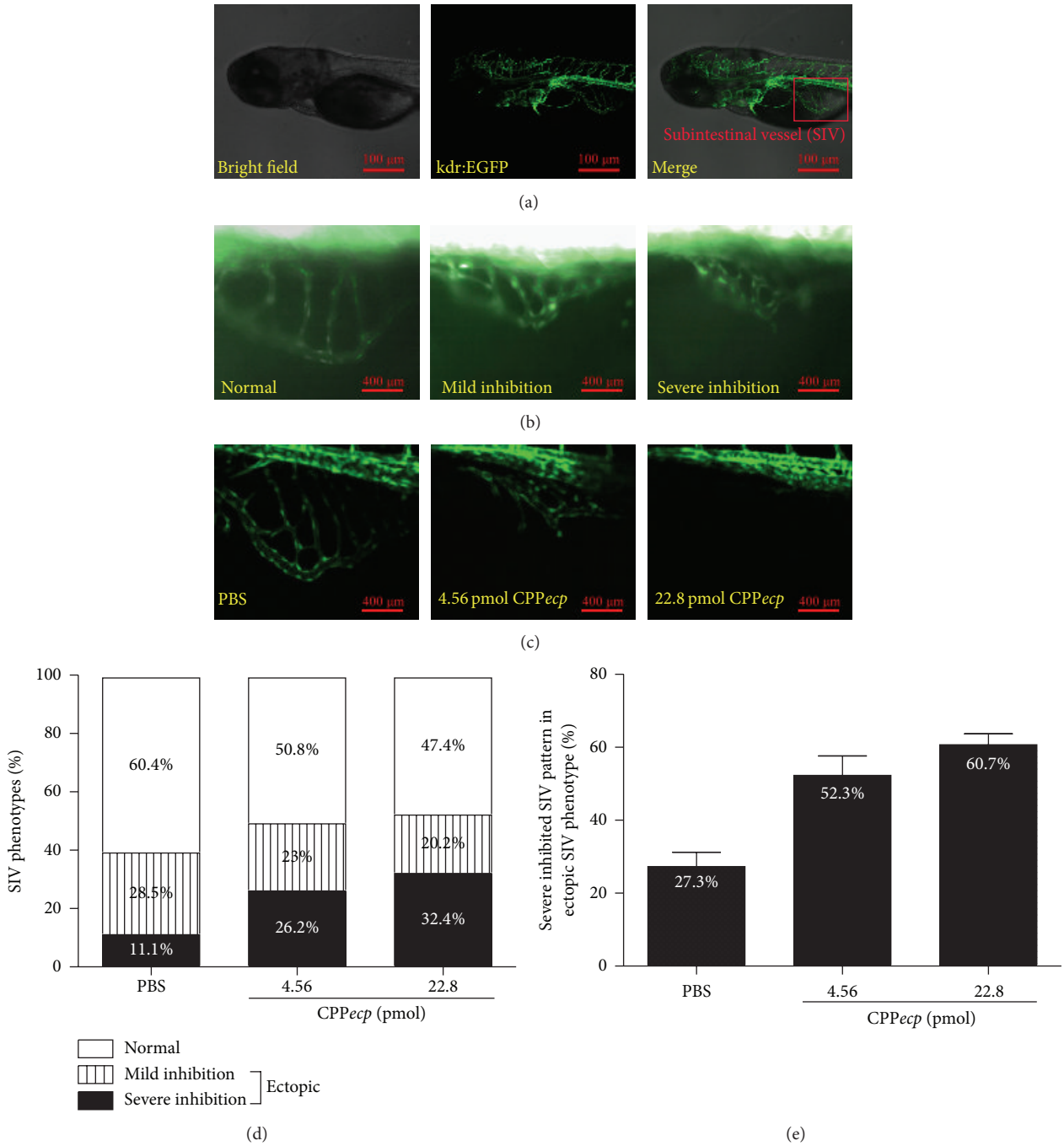
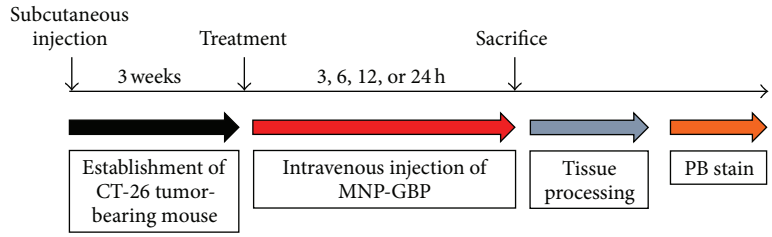
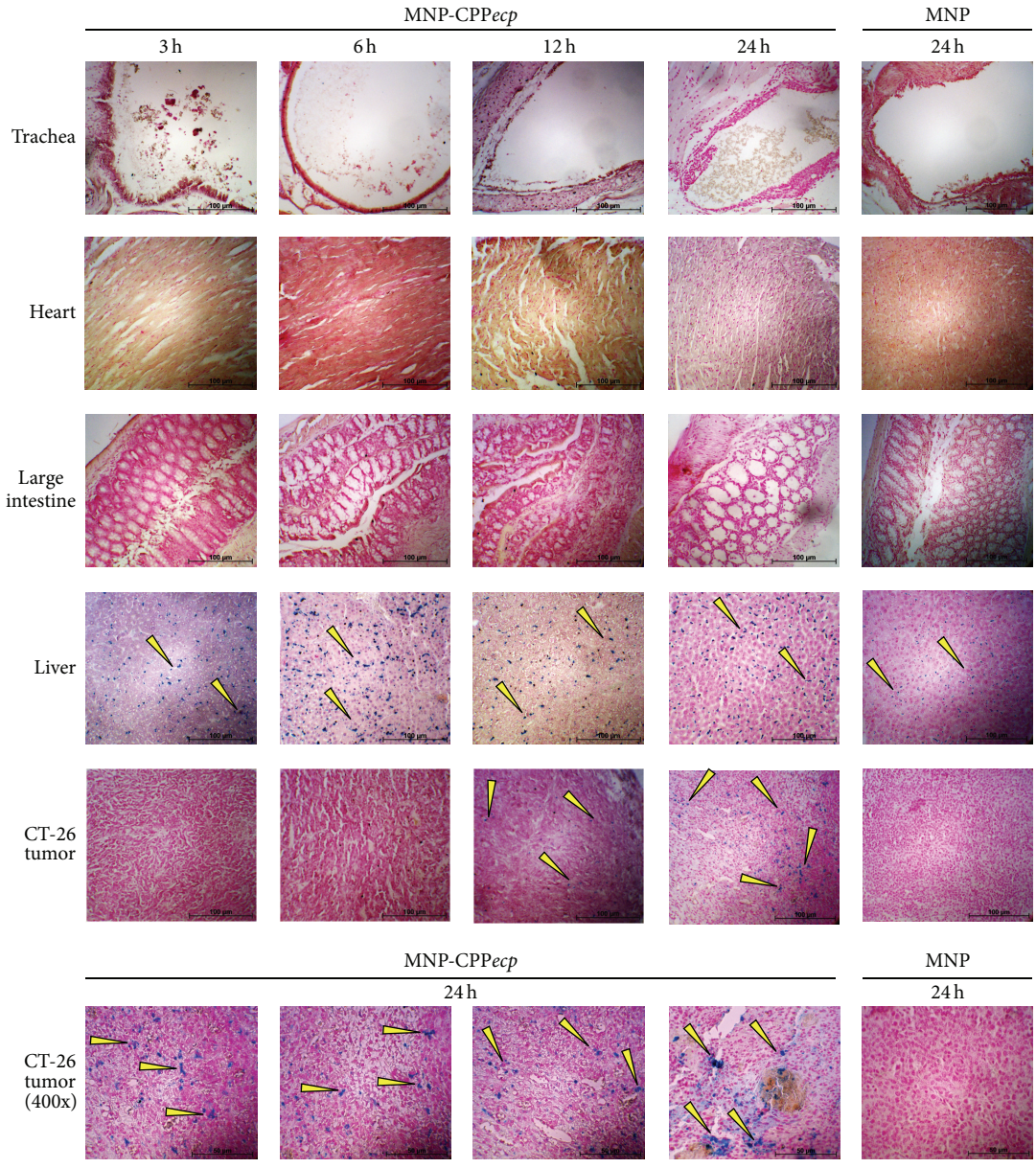


FIGURE 4: Inhibitory effect of CPP*ecp* on angiogenesis in Tg(kdr:EGFP) zebrafish. (a) Morphology and green-labeled vessels in Tg(kdr:EGFP) zebrafish. The red rectangle represents the area of subintestinal vessel (SIV) network. Magnification: 100x. Scale bar: 400 μm. (b) Development of SIV network in the zebrafish yolk sac could be classified into three groups: normal, mild inhibition, and severe inhibition pattern. Magnification: 400x. Scale bar: 100 μm. (c) Development of SIV network in the zebrafish yolk sac at 24 h postinjection (hpi). Magnification: 400x. Scale bar: 100 μm. (d) Percentage of different SIV phenotypes in the zebrafish yolk sac at 24 hpi. (e) Percentage of severe inhibited SIV phenotype in ectopic SIV phenotype in the zebrafish yolk sac at 24 hpi. 16–20 zebrafish were used for each treatment group. The data represents means ± SD (standard deviation) of three independent experiments.



(a)



(b)

Figure 5: Localization of MNP-CPP*ecp* in CT-26 tumor-bearing mouse. To investigate the biodistribution of CPP*ecp* *in vivo*, CT-26 tumor-bearing mice were intravenously injected with 0.06 emu/g MNP-CPP*ecp* and sacrificed at a time point of 3, 6, 12, and 24 h after injection (a). Signal of MNP-CPP*ecp* was visualized using Prussian blue staining to indicate ferric iron in tissue section (blue color, yellow arrow). Represented staining patterns of trachea, heart, large intestine, liver, and CT-26 tumor were shown (b). MNP injection was set as negative control. Nuclear fast red staining was set as counterstain (red color). Magnification: 200x and 400x. Scale bar: 100 μm and 50 μm.

Table 2: Multifunctional CPPs for tumor suppression.

Name/sequence	Function	Mechanism	Cell line	Tumor mouse model	Ref.
CPP*ecp*/**NYRWRCKNQN**	Cell penetrating HS binding Antimigration Antiangiogenesis Tumor targeting	Block putative HS coreceptor for growth factor	CT-26 HUVEC	Murine colon carcinoma CT-26	[12–14]
Crotamine/**YKQCHKKGGHCFPKEKICLPPSSDFGKMDCRWRWKCCKKGSG**	Cell penetrating HS binding Antiproliferation Tumor targeting	Interact with lysosomes to trigger intracellular Ca^{2+} transients and alter mitochondrial membrane potential	B16F10 CHO-K1	Murine melanoma (B16F10) Murine mammary carcinoma (TS/A-pc, TS/A-pc-pGL3)	[34, 57]
NFL-TBS. (40–63)/**YSSYSAPVSSSLSVRRSYSSSSGS**	Cell penetrating Antimigration Antiproliferation Apoptosis-inducing Antitumor growth	Inhibit polymerization of microtubules	Human glioblastoma (T98G) Rat glioblastoma (F98) Rat gliosarcoma (9L)	Murine glioblastoma (F98)	[72, 73]
TAT peptide (46–57)/**SYGRKKRRQRRR**	Cell penetrating HS binding Antiangiogenesis Apoptosis-inducing	Inhibit VEGF binding to HUVEC and inhibit phosphorylation of ERK	HUVEC	×	[25, 74]
p28/**LSTAADMQGVVTDGMASGLDKDYLKPDD**	Cell penetrating Antiangiogenesis Antitumor growth	Inhibit phosphorylation of VEGFR-2, FAK, and Akt	HUVEC	Human melanoma (UISO-Mel-6)	[75]

tumor site. Moreover, CPP*ecp* containing a core RWRXK sequence demonstrates both cell penetrating and epithelial tumor targeting activities. Taken together, our HS-binding CPP*ecp* might be feasible for further application in molecular imaging for tumor homing and selectively targeting drug delivery system.

Conflict of Interests

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

Authors' Contribution

Chien-Jung Chen and Kang-Chiao Tsai contributed equally to this work.

Acknowledgments

The authors thank Drs. Ching-Chuan Kuo, Wun-Shaing Wayne Chang, and Yu-Ting Chou for critical comments. This work was supported by National Tsing Hua University (NTHUIOON705IEI and NTHUIOIN205IEI) and National Science Council (NSC101-2622-B-007-001-CC1 and NSC103-2325-B-007-002) to M. D.-T. Chang. C.-J. Chen and P.-H. Kuo are awarded a scholarship sponsored by Apex Biotechnology Corporation, Taiwan. K.-C. Tsai is supported by Graduate Program of Biotechnology in Medicine sponsored by National Tsing Hua University and National Health Research Institute in Taiwan.

References

[1] K. R. Kampen, "The mechanisms that regulate the localization and overexpression of VEGF receptor-2 are promising therapeutic targets in cancer biology," *Anti-Cancer Drugs*, vol. 23, no. 4, pp. 347–354, 2012.

[2] R. Sasisekharan, Z. Shriver, G. Venkataraman, and U. Narayanasami, "Roles of heparan-sulphate glycosaminoglycans in cancer," *Nature Reviews Cancer*, vol. 2, no. 7, pp. 521–528, 2002.

[3] M. Adachi, T. Taki, M. Higashiyama, N. Kohno, H. Inufusa, and M. Miyake, "Significance of integrin α5 gene expression as a prognostic factor in node-negative non-small cell lung cancer," *Clinical Cancer Research*, vol. 6, no. 1, pp. 96–101, 2000.

[4] E. H. Knelson, J. C. Nee, and G. C. Blobe, "Heparan sulfate signaling in cancer," *Trends in Biochemical Sciences*, vol. 39, no. 6, pp. 277–288, 2014.

[5] A. Ori, M. C. Wilkinson, and D. G. Fernig, "A systems biology approach for the investigation of the heparin/heparan sulfate interactome," *The Journal of Biological Chemistry*, vol. 286, no. 22, pp. 19892–19904, 2011.

[6] N. Sasaki, N. Higashi, T. Taka, M. Nakajima, and T. Irimura, "Cell surface localization of heparanase on macrophages regulates degradation of extracellular matrix heparan sulfate," *The Journal of Immunology*, vol. 172, no. 6, pp. 3830–3835, 2004.

[7] J. Regberg, A. Srimanee, and Ü. Langel, "Applications of cell-penetrating peptides for tumor targeting and future cancer therapies," *Pharmaceuticals*, vol. 5, no. 9, pp. 991–1007, 2012.

[8] V. Kersemans, K. Kersemans, and B. Cornelissen, "Cell penetrating peptides for in vivo molecular imaging applications," *Current Pharmaceutical Design*, vol. 14, no. 24, pp. 2415–2427, 2008.

[9] S. Console, C. Marty, C. García-Echeverría, R. Schwendener, and K. Ballmer-Hofer, "Antennapedia and HIV transactivator of transcription (TAT) "protein transduction domains" promote endocytosis of high molecular weight cargo upon binding to cell surface glycosaminoglycans," *The Journal of Biological Chemistry*, vol. 278, no. 37, pp. 35109–35114, 2003.

[10] S. Deshayes, T. Plénat, P. Charnet, G. Divita, G. Molle, and F. Heitz, "Formation of transmembrane ionic channels of primary amphipathic cell-penetrating peptides. Consequences on the mechanism of cell penetration," *Biochimica et Biophysica Acta*, vol. 1758, no. 11, pp. 1846–1851, 2006.

[11] B. G. Bitler and J. A. Schroeder, "Anti-cancer therapies that utilize cell penetrating peptides," *Recent Patents on Anti-Cancer Drug Discovery*, vol. 5, no. 2, pp. 99–108, 2010.

[12] S.-L. Fang, T.-C. Fan, H.-W. Fu et al., "A novel cell-penetrating peptide derived from human eosinophil cationic protein," *PLoS ONE*, vol. 8, no. 3, Article ID e57318, 2013.

[13] P.-C. Lien, P.-H. Kuo, C.-J. Chen et al., "*In silico* prediction and *in vitro* characterization of multifunctional human RNase3," *BioMed Research International*, vol. 2013, Article ID 170398, 12 pages, 2013.

[14] C.-J. Chen, P.-H. Kuo, T.-J. Hung et al., "In vitro characterization and in vivo application of a dual functional peptide," in *Proceedings of the 7th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS '13)*, pp. 576–581, Taichung, Taiwan, July 2013.

[15] K. R. Kidd and B. M. Weinstein, "Fishing for novel angiogenic therapies," *British Journal of Pharmacology*, vol. 140, no. 4, pp. 585–594, 2003.

[16] G. N. Serbedzija, E. Flynn, and C. E. Willett, "Zebrafish angiogenesis: a new model for drug screening," *Angiogenesis*, vol. 3, no. 4, pp. 353–359, 1999.

[17] S. Y. Yang, J. L. Wu, C. H. Tso et al., "A novel quantitative immunomagnetic reduction assay for Nervous necrosis virus," *Journal of Veterinary Diagnostic Investigation*, vol. 24, no. 5, pp. 911–917, 2012.

[18] I. Capila and R. J. Linhardt, "Heparin-protein interactions," *Angewandte Chemie—International Edition*, vol. 41, no. 3, pp. 391–412, 2002.

[19] C. Combet, C. Blanchet, C. Geourjon, and G. Deléage, "NPS@: network protein sequence analysis," *Trends in Biochemical Sciences*, vol. 25, no. 3, pp. 147–150, 2000.

[20] J. L. Sussman, E. E. Abola, D. Lin, J. Jiang, N. O. Manning, and J. Prilusky, "The protein data bank: Bridging the gap between the sequence and 3D structure world," *Genetica*, vol. 106, no. 1-2, pp. 149–158, 1999.

[21] Y. G. Choi and A. L. N. Rao, "Molecular studies on bromovirus capsid protein: VII. Selective packaging of BMV RNA4 by specific N-terminal arginine residues," *Virology*, vol. 275, no. 1, pp. 207–217, 2000.

[22] S. Futaki, T. Suzuki, W. Ohashi et al., "Arginine-rich peptides. An abundant source of membrane-permeable peptides having potential as carriers for intracellular protein delivery," *Journal of Biological Chemistry*, vol. 276, no. 8, pp. 5836–5840, 2001.

[23] J. P. M. Langedijk, "Translocation activity of C-terminal domain of pestivirus Erns and ribotoxin L3 loop," *Journal of Biological Chemistry*, vol. 277, no. 7, pp. 5308–5314, 2002.

[24] L. Chaloin, P. Vidal, P. Lory et al., "Design of carrier peptide-oligonucleotide conjugates with rapid membrane translocation

and nuclear localization properties," *Biochemical and Biophysical Research Communications*, vol. 243, no. 2, pp. 601–608, 1998.

[25] E. Vivès, P. Brodin, and B. Lebleu, "A truncated HIV-1 Tat protein basic domain rapidly translocates through the plasma membrane and accumulates in the cell nucleus," *Journal of Biological Chemistry*, vol. 272, no. 25, pp. 16010–16017, 1997.

[26] J. S. Wadia, R. V. Stan, and S. F. Dowdy, "Transducible TAT-HA fusogenic peptide enhances escape of TAT-fusion proteins after lipid raft macropinocytosis," *Nature Medicine*, vol. 10, no. 3, pp. 310–315, 2004.

[27] T. Sugita, T. Yoshikawa, Y. Mukai et al., "Comparative study on transduction and toxicity of protein transduction domains," *British Journal of Pharmacology*, vol. 153, no. 6, pp. 1143–1152, 2008.

[28] S. Kameyama, M. Horie, T. Kikuchi et al., "Acid wash in determining cellular uptake of fab/cell-permeating peptide conjugates," *Biopolymers*, vol. 88, no. 2, pp. 98–107, 2007.

[29] G. Elliott and P. O'Hare, "Intercellular trafficking and protein delivery by a herpesvirus structural protein," *Cell*, vol. 88, no. 2, pp. 223–233, 1997.

[30] F. J. Byfield, Q. Wen, K. Leszczyńska et al., "Cathelicidin LL-37 peptide regulates endothelial cell stiffness and endothelial barrier permeability," *The American Journal of Physiology—Cell Physiology*, vol. 300, no. 1, pp. C105–C112, 2011.

[31] S. Pochet, S. Tandel, S. Querriére et al., "Modulation by LL-37 of the responses of salivary glands to purinergic agonists," *Molecular Pharmacology*, vol. 69, no. 6, pp. 2037–2046, 2006.

[32] G. Drin, S. Cottin, E. Blanc, A. R. Rees, and J. Temsamani, "Studies on the internalization mechanism of cationic cell-penetrating peptides," *The Journal of Biological Chemistry*, vol. 278, no. 33, pp. 31192–31201, 2003.

[33] M. Magzoub, S. Sandgren, P. Lundberg et al., "N-terminal peptides from unprocessed prion proteins enter cells by macropinocytosis," *Biochemical and Biophysical Research Communications*, vol. 348, no. 2, pp. 379–385, 2006.

[34] F. D. Nascimento, M. A. F. Hayashi, A. Kerkis et al., "Crotamine mediates gene delivery into cells through the binding to heparan sulfate proteoglycans," *The Journal of Biological Chemistry*, vol. 282, no. 29, pp. 21349–21360, 2007.

[35] Z. Fajloun, R. Kharrat, L. Chen et al., "Chemical synthesis and characterization of maurocalcine, a scorpion toxin that activates Ca^{2+} release channel/ryanodine receptors," *FEBS Letters*, vol. 469, no. 2-3, pp. 179–185, 2000.

[36] A. Mosbah, R. Kharrat, Z. Fajloun et al., "A new fold in the scorpion toxin family, associated with an activity on a ryanodine-sensitive calcium channel," *Proteins*, vol. 40, no. 3, pp. 436–442, 2000.

[37] E. Vives, "Cellular utake of the Tat peptide: an endocytosis mechanism following ionic interactions," *Journal of Molecular Recognition*, vol. 16, no. 5, pp. 265–271, 2003.

[38] M. Tyagi, M. Rusnati, M. Presta, and M. Giacca, "Internalization of HIV-1 tat requires cell surface heparan sulfate proteoglycans," *The Journal of Biological Chemistry*, vol. 276, no. 5, pp. 3254–3261, 2001.

[39] T.-W. Pai, M. D.-T. Chang, W.-S. Tzou et al., "REMUS: a tool for identification of unique peptide segments as epitopes," *Nucleic Acids Research*, vol. 34, pp. W198–W201, 2006.

[40] D. Jha, R. Mishra, S. Gottschalk et al., "CyLoP-1: a novel cysteine-rich cell-penetrating peptide for cytosolic delivery of cargoes," *Bioconjugate Chemistry*, vol. 22, no. 3, pp. 319–328, 2011.

[41] G. Dom, C. Shaw-Jackson, C. Matis et al., "Cellular uptake of Antennapedia Penetratin peptides is a two-step process in which phase transfer precedes a tryptophan-dependent translocation," *Nucleic Acids Research*, vol. 31, no. 2, pp. 556–561, 2003.

[42] H. Park, Y. Kim, Y. Lim, I. Han, and E.-S. Oh, "Syndecan-2 mediates adhesion and proliferation of colon carcinoma cells," *The Journal of Biological Chemistry*, vol. 277, no. 33, pp. 29730–29736, 2002.

[43] K. Nackaerts, E. Verbeken, G. Deneffe, B. Vanderschueren, M. Demedts, and G. David, "Heparan sulfate proteoglycan expression in human lung-cancer cells," *International Journal of Cancer*, vol. 74, no. 3, pp. 335–345, 1997.

[44] R. D. Sanderson, "Heparan sulfate proteoglycans in invasion and metastasis," *Seminars in Cell and Developmental Biology*, vol. 12, no. 2, pp. 89–98, 2001.

[45] S. Hibino, M. Shibuya, M. P. Hoffman et al., "Laminin α5 chain metastasis- and angiogenesis-inhibiting peptide blocks fibroblast growth factor 2 activity by binding to the heparan sulfate chains of CD44," *Cancer Research*, vol. 65, no. 22, pp. 10494–10501, 2005.

[46] M. M. Fuster, L. Wang, J. Castagnola et al., "Genetic alteration of endothelial heparan sulfate selectively inhibits tumor angiogenesis," *The Journal of Cell Biology*, vol. 177, no. 3, pp. 539–549, 2007.

[47] L. Jakobsson, J. Kreuger, K. Holmborn et al., "Heparan sulfate in trans potentiates VEGFR-mediated angiogenesis," *Developmental Cell*, vol. 10, no. 5, pp. 625–634, 2006.

[48] L. Lamalice, F. Le Boeuf, and J. Huot, "Endothelial cell migration during angiogenesis," *Circulation Research*, vol. 100, no. 6, pp. 782–794, 2007.

[49] S. Hibino, M. Shibuya, J. A. Engbring, M. Mochizuki, M. Nomizu, and H. K. Kleinman, "Identification of an active site on the laminin α5 chain globular domain that binds to CD44 and inhibits malignancy," *Cancer Research*, vol. 64, no. 14, pp. 4810–4816, 2004.

[50] T.-Y. Lee, J. Folkman, and K. Javaherian, "HSPG-Binding peptide corresponding to the exon 6a-encoded domain of VEGF inhibits tumor growth by blocking angiogenesis in Murine model," *PLoS ONE*, vol. 5, no. 4, Article ID e9945, 2010.

[51] S. Nicoli, G. De Sena, and M. Presta, "Fibroblast growth factor 2-induced angiogenesis in zebrafish: the zebrafish yolk membrane (ZFYM) angiogenesis assay," *Journal of Cellular and Molecular Medicine*, vol. 13, no. 8, pp. 2061–2068, 2009.

[52] M.-W. Kuo, C.-H. Wang, H.-C. Wu, S.-J. Chang, and Y.-J. Chuang, "Soluble THSD7A is an N-glycoprotein that promotes endothelial cell migration and tube formation in angiogenesis," *PLoS ONE*, vol. 6, no. 12, Article ID e29000, 2011.

[53] R. Kumar, I. Roy, T. Y. Ohulchanskky et al., "*In vivo* biodistribution and clearance studies using multimodal organically modified silica nanoparticles," *ACS Nano*, vol. 4, no. 2, pp. 699–708, 2010.

[54] S. Nagayama, K.-I. Ogawara, Y. Fukuoka, K. Higaki, and T. Kimura, "Time-dependent changes in opsonin amount associated on nanoparticles alter their hepatic uptake characteristics," *International Journal of Pharmaceutics*, vol. 342, no. 1-2, pp. 215–221, 2007.

[55] B. Zhang, B. Yang, C. Zhai, B. Jiang, and Y. Wu, "The role of exendin-4-conjugated superparamagnetic iron oxide nanoparticles in beta-cell-targeted MRI," *Biomaterials*, vol. 34, no. 23, pp. 5843–5852, 2013.

[56] D. Sarko, B. Beijer, R. G. Boy et al., "The pharmacokinetics of cell-penetrating peptides," *Molecular Pharmaceutics*, vol. 7, no. 6, pp. 2224–2231, 2010.

[57] F. D. Nascimento, L. Sancey, A. Pereira et al., "The natural cell-penetrating peptide crotamine targets tumor tissue in vivo and triggers a lethal calcium-dependent pathway in cultured cells," *Molecular Pharmaceutics*, vol. 9, no. 2, pp. 211–221, 2012.

[58] D. Görlich and I. W. Mattaj, "Nucleocytoplasmic transport," *Science*, vol. 271, no. 5255, pp. 1513–1518, 1996.

[59] D. Kalderon, B. L. Roberts, W. D. Richardson, and A. E. Smith, "A short amino acid sequence able to specify nuclear location," *Cell*, vol. 39, no. 3, pp. 499–509, 1984.

[60] D. A. Jans, D. A. Jans, P. Jans, and P. Jans, "Negative charge at the casein kinase II site flanking the nuclear localization signal of the SV40 large T-antigen is mechanistically important for enhanced nuclear import," *Oncogene*, vol. 9, no. 10, pp. 2961–2968, 1994.

[61] D. Derossi, A. H. Joliot, G. Chassaing, and A. Prochiantz, "The third helix of the Antennapedia homeodomain translocates through biological membranes," *The Journal of Biological Chemistry*, vol. 269, no. 14, pp. 10444–10450, 1994.

[62] T.-C. Fan, H.-T. Chang, I.-W. Chen, H.-Y. Wang, and M. D.-T. Chang, "A heparan sulfate-facilitated and raft-dependent macropinocytosis of eosinophil cationic protein," *Traffic*, vol. 8, no. 12, pp. 1778–1795, 2007.

[63] M. C. Schmidt, B. Rothen-Rutishauser, B. Rist et al., "Translocation of human calcitonin in respiratory nasal epithelium is associated with self-assembly in lipid membrane," *Biochemistry*, vol. 37, no. 47, pp. 16582–16590, 1998.

[64] N. Sakamoto and A. S. Rosenberg, "Apolipoprotein B binding domains: evidence that they are cell-penetrating peptides that efficiently deliver antigenic peptide for cross-presentation of cytotoxic T cells," *Journal of Immunology*, vol. 186, no. 8, pp. 5004–5011, 2011.

[65] S. Lang, B. Rothen-Rutishauser, J. C. Perriard, M. C. Schmidt, and H. P. Merkle, "Permeation and pathways of human calcitonin (hCT) across excised bovine nasal mucosa," *Peptides*, vol. 19, no. 3, pp. 599–607, 1998.

[66] A. Elmquist, M. Lindgren, T. Bartfai, and Ü. Langel, "Vecadherin-derived cell-penetrating peptide, pVEC with carrier functions," *Experimental Cell Research*, vol. 269, no. 2, pp. 237–244, 2001.

[67] E. Eiríksdóttir, I. Mäger, T. Lehto, S. El Andaloussi, and Ü. Langel, "Cellular internalization kinetics of (luciferin-)cell-penetrating peptide conjugates," *Bioconjugate Chemistry*, vol. 21, no. 9, pp. 1662–1672, 2010.

[68] I. Mäger, E. Eiríksdóttir, K. Langel, S. EL Andaloussi, and Ü. Langel, "Assessing the uptake kinetics and internalization mechanisms of cell-penetrating peptides using a quenched fluorescence assay," *Biochimica et Biophysica Acta*, vol. 1798, no. 3, pp. 338–343, 2010.

[69] F. Duchardt, I. R. Ruttekolk, W. P. R. Verdurmen et al., "A cell-penetrating peptide derived from human lactoferrin with conformation-dependent uptake efficiency," *The Journal of Biological Chemistry*, vol. 284, no. 52, pp. 36099–36108, 2009.

[70] H. Noguchi, S. Matsumoto, T. Okitsu et al., "PDX-1 protein is internalized by lipid raft-dependent macropinocytosis," *Cell Transplantation*, vol. 14, no. 9, pp. 637–645, 2005.

[71] F. J. Byfield, Q. Wen, K. Leszczyńska et al., "Cathelicidin LL-37 peptide regulates endothelial cell stiffness and endothelial barrier permeability," *American Journal of Physiology—Cell Physiology*, vol. 300, no. 1, pp. C105–C112, 2011.

[72] R. Berges, J. Balzeau, A. C. Peterson, and J. Eyer, "A tubulin binding peptide targets glioma cells disrupting their microtubules, blocking migration, and inducing apoptosis," *Molecular Therapy*, vol. 20, no. 7, pp. 1367–1377, 2012.

[73] C. Lépinoux-Chambaud and J. Eyer, "The NFL-TBS.40–63 anti-glioblastoma peptide enters selectively in glioma cells by endocytosis," *International Journal of Pharmaceutics*, vol. 454, no. 2, pp. 738–747, 2013.

[74] H. Jia, M. Lohr, S. Jezequel et al., "Cysteine-rich and basic domain HIV-1 Tat peptides inhibit angiogenesis and induce endothelial cell apoptosis," *Biochemical and Biophysical Research Communications*, vol. 283, no. 2, pp. 469–479, 2001.

[75] R. R. Mehta, T. Yamada, B. N. Taylor et al., "A cell penetrating peptide derived from azurin inhibits angiogenesis and tumor growth by inhibiting phosphorylation of VEGFR-2, FAK and Akt," *Angiogenesis*, vol. 14, no. 3, pp. 355–369, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 239654, 9 pages
http://dx.doi.org/10.1155/2015/239654

Review Article

A Survey on the Computational Approaches to Identify Drug Targets in the Postgenomic Era

Yan-Fen Dai[1,2] and Xing-Ming Zhao[3,4]

[1] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[2] *Department of Mathematics, Shanghai University, Shanghai 200444, China*
[3] *Department of Computer Science, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China*
[4] *Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*

Correspondence should be addressed to Xing-Ming Zhao; zhaoxingming@gmail.com

Received 1 July 2014; Accepted 27 August 2014

Academic Editor: Hao-Teng Chang

Identifying drug targets plays essential roles in designing new drugs and combating diseases. Unfortunately, our current knowledge about drug targets is far from comprehensive. Screening drug targets in the lab is an expensive and time-consuming procedure. In the past decade, the accumulation of various types of omics data makes it possible to develop computational approaches to predict drug targets. In this paper, we make a survey on the recent progress being made on computational methodologies that have been developed to predict drug targets based on different kinds of omics data and drug property data.

1. Introduction

In the past decades, the time and cost of developing new drugs have soared significantly. In general, it takes about 15 years and up to 800 million dollars to convert a promising new compound into a drug in the market [1]. In the procedure of drug discovery, the identification of drug targets is the first and one of the most important steps. With the therapeutic targets, the optimal compounds with expected effects can be designed and new indications of old drugs may be discovered. For example, mitoxantrone was originally designed as a type II topoisomerase inhibitor. Recently, Wan et al. [2] found that mitoxantrone can inhibit the PIM1-mediated phosphorylation in cancer cells by binding to PIM1 kinase. Another example is ellipticine that was designed to target Top2 protein, but recent *in vitro* experiments indicate that ellipticine is able to decrease the proliferation rate in cancers by selectively targeting Pol-1 [3]. The targets of drugs also provide insights into the mechanism of actions (MOAs) of these drugs. Therefore, large efforts have been made to screen drug targets in lab. Accordingly, the information about drug targets has been deposited in many public databases (see Table 1), for example, STITCH [4] and DrugBank [5]. These valuable resources make it much easier to design new drugs. However, the knowledge about drug targets is far from comprehensive, which hampers the discovery of new drugs. Considering the cost and time spent in searching for drug targets, it is not feasible to screen all possible molecules targeted by drugs in lab.

Under these circumstances, some computational approaches have been proposed to identify or predict drug targets *in silico*. In particular, the accumulation of various types of omics data, such as gene expression and protein structure, makes it possible to develop more efficient computational methodologies to predict drug targets. For example, with the assumption that the drugs with the same MOAs will induce similar gene expressions, Iorio et al. [6] proposed a new approach to identify drugs that may target the same proteins. Assuming that drugs with similar MOA bind to similar pockets on the protein surfaces, some computational approaches have been developed to predict drug-protein interactions by investigating the similarity between binding profiles of candidate ligands and known drugs [7, 8]. Supposing that proteins with similar functions

Table 1: Popular drug target databases.

Drug target databases	Websites
DrugBank	http://www.drugbank.ca/
STITCH	http://stitch.embl.de/
ChEMBL	https://www.ebi.ac.uk/chembldb/
Superdrug	http://bioinformatics.charite.de/superdrug2_/
DGIdb	http://dgidb.genome.wustl.edu/
Binding DB	http://www.bindingdb.org/bind/index.jsp
CLiBE	http://xin.cz3.nus.edu.sg/group/clibe/clibe.asp
The TDR Targets database	http://tdrtargets.org/
Comparative Toxicogenomics Database (CTD)	http://ctdbase.org/
IUPHAR-DB	http://www.iuphar-db.org/index.jsp
PROMISCUOUS	http://bioinformatics.charite.de/promiscuous/
KEGG BRITE	http://www.genome.jp/kegg/brite.html
Potential Drug Target Database (PDTD)	http://www.dddc.ac.cn/pdtd/
Therapeutic Target Database (TTD)	http://bidd.nus.edu.sg/group/ttd/ttd.asp

may be bound by same drugs while drugs with similar chemical structures possibly target same proteins, Yamanishi et al. [9, 10] proposed a novel model to predict drug-protein interactions by integrating chemical structure and genomic sequence information, and they later further took into account the pharmacological information to improve prediction accuracy.

In this review, we present the recent progresses on computational methodologies that have been developed to identify drug targets. In particular, we focus on those methodologies based on gene expression data, molecular networks, and pharmacological information due to the rich resources of these types of data. As a well studied topic, those computational approaches that have been developed to predict drug targets based on protein structures are referred to in a recent review paper by Tan et al. [8]. Furthermore, we introduce popular public resources about drug target information, which can significantly facilitate the discovery of new drugs. Note that this survey aims to summarize the recent progress on computational approaches for prediction of drug targets; however, it is by no means comprehensive due to the rapid evolvement of the field.

2. Predicting Drug Targets Based on Gene Expression Profiles

A large part of known drugs target certain proteins to exert their functions after they are administered. Therefore, the gene expression profiles induced by drugs can provide insights into the mechanisms of action of these drugs to some extent, where the transcriptome data is able to monitor the expression dynamics of tens of thousands of genes simultaneously. Recently, the publicly accessible gene expression profiles, for example, Connectivity Map (CMap) (http://www.broadinstitute.org/ccle/home), NCI-60 cell lines (http://dtp.nci.nih.gov/), LINCS (http://lincs.hms.harvard.edu/db/), and CCLE (http://www.broadinstitute.org/ccle/home), make it possible to predict drug targets based on the transcriptome data. As shown in Figure 1, some computational approaches have been presented to define expression signatures that are able to characterize the MOAs of corresponding drugs, and these signatures can in turn be utilized to predict targets of novel compounds, where it is assumed that the drugs binding to the same proteins will induce similar gene expression profiles.

In their pioneering work, Lamb et al. [11] established the CMap (Connectivity Map) database that is composed of the genome-wide gene expression profiles induced by more than one thousand compounds across four cell lines. Furthermore, they defined gene signatures from these expression data to characterize the MOAs of those compounds and in turn utilized these signatures to connect small molecules with genes and diseases. The new indications of some drugs were discovered based on the alignment of drug signatures with the assumption that drugs with similar signatures may have similar therapeutic effects [12]. Based on the gene expression profiles from CMap, Iorio et al. [6, 13] constructed a drug-drug network (DDN), where drugs with similar signatures were connected. Furthermore, they extracted network communities from the DDN, and drugs with similar MOAs were found to be enriched in each community. Accordingly, the drugs in the same community are more likely to target the same proteins or pathways. They provided a computational tool, called MANTRA (http://mantra.tigem.it/), to facilitate the analysis of drug-induced gene expression profiles. Iskar et al. [14] presented a new strategy to normalize the gene expression profiles from CMap, which significantly removed the batch effect inherited in the datasets. With the signature defined similar to GSEA [15] for each drug, they successfully identified drugs with similar mechanisms and found new targets for some drugs. Analysis of characterized modules constructed with drug-induced coregulated genes reveals that zaprinast, a drug that had been previously reported to be clinically unsuccessful, is refereed interacting with new target PARAγ and has been experimentally validated successfully [16].

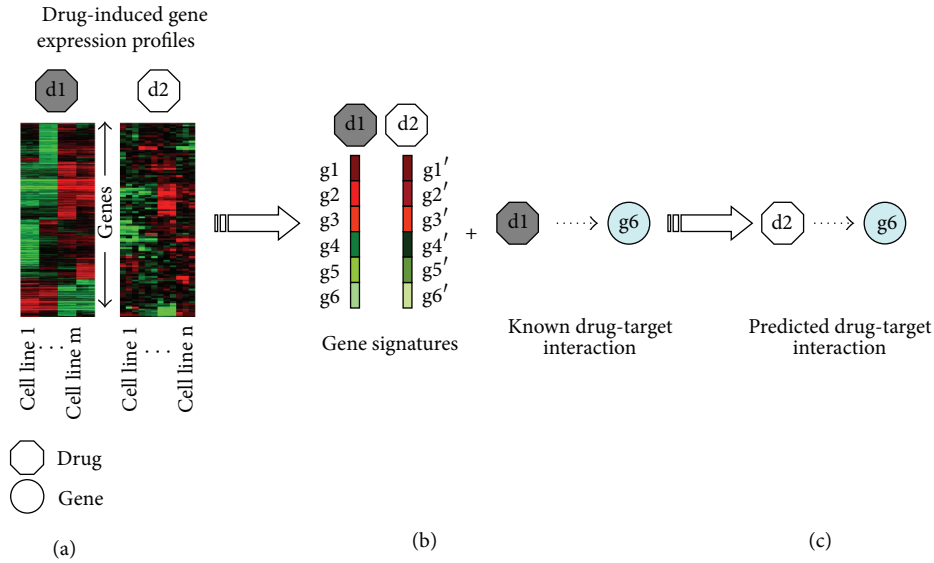


Figure 1: A schematic view of identifying drug-target interactions based on drug-induced gene expression profiles. (a) The drug-induced gene expression profiles across cell lines. (b) Define a gene signature for each compound and calculate the MOA similarity between each pair of drugs. (c) Predict targets for novel drugs with the assumption that drugs with similar MOAs are likely to target same proteins.

The NCI-60 cell line dataset [17] generated by the Development Therapeutics Program of the National Cancer Institute (NCI) is another commonly used valuable resource that contains expression profiles of genes and miRNAs induced by ~400,000 compounds across 60 cell lines. With the assumption that compounds with similar activity profiles may target similar proteins, new possible drug-protein interactions can be predicted by clustering analysis of compound bioactivity profiles across cell lines. To facilitate discovery of anticancer drugs based on the NCI-60 dataset, Reinhold et al. [18] developed a web-based tool called CellMiner along with the expression profiles of 22,217 genes and 360 microRNAs across 60 cell lines perturbed by 18,549 compounds. They identified Tdp1 as the new target of indenoisoquinoline that was originally thought to target Top1 only [19]. Yan et al. [20] identified thioredoxin reductase as a potential target of indolequinone by screening drugs in pancreatic cancer cell line and compared the compounds' bioactivity profiles against those from the NCI-60 cell line panel. Cheng et al. [21] presented a computational approach, namely, BASS, to calculate drug similarities based on their bioactivity profiles, which can in turn be utilized to predict new target(s) for known drugs or targets for novel compounds.

Beyond the above compound-centered large datasets, the accumulation of huge amount of gene expression profiles deposited in the Gene Expression Omnibus (GEO) also significantly facilitates the identification of drug targets. For example, utilizing the transcriptome profiles treated with letrozolein, the ER^+ breast tumors, Penrod and Moore [22] proposed an influence network approach that can not only identify promising targets but also suggest potential target combinations. The publicly available huge amount of transcriptome data is making it an attractive field to predict drug targets and reposition known drugs based on the gene expression profiles. In addition, the genome-wide gene expression profiles provide new insights into the drug MOAs from a systematic perspective.

3. Identifying Drug-Target Interactions from Molecular Networks

Despite the usefulness of the transcriptome data, most drugs exert their functions by affecting the activity of proteins, whereas it is known that there is a gap between the transcriptome and proteome [23]. The biological systems consist of various molecular interactions, for example, protein-protein interactions, and these interactions can be represented as distinct molecular networks depending on the interaction nature. The molecular networks can provide insights into the context in which the drug target works and can therefore help understand the drug mechanisms of action.

Among various types of molecular networks, the protein-protein interaction network (PPIN) is well studied. Since the PPIN provides the context in which the target protein works, the PPIN is also utilized to predict drug targets with the assumption that the proteins targeted by drugs of similar MOAs tend to be functionally associated and be close in the PPIN [24, 25]. As shown in Figure 2(a), if a protein is close to the one targeted by a drug, this protein is more likely to be targeted by the drug or a drug with similar therapeutic effects. Based on this idea, Zhao and Li [26] proposed a novel method named drugCIPHER to predict drug-target interactions by integrating drug therapy information, chemical structure information, and PPIN. Later, drugCIPHER has been successfully applied to predict targets of traditional Chinese medicine (TCM). For example, AKT and SRC were identified as targets of vitexicarpin [27], CCR2 was identified as the target of three compounds

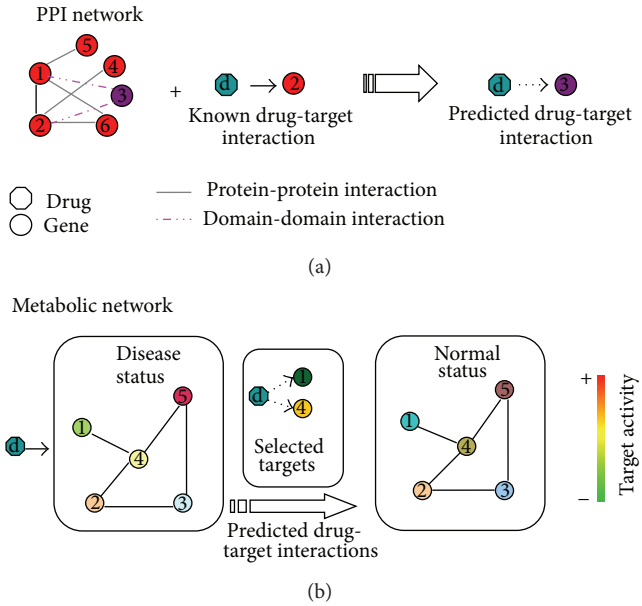


Figure 2: A schematic view of identifying drug-target interaction from molecular networks. (a) Identify drug targets from PPIN supposing that proteins in close proximity of the PPIN are more likely targeted by the same drug(s). (b) Predict drug targets based on metabolic networks assuming that the targets are able to interrupt the pathological procedure so that the disease status can be reversed to normal status.

betulin, fucosterol, and amyrin [28], and IL1R1 was the target of matrine, a bioactive compound of the herbal formula Qing-Luo-Yin [29]. Considering that some nodes play more important roles than others in a complex network, some computational approaches have been proposed by taking into account some network attributes, for example, degree and centrality, to characterize the drug targets. The degree of a protein in the PPIN is the number of interactions in which this protein is involved, while centrality indexes quantify the relative importance of a protein. For instance, Yao and Rzhetsky [25] utilized the protein betweenness centrality in a PPIN to predict drug-target interactions (DTIs) with the assumption that good targets should be of low "betweenness centrality" since the interruption of those highly connected nodes in the PPIN may cause broad and often unintended consequences. Hwang et al. [30] investigated DTIs from the perspective of bridging centrality. Using degree and centrality as features, Zhu et al. [31] trained a SVM classifier to rank potential drug targets and achieved promising results. Among their top 200 predictions, 94 proteins were validated as drug targets in DrugBank [5] database while some novel predictions can find supporting evidences in literature and other public databases.

Considering that the structure and function of a protein are generally determined by its component domains, we proposed a novel computational approach to predict drug targets supposing that drug-protein interactions are dominated by drug-domain interactions even if the drug-domain interactions are not necessarily physical binding interactions [32]. In our approach, the drug-domain interactions were first inferred from known drug-protein interactions as below:

$$P\left(m_i\_d_{\mathrm{ATC}_j}\right) = \frac{N\left(p \mid m_i\right)}{N\left(p' \mid m_i\right)}, \quad (1)$$

where ATC code is the abbreviation of "Anatomical Therapeutic Chemical," a classification system used for the classification of drugs, $\mathrm{ATC}_{(j)}$ means ATC code j, $P(m_i\_d_{\mathrm{ATC}_j})$ is the probability that domain m_i interacts with drugs annotated with ATC_j, $N(p \mid m_{(i)})$ denotes the number of proteins that are bound by drugs belonging to ATC_j and contain domain m_i as well, and $N(p' \mid m_i)$ is the number of all human proteins that contain domain m_i. After obtaining the probability of drug-domain interactions, we can determine whether a pair of drugs and domain interact by setting a threshold, where those drug-domain pairs with probabilities above the threshold were treated as drug-domain pair interactions. Accordingly, we can predict drug-protein interactions based on the drug-domain interactions as follows:

$$P\left(p_i\_d_{\mathrm{ATC}_j}\right) = 1 - \prod\left(1 - P\left(m_k\_d_{\mathrm{ATC}_j}\right)\right), \quad (2)$$

where $P(p_i\_d_{\mathrm{ATC}_j})$ is the probability of protein p_i interacting with drugs belonging to $\mathrm{ATC}_{(j)}$, $P(m_k\_d_{\mathrm{ATC}_j})$ is the probability that domain m_k interacts with drugs from $\mathrm{ATC}_{(j)}$, and p_i is a protein that contains domain m_k. The results on benchmark dataset show that our proposed approach can improve prediction accuracy compared with other popular methods. Later, with the drug-domain interaction network, Moya-García and Ranea [33] found that drugs are organized around a privileged set of druggable domains, which can help explain drug polypharmacology.

Except for PPIN, the metabolic networks are also widely used to predict drug targets. In the metabolic network based approach, it is assumed that the disruption of pathogenic pathways or inhibition of certain molecules can help reverse the disease state to normal state. With flux balance analysis (FBA) of metabolic networks, Li et al. [34, 35] developed a new approach to identify potential therapeutic drug targets by comparing the fluxes of reactions and metabolites in pathologic and medication states based on linear programming. By simulating the flux distribution in the metabolic network, Folger et al. [36] successfully identified some targets of anticancer drugs. With a detailed disease network, Yang et al. [37] proposed a computational framework to identify optimal multiple target intervention (MTOI) solution by simulating the dynamics of the system with mass action modeling along with simulated annealing. The optimal target combinations detected by this promising method not only overcome the compensatory mechanisms in diseases but also avoid unwanted side effects caused by possible off-targets. By integrating the gene expression profiles across cell lines and human metabolic networks, Li et al. [38] identified new enzyme targets with kernel k-nearest neighbor (kNN) classifiers by comparing the reaction flux of novel compound-reaction against that of known drug-reaction. Furthermore, utilizing the genome-scale metabolic models (GSMMs),

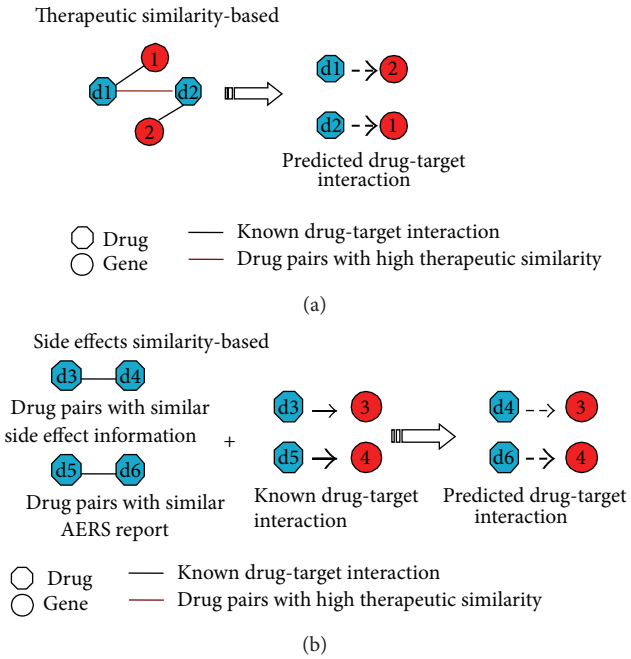


FIGURE 3: A schematic view of identifying drug-target interactions based on drug effect profiles. (a) Identify drug-target interaction based on therapy information by assuming that drugs with similar therapy may target same protein(s). (b) Predict drug targets based on side effects supposing that drugs with similar side effect have common target(s).

Yizhak et al. [39] proposed a metabolic transformation algorithm (MTA) to search for targets that could restore the metabolism within the cell from the source (disease) state to the target (healthy) state.

Since the molecular networks are able to provide the circuit context in which the drug target protein works, they provide a straightforward way to understand how the drugs affect or regulate the biological systems. Unfortunately, our current knowledge about the molecular interactomes at different levels is far from complete. Even though large-scale interactomes have been detected or predicted, they are just static snapshots of the biological systems, whereas the real biological systems are spatially and temporally dynamic. Furthermore, little is known about the detailed interaction kinetics. All these limit the application of the molecular networks in the identification of drug targets.

4. Identifying Drug Targets with Drug Effects

Except for the omics data from molecular space, a straightforward way to understand the drug MOAs is to explore the drug effects in the pharmacological space, which can in turn help predict the drug targets. Similar to the approaches based on gene expression profiles, the drug effect based approaches assume that the drugs with similar therapeutic effect may target the same protein(s) (see Figure 3). For example, Yamanishi et al. [10] found that the drug therapy information can better characterize drug targets compared against the commonly used chemical structure information. With the drug pharmacological information predicted with chemical structures, they significantly improved the prediction accuracy with a supervised bipartite graph model. Cheng et al. [40] integrated the chemical structure information with pharmacological information, to predict DTIs, and obtained promising results, where the drug therapeutic similarity they used was defined by Xu et al. [41] as shown below:

$$\mathrm{TS}(d_1, d_2) = \frac{\sum_{k=1}^{3} S_k(d_1, d_2)}{n}, \tag{3}$$

where n ranges from 1 to 5 and $S_k(d_1, d_2)$ is defined as

$$S_k(d_1, d_2) = \frac{\mathrm{ATC}_k(d_1) \cap \mathrm{ATC}_k(d_2)}{\mathrm{ATC}_k(d_1) \cup \mathrm{ATC}_k(d_2)}, \tag{4}$$

where $\mathrm{ATC}_k(d)$ denotes all the ATC codes at the kth level of drug d. Note that a drug has five levels of ATC codes.

In contrast to the therapy information, little attention has been paid to the adverse effects caused by drugs when predicting the drug targets. It is known that the unexpected drug side effects may be caused because of the off-targets [42, 43] and these off-targets may help to predict therapeutic targets. Recently, Campillos et al. [44] proposed a novel approach to predict the drug targets based on the drug side effects, where they assumed that the drugs with similar side effects will share common target proteins. To calculate the drug similarity based on their side effect profiles, they first extracted drug associated adverse effects from FDA adverse event reporting system and formalized them with the Unified Medical Language System (UMLS) ontology [45]. The drug side effect information has been deposited in the resource of SIDER [46]. With the drug adverse reaction information, they discovered unexpected connections among drugs with different chemical structures and therapeutic indications. By integrating the chemical structures and side effects, they significantly improved the prediction accuracy and identified some novel predictions which otherwise will not be found with only chemical structures. In addition, some of their predictions were experimentally validated, implying the predictive power of the side effects. With novel targets identified for old drugs, new potential indications can be found for these known drugs. For instance, the authors found that the nervous system drugs pergolide, paroxetine, and fluoxetine share the same targets with the drug rabeprazole that is an approved drug for relieving duodenal ulcer symptoms and treating ulcerative gastroesophageal reflux disease, indicating that these drugs may be repositioned for treating new diseases. Due to the scarceness of drugs' side effect information, Takarabe et al. [47] proposed a new approach, to predict novel drug-target interactions by integrating pharmacological information from AERS (adverse event reporting system) and genomic information for proteins, and found some novel targets.

The pharmacological information associated with drugs provides an alternative way to predict drug targets and has been proved to be complementary with the commonly used molecular information, for example, genome sequence or

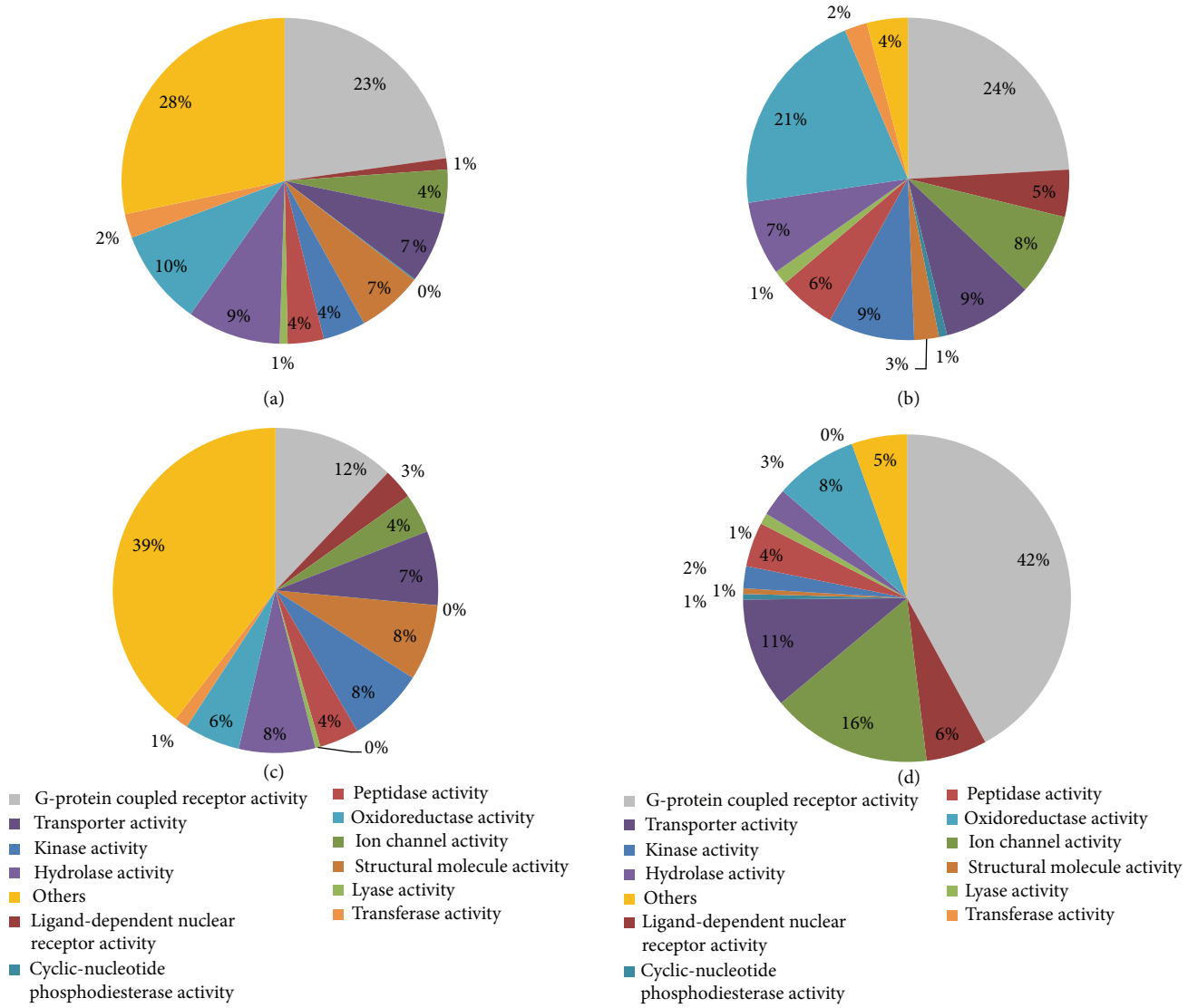


Figure 4: Functional distribution of human proteins (a), drug targets (b), neighbor proteins of drug targets (c) in the PPIN, and drug therapeutic targets (d).

transcriptome data. Unfortunately, the scarceness of drug package and adverse reaction information limits the application of above-mentioned approaches to those well studied drugs. In addition, the drug effects are determined by the molecular context of their target proteins and those drugs with similar effects may not share any target proteins in fact. For example, Brouwers et al. [48] found that the drug side effects are determined by the neighborhood of their targets in a PPIN, where the same neighborhood does not necessarily mean same target proteins.

5. Discussion and Conclusion

The identification of drug targets plays essential roles in understanding the drug MOAs and designing new drugs with expected therapy. In this review, we summarized the recent progress on computational methodologies that have been developed to identify drug-target interactions. We summarized some recent popular tools or algorithms for drug target prediction in Table 2. Furthermore, we categorized these approaches according to the high-throughput data on which they work. In particular, we focused on those approaches that explore transcriptome, molecular network, and drug effect data due to their public availability. The transcriptome data provides a snapshot of the whole-genome dynamics and can help understand the mechanisms of action of drugs. The transcriptome-driven computational approaches assume that the drugs with similar gene expression signature will target the same protein. However, it is not easy to define a robust gene signature due to the noise and batch effects inherited in the gene expression data. The molecular network provides the circuit context in which the drug targets work, which makes the network approaches promising. Unfortunately, the

Table 2: Popular software/algorithms for identifying drug target.

Reference	Data used
Iskar et al. [14]	Transcriptome profiles
Reinhold et al. [18]	Transcriptome profiles
Cheng et al. [21]	Transcriptome profiles
Carrella et al. [49]	Transcriptome profiles
Xu et al. [50]	Transcriptome profiles
Zhao and Li [26]	Molecular networks
Li et al. [35]	Molecular networks
Gönen [51]	Molecular networks
Wang et al. [52]	Molecular networks
Yizhak et al. [39]	Molecular networks
Yang et al. [53]	Molecular networks
Takarabe et al. [47]	Drug effects
Campillos et al. [44]	Drug effects
Mizutani et al. [54]	Drug effects
Iwata et al. [55]	Drug effects

incompleteness of the network knowledge and the network dynamics induced by drugs limit the application of these methods. Compared with the molecular data, the drug therapy and side effect information are more difficult to get. Therefore, the integration of distinct types and complementary data will be a promising direction in the future.

Except for the above-mentioned data, the functions of the proteins targeted by drugs should also be taken into account. Figure 4 shows the functional distribution of human proteins, drug targets, neighborhood proteins of drug targets in PPIN, and drug therapeutic targets. The drug target information was extracted from DrugBank [5] database, the neighborhood proteins are those direct neighbors of drug targets in the PPIN that was extracted from Entrez Gene Database [56], and the therapeutic targets were retrieved from [57]. All the proteins were grouped according to the molecular function annotations from Gene Ontology [58]. We can clearly see that the drug targets have different functions compared with the human genome background. On the other hand, the therapeutic targets have different functions from all proteins that can be targeted by drugs, implying that the off-targets may have specific functions. What is interesting is that unlike the drug targets, most of which belong to the GPCR family, the neighborhood proteins of drug targets belong to transferase. This information should be utilized to improve prediction accuracy when developing new methodologies 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 is partly supported by the National Nature Science Foundation of China (91130032, 61103075), Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040700), Innovation Program of Shanghai Municipal Education Commission (13ZZ072), and Shanghai Pujiang Program (13PJD032).

References

[1] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, "The price of innovation: new estimates of drug development costs," *Journal of Health Economics*, vol. 22, no. 2, pp. 151–185, 2003.

[2] X. Wan, W. Zhang, L. Li, Y. Xie, W. Li, and N. Huang, "A new target for an old drug: Identifying mitoxantrone as a nanomolar inhibitor of PIM1 kinase via kinome-wide selectivity modeling," *Journal of Medicinal Chemistry*, vol. 56, no. 6, pp. 2619–2629, 2013.

[3] W. J. Andrews, T. Panova, C. Normand, O. Gadal, I. G. Tikhonova, and K. I. Panov, "Old drug, new target: ellipticines selectively inhibit RNA polymerase I transcription," *Journal of Biological Chemistry*, vol. 288, no. 7, pp. 4567–4582, 2013.

[4] M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild et al., "STITCH 4: integration of protein-chemical interactions with user data," *Nucleic Acids Research*, vol. 42, no. 1, pp. D401–D407, 2014.

[5] C. Knox, V. Law, T. Jewison et al., "DrugBank 3.0: a comprehensive resource for "Omics" research on drugs," *Nucleic Acids Research*, vol. 39, no. 1, pp. D1035–D1041, 2011.

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

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

[8] H. Tan, X. Ge, and L. Xie, "Structural systems pharmacology: a new frontier in discovering novel drug targets," *Current Drug Targets*, vol. 14, no. 9, pp. 952–958, 2013.

[9] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.

[10] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. 12, Article ID btq176, pp. i246–i254, 2010.

[11] J. Lamb, E. D. Crawford, D. Peck et al., "The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[12] J. Lamb, "The Connectivity Map: a new tool for biomedical research," *Nature Reviews Cancer*, vol. 7, no. 1, pp. 54–60, 2007.

[13] F. Iorio, R. Tagliaferri, and D. di Bernardo, "Identifying network of drug mode of action by gene expression profiling," *Journal of Computational Biology*, vol. 16, no. 2, pp. 241–251, 2009.

[14] M. Iskar, M. Campillos, M. Kuhn, L. J. Jensen, V. van Noort, and P. Bork, "Drug-induced regulation of target expression," *PLoS Computational Biology*, vol. 6, no. 9, Article ID e1000925, 2010.

[15] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

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

[17] R. H. Shoemaker, "The NCI60 human tumour cell line anticancer drug screen," *Nature Reviews Cancer*, vol. 6, no. 10, pp. 813–823, 2006.

[18] W. C. Reinhold, M. Sunshine, H. Liu et al., "CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set," *Cancer Research*, vol. 72, no. 14, pp. 3499–3511, 2012.

[19] M. Conda-Sheridan, P. V. N. Reddy, A. Morrell et al., "Synthesis and biological evaluation of indenoisoquinolines that inhibit both tyrosyl-DNA phosphodiesterase i (Tdp1) and topoisomerase i (Top1)," *Journal of Medicinal Chemistry*, vol. 56, no. 1, pp. 182–200, 2013.

[20] C. Yan, B. Shieh, P. Reigan et al., "Potent activity of indolequinones against human pancreatic cancer: identification of thioredoxin reductase as a potential target," *Molecular Pharmacology*, vol. 76, no. 1, pp. 163–172, 2009.

[21] T. Cheng, Q. Li, Y. Wang, and S. H. Bryant, "Identifying compound-target associations by combining bioactivity profile similarity search and public databases mining," *Journal of Chemical Information and Modeling*, vol. 51, no. 9, pp. 2440–2448, 2011.

[22] N. M. Penrod and J. H. Moore, "Influence networks based on coexpression improve drug target discovery for the development of novel cancer therapeutics," *BMC Systems Biology*, vol. 8, no. 1, article 12, 2014.

[23] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold, "Correlation between protein and mRNA abundance in yeast," *Molecular and Cellular Biology*, vol. 19, no. 3, pp. 1720–1730, 1999.

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

[25] L. Yao and A. Rzhetsky, "Quantitative systems-level determinants of human genes targeted by successful drugs," *Genome Research*, vol. 18, no. 2, pp. 206–213, 2008.

[26] S. Zhao and S. Li, "Network-based relating pharmacological and genomic spaces for drug target identification," *PLoS ONE*, vol. 5, no. 7, Article ID e11764, 2010.

[27] B. Zhang, L. Liu, S. Zhao, X. Wang, and S. Li, "Vitexicarpin acts as a novel angiogenesis inhibitor and its target network," *Evidence-Based Complementary and Alternative Medicine*, vol. 2013, Article ID 278405, 13 pages, 2013.

[28] X. Liang, H. Li, and S. Li, "A novel network pharmacology approach to analyse traditional herbal formulae: the Liu-Wei-Di-Huang pill as a case study," *Molecular BioSystems*, vol. 10, no. 5, pp. 1014–1022, 2014.

[29] B. Zhang, X. Wang, and S. Li, "An integrative platform of TCM network pharmacology and its application on a herbal formula, Qing-Luo-Yin," *Evidence-Based Complementary and Alternative Medicine*, vol. 2013, Article ID 456747, 12 pages, 2013.

[30] W.-C. Hwang, A. Zhang, and M. Ramanathan, "Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery," *Clinical Pharmacology & Therapeutics*, vol. 84, no. 5, pp. 563–572, 2008.

[31] M. Zhu, L. Gao, X. Li et al., "The analysis of the drug-targets based on the topological properties in the human protein-protein interaction network," *Journal of Drug Targeting*, vol. 17, no. 7, pp. 524–532, 2009.

[32] L. Wang, Z.-P. Liu, X.-S. Zhang, and L. N. Chen, "Prediction of hot spots in protein interfaces using a random forest model with hybrid features," *Protein Engineering, Design & Selection*, vol. 25, no. 3, pp. 119–126, 2012.

[33] A. A. Moya-García and J. A. G. Ranea, "Insights into polypharmacology from drug-domain associations," *Bioinformatics*, vol. 29, no. 16, pp. 1934–1937, 2013.

[34] Z. P. Li, R. S. Wang, and X. S. Zhang, "Drug target identification based on flux balance analysis of metabolic networks," *Computational Systems Biology*, vol. 13, pp. 331–338, 2010.

[35] Z. Li, R.-S. Wang, and X.-S. Zhang, "Two-stage flux balance analysis of metabolic networks for drug target identification," *BMC Systems Biology*, vol. 5, no. 1, article S11, 2011.

[36] O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, and T. Shlomi, "Predicting selective drug targets in cancer through metabolic networks," *Molecular Systems Biology*, vol. 7, article 501, 2011.

[37] K. Yang, H. Bai, Q. Ouyang, L. Lai, and C. Tang, "Finding multiple target optimal intervention in disease-related molecular network," *Molecular Systems Biology*, vol. 4, p. 228, 2008.

[38] L. Li, X. Zhou, W.-K. Ching, and P. Wang, "Predicting enzyme targets for cancer drugs by profiling human metabolic reactions in NCI-60 cell lines," *BMC Bioinformatics*, vol. 11, article 501, 2010.

[39] K. Yizhak, O. Gabay, H. Cohen, and E. Ruppin, "Model-based identification of drug targets that revert disrupted metabolism and its application to ageing," *Nature Communications*, vol. 4, article 2632, 2013.

[40] F. Cheng, W. Li, Z. Wu et al., "Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space," *Journal of Chemical Information and Modeling*, vol. 53, no. 4, pp. 753–762, 2013.

[41] K.-J. Xu, J. Song, and X.-M. Zhao, "The drug cocktail network," *BMC Systems Biology*, vol. 6, supplement 1, article S5, 2012.

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

[43] S. Whitebread, J. Hamon, D. Bojanic, and L. Urban, "Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development," *Drug Discovery Today*, vol. 10, no. 21, pp. 1421–1433, 2005.

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

[45] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, pp. D267–D270, 2004.

[46] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs," *Molecular Systems Biology*, vol. 6, p. 343, 2010.

[47] M. Takarabe, M. Kotera, Y. Nishimura, S. Goto, and Y. Yamanishi, "Drug target prediction using adverse event report systems: a pharmacogenomic approach," *Bioinformatics*, vol. 28, no. 18, Article ID bts413, pp. i611–i618, 2012.

[48] L. Brouwers, M. Iskar, G. Zeller, V. van Noort, and P. Bork, "Network neighbors of drug targets contribute to drug side-effect similarity," *PLoS ONE*, vol. 6, no. 7, Article ID e22187, 2011.

[49] D. Carrella, F. Napolitano, R. Rispoli et al., "Mantra 2.0: an online collaborative resource for drug mode of action and repurposing by network analysis," *Bioinformatics*, vol. 30, no. 12, pp. 1787–1788, 2014.

[50] T. Xu, R. Zhu, Q. Liu, and Z. Cao, "Quantitatively integrating molecular structure and bioactivity profile evidence into drug-target relationship analysis," *BMC Bioinformatics*, vol. 13, no. 1, article 75, 2012.

[51] M. Gönen, "Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization," *Bioinformatics*, vol. 28, no. 18, Article ID bts360, pp. 2304–2310, 2012.

[52] Y.-Y. Wang, J. C. Nacher, and X.-M. Zhao, "Predicting drug targets based on protein domains," *Molecular BioSystems*, vol. 8, no. 5, pp. 1528–1534, 2012.

[53] K. Yang, H. Bai, Q. Ouyang, L. Lai, and C. Tang, "Finding multiple target optimal intervention in disease-related molecular network," *Molecular Systems Biology*, vol. 4, article 228, 2008.

[54] S. Mizutani, E. Pauwels, V. Stoven, S. Goto, and Y. Yamanishi, "Relating drug-protein interaction network with drug side effects," *Bioinformatics*, vol. 28, no. 18, Article ID bts383, pp. i522–i528, 2012.

[55] H. Iwata, S. Mizutani, Y. Tabei, M. Kotera, S. Goto, and Y. Yamanishi, "Inferring protein domains associated with drug side effects based on drug-target interaction network," *BMC Systems Biology*, vol. 7, supplement 6, p. S18, 2013.

[56] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez gene: gene-centered information at NCBI," *Nucleic Acids Research*, vol. 39, pp. D52–D57, 2011.

[57] E. Gregori-Puigjané, V. Setola, J. Hert et al., "Identifying mechanism-of-action targets for drugs and probes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 28, pp. 11178–11183, 2012.

[58] 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.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 475062, 6 pages
http://dx.doi.org/10.1155/2015/475062

Resource Review

A Large-Scale Structural Classification of Antimicrobial Peptides

Hao-Ting Lee, Chen-Che Lee, Je-Ruei Yang, Jim Z. C. Lai, and Kuan Y. Chang

Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung 202, Taiwan

Correspondence should be addressed to Kuan Y. Chang; kchang@ntou.edu.tw

Received 1 September 2014; Accepted 23 February 2015

Academic Editor: Oliver Ray

Antimicrobial peptides (AMPs) are potent drug candidates against microbial organisms such as bacteria, fungi, parasites, and viruses. AMPs have abundant sequences and structures, two fundamental resources for bioinformatics researches, but analyses on how they associate with each other are either nonexistent or limited to partial classification and data. We thus present A Database of Anti-Microbial peptides (ADAM), which contains 7,007 unique sequences and 759 structures, to systematically establish comprehensive associations between AMP sequences and structures through structural folds and to provide an easy access to view their relationships. 30 distinct AMP structural fold clusters with more than one structure are detected and about a thousand AMPs are associated with at least one structural fold cluster. According to ADAM, AMP structural folds are limited—AMPs only cover about 3% of the overall protein fold space.

1. Introduction

Antimicrobial peptides (AMPs) are potent drug candidates against microbial organisms such as bacteria, fungi, parasites, and viruses. Up to date, more than 10 AMPs have entered clinical trials [1]. Due to the importance, several databases dedicated to AMPs were released in the past few years. Some databases are species-specific such as BACTIBASE [2], BAGEL2 [3], DADP [4], PenBase [5], and PhytAMP [6]; some curate a broad spectrum of species such as AMPer [7], APD2 [8], CAMP2 [9], DAMPD [10], Defensins Knowledgebase [11], and YADAMP [12]. The sizes of these databases range from hundreds to a couple of thousand AMP sequences. However, none of these databases contains all.

Understanding sequence-structure relationships is important for AMP-based drug design. However, one of the major limitations in AMP databases is poorly utilizing structural information. Like AMP sequences, various AMP structures have been resolved. Classified by secondary structures, four traditional AMP structures are alpha helices, beta strands, loop structures, and extended structures [13, 14]. An alternative structural classification using peptide backbone torsion angles also shows many different AMP folds [1]. Few AMP databases such as APD2 have attempted to associate AMP sequences with their secondary structures. However, none has established associations between AMP sequences and AMP structural folds. Examining AMP tertiary structures would help us understand AMPs better and enhance potential antimicrobial drug discovery.

In this work, we present A Database of Anti-Microbial peptides (ADAM) (available at http://bioinformatics.cs.ntou.edu.tw/ADAM). ADAM collects AMPs comprehensively and establishes associations systematically between AMP sequences and structures. Integrated from various sources, ADAM contains the most complete AMP sequences and structures. ADAM not only allows biomedical researchers to search basic AMP information but also provides an easy access to link AMP sequences to structures and vice versa.

2. Data Collection and Methods

2.1. AMP Sequences. ADAM contains 7,007 unique AMP sequences extracted from twelve databases (Figure 1). The twelve databases include APD2 [8], AVPpred [15], BACTIBASE [2], BAGEL3 [3], CAMP2 [9], DADP [4], DAMPD [10], HIPdb [16], PenBase [5], PhytAMP [6], RAPD [17], and YADAMP [12]. The AMP sequences in ADAM were mostly derived from natural sources, covering a broad spectrum of species such as archaea, bacteria, plants, and animals. 2497 out of the 7,007 sequences have been validated

Table 1: Comparison of overlapping identical sequence counts of the twelve AMP databases.

	APD	CAMP	DADP	DAMPD	YAD	BACTIB	BAGEL	PenBase	PhytAMP	RAPD	AVP	HIPdb
APD	**2436**	2100	744	376	1601	86	39	1	107	55	56	33
CAMP	2100	**3052**	858	586	1994	122	56	1	145	71	65	33
DADP	744	858	**1792**	220	772	0	0	0	0	5	13	11
DAMPD	376	586	220	**1068**	528	31	70	5	19	10	18	11
YADAMP	1601	1994	772	528	**2782**	113	43	1	60	67	76	49
BACTIBASE	86	122	0	31	113	**204**	52	0	0	10	0	1
BAGEL	39	56	0	70	43	52	**431**	0	0	0	1	0
PenBase	1	1	0	5	1	0	0	**28**	0	0	0	0
PhytAMP	107	145	0	19	60	0	0	0	**272**	3	10	0
RAPD	55	71	5	10	67	10	0	0	3	**119**	9	5
AVP	56	65	13	18	76	0	1	0	10	9	**604**	156
HIPdb	33	33	11	11	49	1	0	0	0	5	156	**744**

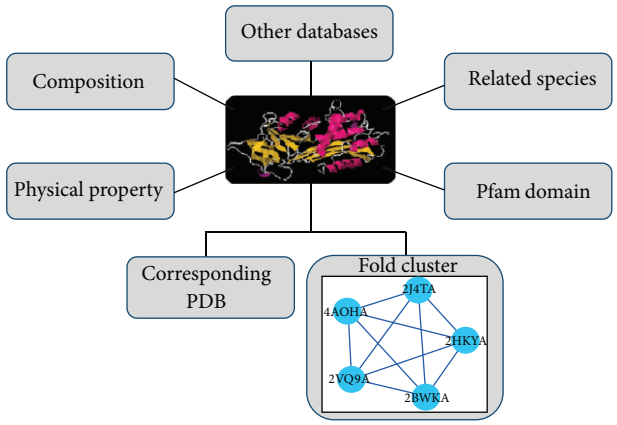


Figure 1: Simplified conceptual diagram of ADAM.

Table 2: Structural classification of the AMPs according to CATH v4.0 classification.

	Class	Architecture	Topology	Homologous superfamily
ADAM	4	11	40	41
CATH 4.0	4	40	1375	2738

Table 3: Structural classification of the AMPs according to SCOP v1.75B.

	Class	Fold	Superfamily	Family
ADAM	7	47	53	72
SCOP 1.75B	11	1390	2220	4609

experimentally and recorded in literature. Table 1 compares the AMPs of the twelve databases. The CAMP2 contains the most overlapping sequences among the large AMP databases such as APD2, CAMP2, DAMPD, DADP, and YADAMP. For species-specific AMP databases, AVP and HIPdb are found to contain less overlapping sequences.

Each unique AMP sequence was assigned an ADAM ID. The ADAM ID is linked to the basic AMP information, structural view, physicochemical properties, amino acid composition, and external resources. The structural view displays the best corresponding PDB structure and, if available, the representative PDB structure of the fold cluster which this AMP sequence belongs to. The physicochemical properties list peptide length, net charge, instability index**, aliphatic index**, and grand average of hydropathicity index**. The composition is the ratio of each amino acid in the AMP. The other resources are linked to PDB, CATH, SCOP, Pfam, and other AMP databases associated with this AMP (**see Supplementary Material available online at http://dx.doi.org/10.1155/2015/475062).

2.2. AMP Structures. The AMP structures were obtained by running BLAST of the experimentally validated AMP sequences against the Protein Data Bank [18]. 408 sequences had 759 matching structures with either 100% sequence identity or at least 90% identity sequence with the E-value < 10^{-5}. Each matching structure was annotated by SCOP v1.75B [19] and CATH v4.0 [20]. Because not every AMP structure had CATH or SCOP annotation, one could not determine all unique AMP structural folds simply based on these annotations.

Tables 2 and 3 record the number of the AMP structures according to CATH v4.0 and SCOP v1.75B, respectively. Four hierarchical levels of CATH are class, architecture, topology, and homologous superfamily; four levels of SCOP are class, fold, superfamily, and family. The topology level of CATH corresponds to the fold level of SCOP. The AMP structures appear at the entire four fundamental CATH classes (Table 2) and seven SCOP classes (Table 3). Within 759 AMP structures, 40 out of 1375 CATH folds (Table 2) and 47 out of 1390 SCOP folds (Table 3) are found. These AMP structures cover about 3% of the protein fold space defined by CATH and SCOP.

2.3. AMP Structural Fold Clusters. A graph-based clustering procedure was applied for accessing the unique AMP folds. In this graph, the vertices represent AMP structures and there is an edge between two vertices if the two AMP structures are

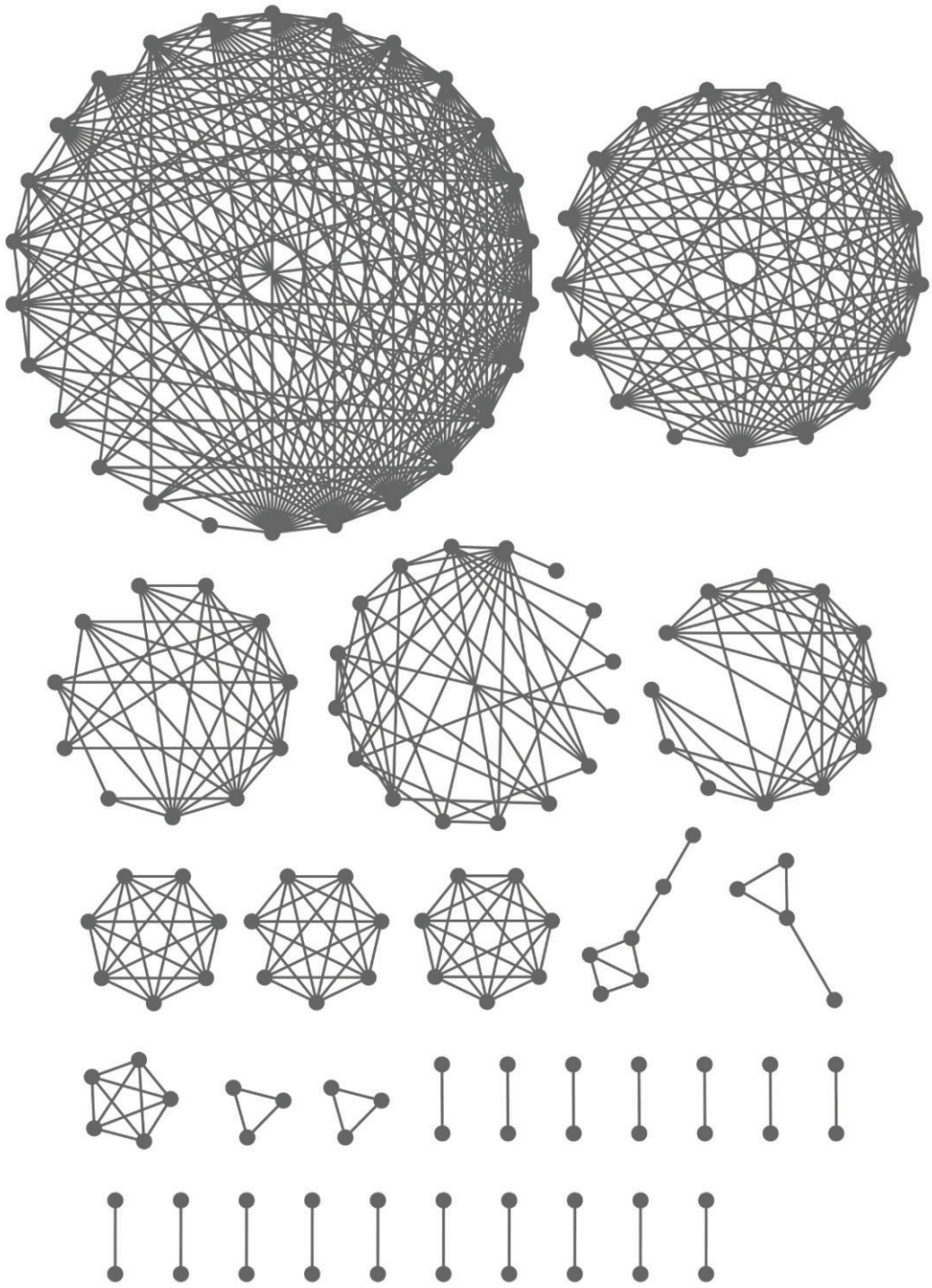


Figure 2: Nework representation of AMP structral fold clusters.

similar. The AMP structures came from the previous BLAST results. Only 264 best matching structures were collected under more stringent selection conditions. Each AMP is allowed to have at most one best matching structure, and multiple AMPs can map to the same AMP structure. The similarity of two AMP structures was then measured by TM-score, whose value ranges from 0 to 1 [21]. An edge exits if its TM-score > 0.5, which indicates that the two structures should belong to the same fold [22]. 136 AMP fold clusters were formed with 30 clusters containing more than one AMP structure, as shown in Figure 2. The top 10 common AMP structural folds with CATH and SCOP annotations are listed in Table 4. The structural fold clusters can have the same CATH and SCOP annotations as cluster #1 in Table 4. One CATH fold can map to multiple SCOP folds as cluster #4 in Table 4; one SCOP fold can also map to multiple CATH folds

Table 4: Top 10 common AMP structural folds annotated by CATH and SCOP.

AMP structural folds	CATH			SCOP	
Fold cluster ID	Class	Architecture	Topology	Class	Fold
1	Alpha beta	2-layer sandwich	Defensin A-like	Small proteins	Knottins
2	Mainly beta	Beta barrel	OB fold	Alpha and beta proteins (a + b)	IL8-like
3	Mainly alpha	Up-down bundle	Single alpha-helices involved in coiled-coils or other helix-helix interfaces	Peptides	Antimicrobial helix
3	—	—	—	Peptides	Liposaccharide-binding protein CAP18
3	—	—	—	Peptides	Peptide hormones
4	Alpha beta	Roll	Antimicrobial peptide, beta-defensin 2; chain A	Small proteins	Defensin-like
5	—	—	—	Small proteins	Knottins
6	Mainly alpha	Orthogonal bundle	Histone, subunit A	All alpha proteins	Histone-fold
7	Mainly alpha	Orthogonal bundle	Lysozyme	Alpha and beta proteins (a + b)	Lysozyme-like
8	Alpha beta	2-layer sandwich	Crambin	Small proteins	Crambin-like
9	Mainly alpha	Orthogonal bundle	NK-lysin	All alpha proteins	Saposin-like
9	Mainly alpha	Up-down bundle	Bacteriocin As-48; chain A	All alpha proteins	Saposin-like
10	Alpha beta	Roll	P-30 protein	Alpha and beta proteins (a + b)	RNase A-like

as cluster #9 in Table 4. Note that some AMP structures have neither CATH nor SCOP annotation.

The vertices represent the AMP structures and an edge between two vertices exists if the TM-score > 0.5, indicating the two structures as the two verctices fall into the same fold [22]. Among the 136 fold clusters in ADAM, 30 of them which contain more than one structure are displayed here.

2.4. AMP Structures Associated with ADAM Sequences. From AMP sequences to structures, AMP structures were obtained by performing BLAST on the experimentally validated AMPs against PDB. From AMP structures to ADAM sequences, about one-eighth of the ADAM sequences, over a thousand sequences, were found to associate with the AMP structures, which were determined by running BLAST against the best matching AMP structures with the E-value < 10^{-5}. Here we list the top 10 common Pfam domains and families [23] found in the experimentally validated AMPs and their associations with the AMP structural fold clusters (Table 5). Out of these common Pfam domains and families, seven of them fall within the top 10 AMP structural folds. Table 5 also indicates that no structures are available for the AMPs with Pfam family antimicrobial_1.

Table 5: Top 10 common Pfam domains and families associated with the AMP structural folds.

	Pfam	AMP structural fold cluster ID
1	Antimicrobial_2	3
2	Antimicrobial_1	NA
3	Defensin_beta	4
4	Gamma-thionin	1
5	Cyclotide	5
6	Defensin_2	1
7	Defensin_1	4
8	Bacteriocin_II	33
9	Cecropin	106
10	DD_K	3

3. Implementations and Results

ADAM was built using AppServ 2.6.0. The Apache HTTP server was applied, the server-side scripts were written in PHP, and the database was built by MySQL.

3.1. Multiple Search Capabilities. ADAM offers multiple search capabilities, which can be classified into two basic categories: sequence search and structural search. Each AMP entry is assigned an ADAM ID, which would have a unique sequence and, if found, a corresponding structure. The sequence search covers the direct information of an AMP sequence, including the description, source species, sequence length, and Pfam domain. ADAM which focuses on AMP structure and sequence information does not contain all of the information that other AMP databases provide. Therefore, external links to other AMP databases are also provided in ADAM. In addition, the structural search allows users to

retrieve the AMP information associated with specific PDB structures or ADAM fold clusters.

3.2. Structure-Sequence Cluster Browsing. ADAM offers 136 AMP fold clusters built by TM-score for browsing. Each structure in the AMP cluster is annotated by CATH, SCOP, and Pfam, if available. The AMP structures from all of the clusters occupy about 3% of the protein fold space defined by CATH and SCOP. Each cluster would list the associated AMP sequences.

For example, ADAM cluster #1 (AC_001) is a cluster of 26 structures associated with 207 AMP sequences. Detailed information can be found at Table S1. These structures in this cluster gathered by TM-score are consistently classified into the same CATH fold, alpha-beta 2-layer sandwich defensin A-like structure, and the same SCOP fold, small protein knottins. SCOP further classifies these structures into four different SCOP families. In addition, this AMP structural fold is found to associate with six different Pfam domains, including antimicrobial_6, defensin_2, gamma-thionin, toxin_2, toxin_3, and toxin_37, which supports that different sequences which fold into the same structure could behave similarly. Another interesting example is ADAM cluster #5 (AC_005), which contains 53 AMP sequences involved with cyclotide Pfam family. Within this cluster, only four structures are annotated by SCOP. All of the four structures are again classified into the same SCOP fold, knottins, but fall into multiple SCOP families.

ADAM also allows users to extract the relevant AMP structures according to CATH or SCOP classification by the underneath hyperlinks. In fact, both structure-to-sequence and sequence-to-structure browsing can be performed in ADAM.

Each AMP cluster is further examined. An interesting phenomenon is observed that peptides in one AMP cluster consistently belong to the same mechanism of microbial killing, either transmembrane pore formation or metabolic inhibition of intracellular targets [24], suggesting that AMP structures may play a role in the killing action. For example, the AMPs in ADAM cluster #3 (AC_003) belong to the mechanism of transmembrane pore formation; those in ADAM cluster #6 (AC_006) are the metabolic inhibitors for the intracellular targets.

4. Discussions

ADAM, which is a comprehensive AMP database, provides an easy access to AMP sequences, structures, and their relations. Two distinct characters of ADAM are its size and sequence-structure analysis. ADAM contains 7,007 unique AMP sequences and 759 structures. To our knowledge, this is the first comprehensive study to analyze various AMP structural folds. Our analysis demonstrates that AMP structures cover about ~3% of the overall CATH or SCOP folds. Biologically this infers more than one scheme for AMPs to fight microbes. The results also indicate that AMP structural folds are limited. The majority of the protein structural folds lack antimicrobial activities.

The development of ADAM raises some interesting research topics, which are beyond the scope of this study, still waiting to be explored. To name a few, for example, Table 5 shows that little is known of the structure of Pfam family antimicrobial_1. Such AMP structures need to be resolved by X-ray crystallography or NMR spectroscopy; Table 4 demonstrates a prolonged discussion that CATH and SCOP classifications are not always consistent with each other [21]. The best approach to annotate protein structure is still to be determined. Despite sequence differences between Pfam antimicrobial_2 and DD_K domains, the two domains somehow share the same alpha-helical structural fold: how the two different domains maintain the same structural fold as well as antimicrobial activities still needs more studies.

ADAM, which offers complete AMP sequence and structure information, can benefit a number of different AMP researches such as biomimetics in drug development, comparative immunomics, and structure-function analysis. For example, ADAM cluster #1 (AC_001) has 26 structures associated with 207 AMP sequences (Table S1). Not every structure in the cluster has annotations, but those which do belong to the same CATH and SCOP fold, matching with six different kinds of Pfam families. Such information can help to identify key elements for antimicrobial drug design.

Conflict of Interests

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

Acknowledgments

The work was supported by National Science Council, Taiwan (NSC-102-2221-E-019-060). Hao-Ting Lee, Chen-Che Lee, and Je-Ruei Yang were partially supported by Center for Excellence for the Oceans at National Taiwan Ocean University. In addition, the authors thank Juin-Yiing Huang and De-Hao Chen for technical support.

References

[1] C. D. Fjell, J. A. Hiss, R. E. W. Hancock, and G. Schneider, "Designing antimicrobial peptides: form follows function," *Nature Reviews Drug Discovery*, vol. 11, no. 1, pp. 37–51, 2012.

[2] R. Hammami, A. Zouhir, C. Le Lay, J. Ben Hamida, and I. Fliss, "BACTIBASE second release: a database and tool platform for bacteriocin characterization," *BMC Microbiology*, vol. 10, article 22, 2010.

[3] A. de Jong, A. J. van Heel, J. Kok, and O. P. Kuipers, "BAGEL2: mining for bacteriocins in genomic data," *Nucleic Acids Research*, vol. 38, no. 2, Article ID gkq365, pp. W647–W651, 2010.

[4] M. Novković, J. Simunić, V. Bojović, A. Tossi, and D. Juretić, "DADP: the database of anuran defense peptides," *Bioinformatics*, vol. 28, no. 10, pp. 1406–1407, 2012.

[5] Y. Gueguen, J. Garnier, L. Robert et al., "PenBase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature," *Developmental & Comparative Immunology*, vol. 30, no. 3, pp. 283–288, 2006.

[6] R. Hammami, J. Ben Hamida, G. Vergoten, and I. Fliss, "PhytAMP: a database dedicated to antimicrobial plant peptides," *Nucleic Acids Research*, vol. 37, no. 1, pp. D963–D968, 2009.

[7] C. D. Fjell, R. E. W. Hancock, and A. Cherkasov, "AMPer: a database and an automated discovery tool for antimicrobial peptides," *Bioinformatics*, vol. 23, no. 9, pp. 1148–1155, 2007.

[8] G. Wang, X. Li, and Z. Wang, "APD2: the updated antimicrobial peptide database and its application in peptide design," *Nucleic Acids Research*, vol. 37, no. 1, pp. D933–D937, 2009.

[9] F. H. Waghu, L. Gopi, R. S. Barai, P. Ramteke, B. Nizami, and S. Idicula-Thomas, "CAMP: collection of sequences and structures of antimicrobial peptides," *Nucleic Acids Research*, vol. 42, no. 1, pp. D1154–D1158, 2014.

[10] V. S. Sundararajan, M. N. Gabere, A. Pretorius et al., "DAMPD: a manually curated antimicrobial peptide database," *Nucleic Acids Research*, vol. 40, no. 1, pp. D1108–D1112, 2012.

[11] S. Seebah, A. Suresh, S. Zhuo et al., "Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides," *Nucleic Acids Research*, vol. 35, no. 1, pp. D265–D268, 2007.

[12] S. P. Piotto, L. Sessa, S. Concilio, and P. Iannelli, "YADAMP: yet another database of antimicrobial peptides," *International Journal of Antimicrobial Agents*, vol. 39, no. 4, pp. 346–351, 2012.

[13] H. Jenssen, P. Hamill, and R. E. W. Hancock, "Peptide antimicrobial agents," *Clinical Microbiology Reviews*, vol. 19, no. 3, pp. 491–511, 2006.

[14] L. T. Nguyen, E. F. Haney, and H. J. Vogel, "The expanding scope of antimicrobial peptide structures and their modes of action," *Trends in Biotechnology*, vol. 29, no. 9, pp. 464–472, 2011.

[15] N. Thakur, A. Qureshi, and M. Kumar, "AVPpred: collection and prediction of highly effective antiviral peptides," *Nucleic Acids Research*, vol. 40, no. 1, pp. W199–W204, 2012.

[16] A. Qureshi, N. Thakur, and M. Kumar, "HIPdb: a database of experimentally validated HIV inhibiting peptides," *PLoS ONE*, vol. 8, no. 1, Article ID e54908, 2013.

[17] Y. Li and Z. Chen, "RAPD: a database of recombinantly-produced antimicrobial peptides," *FEMS Microbiology Letters*, vol. 289, no. 2, pp. 126–129, 2008.

[18] P. W. Rose, C. Bi, W. F. Bluhm et al., "The RCSB Protein Data Bank: new resources for research and education," *Nucleic Acids Research*, vol. 41, no. 1, pp. D475–D482, 2013.

[19] A. Andreeva, D. Howorth, J.-M. Chandonia et al., "Data growth and its impact on the SCOP database: new developments," *Nucleic Acids Research*, vol. 36, no. supplement 1, pp. D419–D425, 2008.

[20] I. Sillitoe, A. L. Cuff, B. H. Dessailly et al., "New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures," *Nucleic Acids Research*, vol. 41, no. 1, pp. D490–D498, 2013.

[21] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins: Structure, Function and Genetics*, vol. 57, no. 4, pp. 702–710, 2004.

[22] J. Xu and Y. Zhang, "How significant is a protein structure similarity with TM-score = 0.5?" *Bioinformatics*, vol. 26, no. 7, pp. 889–895, 2010.

[23] M. Punta, P. C. Coggill, R. Y. Eberhardt et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 40, no. 1, pp. D290–D301, 2012.

[24] K. A. Brogden, "Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria?" *Nature Reviews Microbiology*, vol. 3, no. 3, pp. 238–250, 2005.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 402536, 13 pages
http://dx.doi.org/10.1155/2015/402536

Research Article

Predicting Flavin and Nicotinamide Adenine Dinucleotide-Binding Sites in Proteins Using the Fragment Transformation Method

Chih-Hao Lu,[1,2] Chin-Sheng Yu,[3,4] Yu-Feng Lin,[5] and Jin-Yi Chen[1]

[1] *Graduate Institute of Molecular Systems Biomedicine, China Medical University, Taichung 40402, Taiwan*
[2] *Graduate Institute of Basic Medical Science, China Medical University, Taichung 40402, Taiwan*
[3] *Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan*
[4] *Master's Program in Biomedical Informatics and Biomedical Engineering, Feng Chia University, Taichung 40724, Taiwan*
[5] *Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 30068, Taiwan*

Correspondence should be addressed to Chih-Hao Lu; chlu@mail.cmu.edu.tw

Received 16 June 2014; Accepted 21 July 2014

Academic Editor: Hao-Teng Chang

We developed a computational method to identify NAD- and FAD-binding sites in proteins. First, we extracted from the Protein Data Bank structures of proteins that bind to at least one of these ligands. NAD-/FAD-binding residue templates were then constructed by identifying binding residues through the ligand-binding database BioLiP. The fragment transformation method was used to identify structures within query proteins that resembled the ligand-binding templates. By comparing residue types and their relative spatial positions, potential binding sites were identified and a ligand-binding potential for each residue was calculated. Setting the false positive rate at 5%, our method predicted NAD- and FAD-binding sites at true positive rates of 67.1% and 68.4%, respectively. Our method provides excellent results for identifying FAD- and NAD-binding sites in proteins, and the most important is that the requirement of conservation of residue types and local structures in the FAD- and NAD-binding sites can be verified.

1. Background

Over the past 12 years, projects involving structural genomics have generated structural data for ~12,000 proteins within the Protein Data Bank (PDB) [1]. For most of these proteins, however, biological function is unknown. It is therefore important to develop computational methodologies that can identify a protein's function from its structure. Many biochemical processes depend on interactions between proteins and cofactors, such as metal ions, vitamins, and adenine dinucleotides, for example, flavin adenine dinucleotide (FAD) and nicotinamide adenine dinucleotide (NAD). Adenine dinucleotides play important roles in many central biological processes, including DNA repair [2, 3], glycolysis, photosynthesis, and transcription [4–7]. By June 2010, 5293 proteins in PDB were annotated "nucleotide binding," and nucleotides constitute ~15% of biologically relevant ligands [8]. These statistics demonstrate how ubiquitous and essential protein-nucleotide interactions are to biological processes.

Although protein-ligand interactions are fundamental to most biochemical reactions, structural information concerning these binding sites is still inadequate. Once ligand-binding sites can be predicted from structural data, putative functions can be assigned to these proteins. More complete annotation of protein function will benefit both basic science and the pharmaceutical industry. Mutations or deletions within these ligand-binding domains often alter biochemical reactions and are the root causes of many diseases. This makes binding sites attractive targets for drug therapies, including anticancer chemotherapy. In recent years computational methods have been used to identify ligand-binding sites within proteins. These methods include empirical approaches [9], support vector machines (SVM) [8, 10, 11], random forest

[12, 13] and artificial neural networks [14], and structure comparison approaches [15–17]. These prediction methods can be divided into two broad categories: ones that use protein-sequence information, for example, amino acid composition, position-specific scoring matrix, and physicochemical properties, and ones that use protein-structure information, for example, dihedral angles, secondary structure, and 3D-structure comparison. The most effective prediction methodologies, however, tend to use a combination of sequence and structure data.

The structural genomics initiative resolves 20 new protein structures each week, and more than 60,000 structures have been deposited into PDB. The functional surfaces of proteins, which interact with cofactors, tend to be more structurally conserved than internal structures [18]. Residues that form a functional binding region are usually quite close to one another when the three-dimensional structure of a protein is examined. In addition, binding regions typically constitute only 10–30% of the entire protein [19–21]. We took advantage of previously generated structural information and used the fragment transformation method [22] to identify new binding sites for the NAD and FAD ligands.

2. Results

2.1. Residues that Bind NAD or FAD. To characterize the structural environment of NAD-/FAD-binding sites, we compared binding-site residues to whole-protein residues. The three-dimensional structure of the NAD/FAD molecule was divided into three moieties according to function. Within the spherical environment of NAD, the adenosine-binding site typically contained glycine, isoleucine, tyrosine, and aspartic acid residues; the phosphate-binding site contained glycine, isoleucine, serine, threonine, methionine, phenylalanine, tyrosine, tryptophan, arginine, and histidine residues; and the nicotinamide-binding site contained serine, threonine, cysteine, phenylalanine, asparagine, tyrosine, tryptophan, histidine, and asparagine residues. For FAD, adenosine was bound by glycine, valine, cysteine, and tryptophan; phosphate was bound by glycine, serine, and arginine; and flavin was bound by cysteine, methionine, phenylalanine, tyrosine, tryptophan, and histidine. The residue types whose ratio of binding-site residues frequency to whole-protein residues frequency was greater than 1.2 were listed above. As such, the binding residues were primarily polar residues, containing charged groups, amide groups, and nucleophilic groups (Figure 1).

We also characterized the types of atoms that were within 3.5 Å of the three moieties of each NAD/FAD ligand (Figure 2). Nicotinamide and flavin moieties were most commonly associated with nitrogen and oxygen atoms within the backbone and side-chains of the protein. Phosphate moieties were commonly bound by backbone and side-chain nitrogen or side-chain oxygen. Each ligand moiety preferentially bound certain atoms within certain residues.

2.2. Prediction Performance. We chose two criteria to evaluate the performance of our binding-site predictions: performance at less than 5% FPR and the Matthews correlation coefficient (MCC). We used a combination of features that included the number of aligned residues, RMSD, BLOSUM, and DSSP. Using a 5% FPR threshold, NAD-binding sites were predicted with an accuracy of 93.46%, a sensitivity of 67.09%, and an MCC of 0.52. Under these same conditions, FAD-binding-site predictions yielded 93.59% accuracy, 68.43% sensitivity, and an MCC of 0.54 (Table 1). When MCCs were maximized, NAD-binding proteins were identified with 95.34% accuracy, 57.88% sensitivity, 97.64% specificity, and an MCC of 0.57. Under these same conditions, FAD-binding residues were identified with an accuracy of 94.33%, a sensitivity of 64.13%, a specificity of 96.27%, and an MCC of 0.55 (Table 2). These data indicated that our method could predict binding residues for these two ligands.

Table 1: The performance of binding-site predictions at a 5% FPR threshold.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
NAD	93.46	67.09	95.08	0.52
FAD	93.59	68.43	95.22	0.54

Table 2: The performance of binding-site predictions at a maximum MCC threshold.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
NAD	95.34	57.88	97.64	0.57
FAD	94.33	64.13	96.27	0.55

2.3. Comparison with Other Methods. We next compared our results with other prediction methodologies. For these comparisons we chose two published methods that use similar criteria for analyzing these kinds of ligand-protein complexes [10, 11]. These chosen methods assign binding or nonbinding status to each residue within NAD-/FAD-binding proteins. Because these published methods use an equal number of binding and nonbinding residues, we applied our prediction method to a similar dataset to make the results comparable. Random-selection processes were performed five times for all nonbinding residues within ligand-protein complexes to generate the same scale for binding and nonbinding residues within each protein. For NAD-binding proteins, our method predicted binding residues with a sensitivity of 86.21% and an MCC of 0.75 compared with 86.13% and 0.75 for the method developed by Ansari and Raghava [10] (Table 3). For FAD-binding proteins, our method yielded 85.68% sensitivity and an MCC of 0.75. These values compared with the performance of the published method (83.36% and 0.66) developed by Mishra and Raghava [11] (Table 4). Our method, therefore, has similar performance in NAD-binding sites predicted but better in FAD-binding sites. However, in native proteins, the number of binding and nonbinding residues should not be equal. The equal number model needs to be further discussed.

2.4. Template Matching. Figures 3–6 show alignments of predicted NAD-/FAD-binding proteins and corresponding templates. Structures within these figures were drawn using

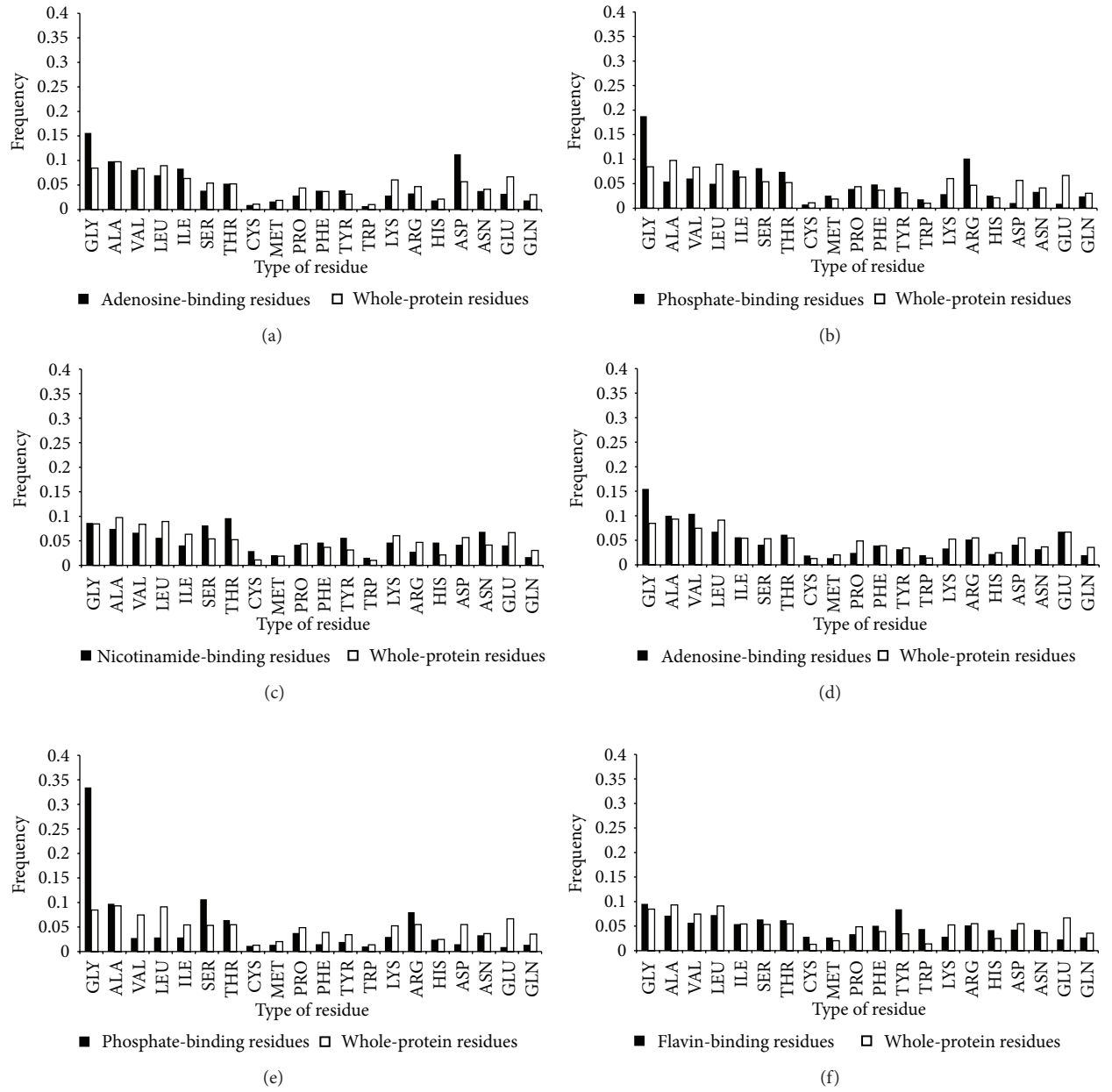


Figure 1: Amino acid frequencies within NAD-/FAD-binding sites. Frequencies within NAD-/FAD-binding sites (black) are compared with whole-protein frequencies (white). (a) Adenosine-binding of NAD. (b) Phosphate-binding of NAD. (c) Nicotinamide-binding of NAD. (d) Adenosine-binding of FAD. (e) Phosphate-binding of FAD. (f) Flavin-binding of FAD. The preferred types of amino acids surrounding the different moiety of NAD/FAD are shown.

Table 3: Comparison between the fragment transformation and SVM methods for predicting NAD-binding-site residues.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
Random 1	87.46	86.45	88.48	0.75
Random 2	87.23	85.79	88.67	0.74
Random 3	87.38	85.65	89.11	0.75
Random 4	87.46	86.91	88.01	0.75
Random 5	87.38	86.25	88.51	0.75
Average	**87.38**	**86.21**	**88.56**	**0.75**
SVM [10]	87.25	86.13	88.37	0.75

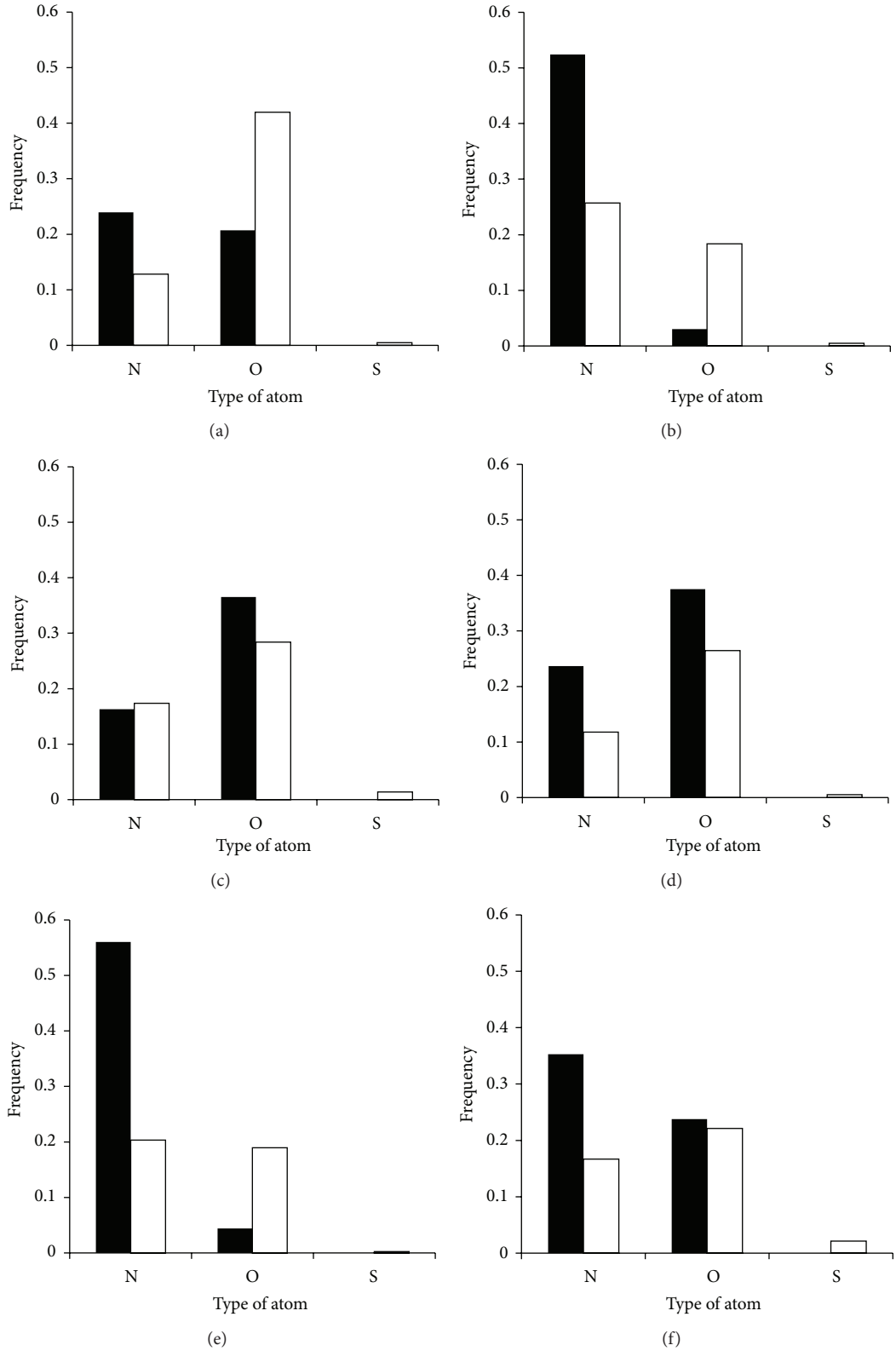


Figure 2: Atom-type frequencies within NAD-/FAD-binding sites. Frequencies for both backbone (black) and side-chain (white) atoms are shown. (a) Adenosine-binding of NAD. (b) Phosphate-binding of NAD. (c) Nicotinamide-binding of NAD. (d) Adenosine-binding of FAD. (e) Phosphate-binding of FAD. (f) Flavin-binding of FAD. The preferred types of atoms surrounding the different moiety of NAD/FAD are shown.

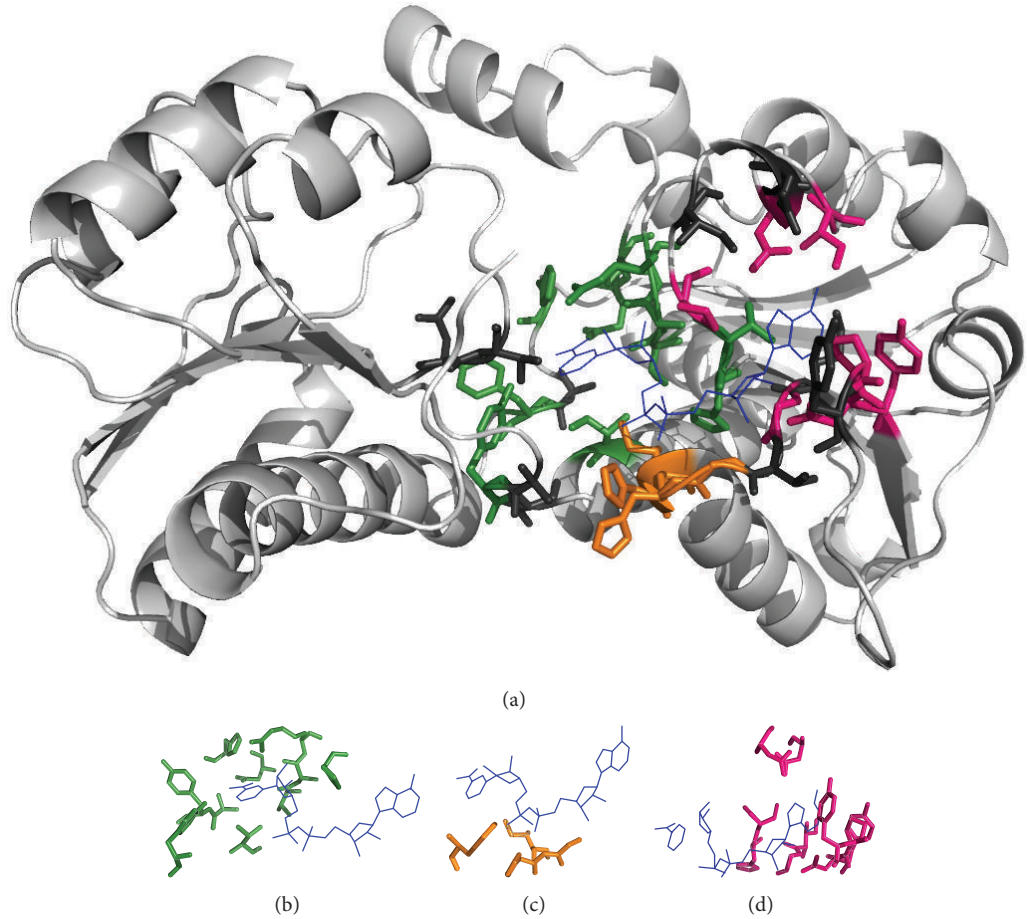


Figure 3: Identification of NAD-binding sites. (a) Chain A of D-2-hydroxyisocaproate dehydrogenase (PDB ID:1DXY) was the query protein. Templates were constructed from (b) D-Lactate dehydrogenase (chain A; PDB ID:3KB6), (c) phosphoglycerate dehydrogenase (chain A; PDB ID:1YBA), and (d) C-terminal-binding protein/brefeldin A-ADP ribosylated substrate (chain A; PDB ID:1HKU).

Table 4: Comparison between the fragment transformation and SVM methods for predicting FAD-binding-site residues.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC
Random 1	87.38	85.68	89.08	0.75
Random 2	87.48	85.73	89.23	0.75
Random 3	87.35	85.55	89.15	0.75
Random 4	87.58	85.73	89.43	0.75
Random 5	87.44	85.73	89.15	0.75
Average	**87.45**	**85.68**	**89.21**	**0.75**
SVM [11]	82.86	83.36	82.36	0.66

PyMOL [23] and color coded: light gray for the query protein; blue lines for the ligand; hot pink, orange, and forest sticks for adenosine-, phosphate-, and nicotinamide-/flavin-binding residues that are predicted correctly; and dark gray sticks for nonbinding residues that are predicted to be binding residues. Our method accurately identified 21 NAD-binding residues within chain A of D-2-hydroxyisocaproate dehydrogenase (PDB ID:1DXY) [24, 25], with ten false positives (Figure 3). Nine nicotinamide-binding residues were identified based on D-Lactate dehydrogenase (chain A; PDB ID:3KB6) [26, 27], three phosphate-binding residues were identified based on phosphoglycerate dehydrogenase (chain A; PDB ID:1YBA) [28], five adenosine-binding residues were identified based on C-terminal-binding protein/brefeldin A-ADP ribosylated substrate (chain A; PDB ID:1HKU) [29], and four were identified based on other protein templates. Our method also accurately predicted 23 NAD-binding residues within chain C of 5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase (PDB ID:2D4E), with only eight false positives (Figure 4). Nine nicotinamide-binding residues were

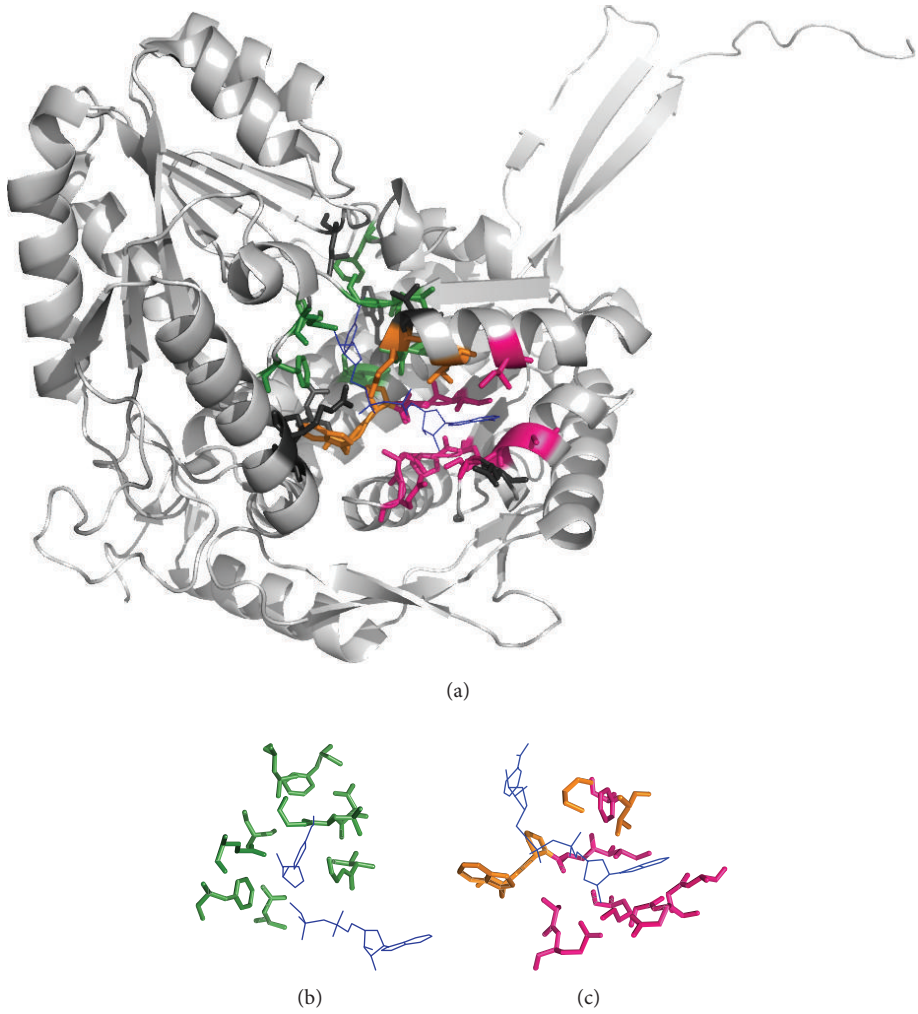


Figure 4: Identification of NAD-binding sites. (a) Chain C of 5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase (PDB ID:2D4E) was the query protein. Templates were constructed from (b) aldehyde dehydrogenase (chain A; PDB ID:3B4W) and (c) l-pyrroline-5-carboxylate dehydrogenase (chain A; PDB ID:2EHU).

identified based on aldehyde dehydrogenase (chain A; PDB ID:3B4W), three phosphate-binding and eight adenosine-binding residues were identified based on l-pyrroline-5-carboxylate dehydrogenase (chain A; PDB ID:2EHU), and three were identified based on other protein templates.

For the FAD-binding proteins, our method accurately predicted chain A of deoxyribodipyrimidine photolyase (PDB ID:1OWL) [30] which contains 24 residues that bind FAD (Figure 5) and only six false positives occurred. Three adenosine-binding residues were identified based on human cryptochrome DASH (chain X; PDB ID:2IJG) [31, 32], six phosphate-binding residues were identified based on photolyase-like domain of cryptochrome 1 (chain A; PDB ID:1U3C) [33], eleven flavin-binding residues were identified based on photolyase (chain A; PDB ID:1IQR) [34], and four were identified based on other protein templates. In addition, 30 FAD-binding residues were accurately predicted within chain H of D-amino acid oxidase (PDB ID:1DDO) [35] with 14 false positives. Five adenosine-binding residues were predicted based on putidaredoxin reductase (chain B; PDB ID:1Q1R) [36, 37], three adenosine-binding and nine flavin-binding residues based on D-amino acid oxidase (chain A; PDB ID:1C0I) [38], three phosphate-binding and five flavin-binding residues based on glycine oxidase (chain B; PDB ID:1NG3) [39], and five based on other protein templates (Figure 6).

3. Discussion

Small molecular cofactors (ligands) are essential for cells to perform numerous biological functions. NAD and FAD, for example, bind to proteins that play critical roles in energy transfer, energy storage, and signal transduction, to name just a few. To understand the mechanism by which these ligands affect protein function, it is important to identify ligand-binding residues within relevant proteins. The experimental identification of these interacting residues is so difficult; however, that computational methods to accomplish this task are in high demand.

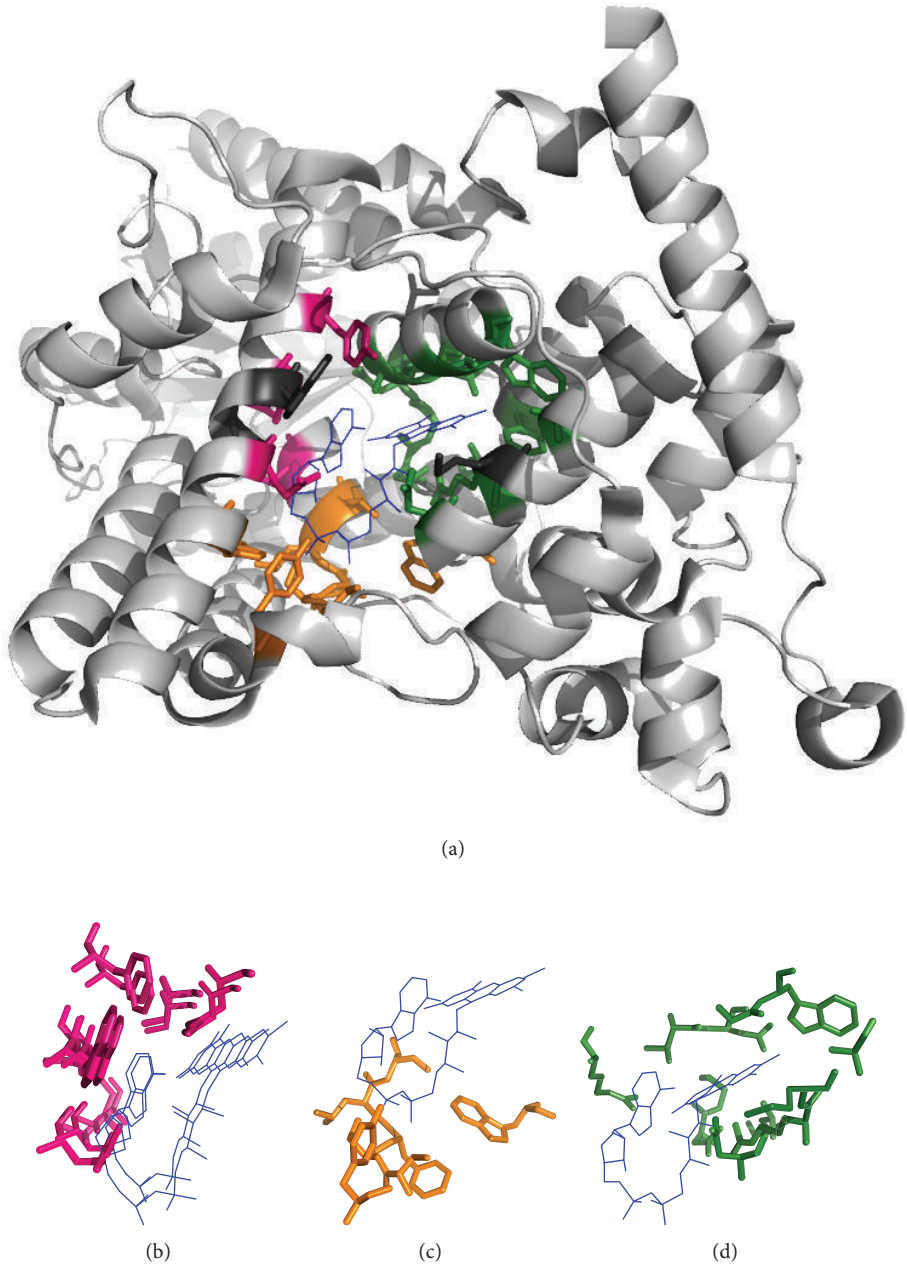


Figure 5: Identification of FAD-binding sites. (a) Chain A of deoxyribodipyrimidine photolyase (PDB ID:1OWL) was the query protein. Templates were constructed from (b) human cryptochrome DASH (chain X; PDB ID:2IJG), (c) photolyase-like domain of cryptochrome 1 (chain A; PDB ID:1U3C), and (d) photolyase (chain A; PDB ID:1IQR).

Here we developed a structure comparison method that uses both sequence and structure information to predict NAD-/FAD-binding residues within proteins. This approach also provides valuable information concerning the microenvironment of the protein-ligand interaction. The composition of NAD-/FAD-binding residues that we identified here is generally similar to previous studies [10, 11]. Interestingly, glycine was the most frequent binding residue, binding to NAD through phosphate or adenosine moieties more often than through the nicotinamide moiety. In contrast, arginine preferentially interacted with phosphate moieties and aspartic acid preferentially interacted with adenosine moieties of NAD, whereas threonine, cysteine, and histidine bound to nicotinamide. The most common residue within FAD-binding sites was also glycine, which preferentially bound phosphate and adenosine moieties. Serine interacted with phosphate moieties, whereas cysteine, tyrosine, and tryptophan primarily bound to nicotinamide. By taking advantage of this kind of structural information, details concerning these critical binding sites may be revealed. To investigate the influence of amino acids on prediction performance, the sensitivity and specificity associated with each

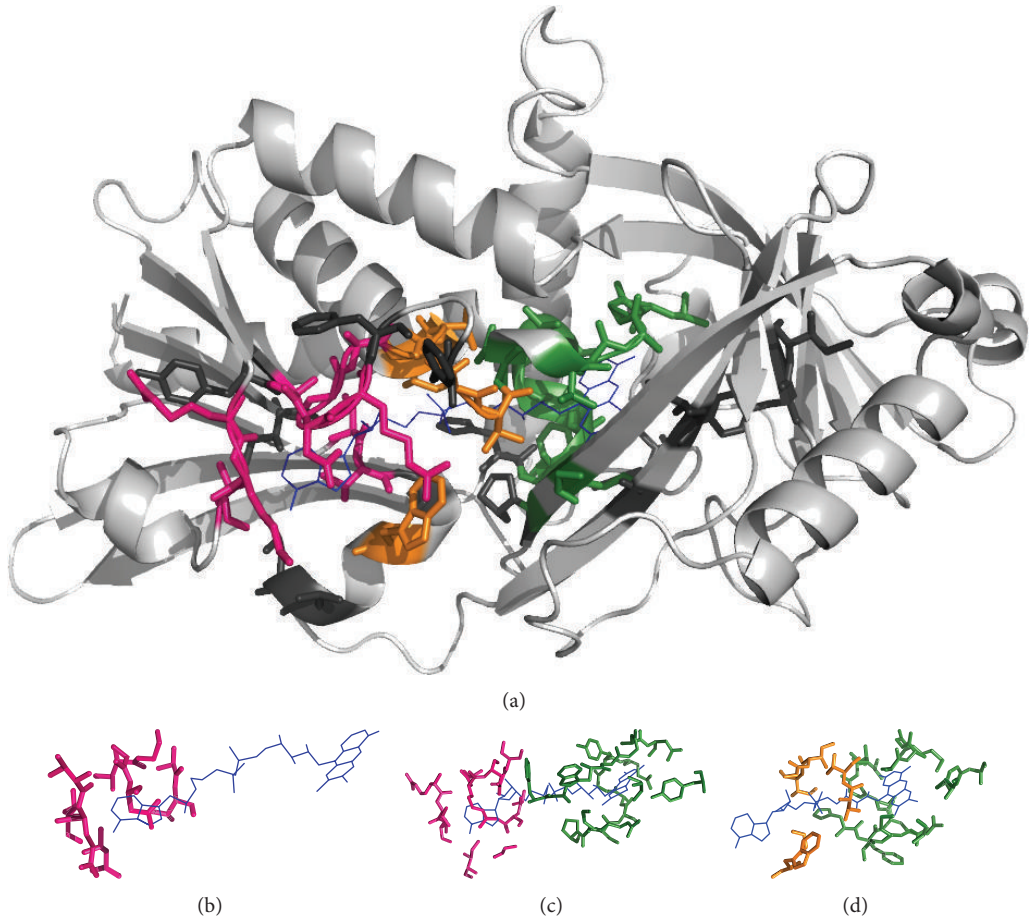


Figure 6: Identification of FAD-binding sites. (a) Chain H of D-amino acid oxidase (PDB ID:1DDO) was the query protein. Templates were constructed from (b) putidaredoxin reductase (chain B; PDB ID:1Q1R), (c) D-amino acid oxidase (chain A; PDB ID:1C0I), and (d) glycine oxidase (chain B; PDB ID:1NG3).

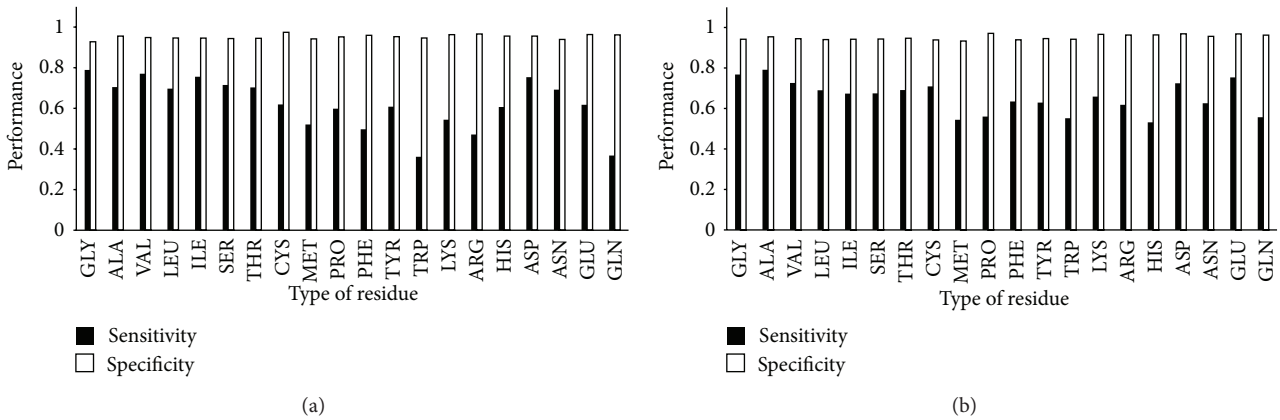


Figure 7: Sensitivity and specificity associated with each amino acid in NAD-/FAD-binding-site predictions. (a) NAD. (b) FAD.

residue were calculated (Figure 7). For NAD-binding-site predictions, specificity for each residue was excellent (0.927–0.966), but sensitivity was relatively low for phenylalanine, tryptophan, arginine, and glutamine which were less than 0.5. For FAD-binding sites, all residues achieved high specificity (0.933–0.971) and sensitivity (0.532–0.791). It should be noted that the ratio of NAD-/FAD-binding residues to nonbinding residues is about 1 to 16 in our dataset. This large difference might cause lots of false positives when predicted. That is the reason for high specificity and accuracy but low sensitivity in our prediction results. Hence, the positions of false positive residues in sequence were also investigated; 20% and 25% of

false positive residues of NAD- and FAD-binding prediction occurred next to the true positive residues in sequence. It was shown that these residues are also located near the ligand in the coordinate space. If these residues were treated as true positive residues, our prediction results of NAD-binding yielded 71.55% sensitivity and 0.61 MCC at a 5% FPR threshold. Under the same conditions, FAD-binding-site predictions yielded 73.34% sensitivity and an MCC of 0.64. Compared with other prediction methods, ours did not use protein evolutionary information but only used protein structure and did not need to use equal number dataset for training but predicted whole-proteins through comparing structures of template database. Our results yielded excellent prediction performance when analyzing NAD-/FAD-binding residues and thus provide important details concerning the binding-site microenvironment. This approach, therefore, may be used to predict putative NAD-/FAD-binding proteins and the specific residues involved in the interaction.

4. Methods

4.1. Overview. We extracted structures of proteins bound to NAD or FAD from PDB and constructed a database of NAD-/FAD-binding residue templates. Residues that were defined as binding residues by the ligand-binding database BioLiP [40] were included in the template. Query protein structures were then compared with each template in the database using a "leave-one-out" comparison method. The fragment transformation method [22] was used to align query and template structures. After comparing the local protein structure, each residue was assigned a score based on both protein sequence and structure. Sequence similarity was calculated using the BLOSUM62 substitution matrix [41], whereas structural similarity was calculated by measuring the root mean square deviation (RMSD) of the Cα carbons from local structure alignments and using a secondary structure substitution matrix [22] according to the Dictionary of Secondary Structure of Proteins' (DSSP) definition of secondary structure [42]. Residues with an alignment score that exceeded a predetermined threshold were predicted to bind NAD/FAD. This method is illustrated in Figure 8.

4.2. NAD-/FAD-Binding Proteins and Binding Residue Templates. We adopted the same datasets with previous research [10, 11]. All protein complexes were collected from PDB and had pairwise sequence identity <40% by using CD-HIT. Proteins chains that are not involved in NAD/FAD binding were excluded. Residues that were defined as binding or nonbinding residues by using the ligand-binding database BioLiP. The main dataset included 184 and 165 polypeptide chains for NAD and FAD, respectively. Because NAD is composed of a nicotinamide moiety, an adenosine moiety, and a phosphate moiety, binding residues were divided into three groups: nicotinamide binding, adenosine binding, and phosphate binding. FAD-binding sitessimilarly contain flavin-binding residues, adenosine-binding residues, and phosphate-binding residues. Groups of residues that contained more than or equal to two binding residues were considered a binding residue template (see Figures 9 and 10).

4.3. The Fragment Transformation Method. We used the fragment transformation method to align NAD-/FAD-binding residues. Each residue was treated as an individual unit and was used to align the query protein S with the binding template T. The structural unit consists of a triplet formed by the N–C$_\alpha$–C atoms within a given residue. S denotes the query protein of length m, and T denotes the template of n residues. The query protein S of length m and the template T of n residues can therefore be expressed in terms of triplets as $S = \{\sigma_1, \sigma_2, \ldots, \sigma_m\}$ and $T = \{\tau_1, \tau_2, \ldots, \tau_n\}$, where $\sigma_i = (p_N, p_{C\alpha}, p_C)$, $\tau_j = (q_N, q_{C\alpha}, q_C)$, and p and q are PDB coordinates for each atom.

A matrix of dimensions $m \times n$ was then constructed for the residues of S and T as

$$M = \begin{vmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,n} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ M_{m,1} & M_{m,2} & \cdots & M_{m,n} \end{vmatrix}, \quad (1)$$

where the element M_{ij} is a rigid-body transformation matrix that transforms the triplet σ_i to τ_j (i.e., $M_{ij}\sigma_i = \tau_j$).

4.4. Performing Triplet Clustering. D_{kl}^{ij} is the Cartesian distance between the target τ_l and the transformed triplet $M_{ij}\sigma_k$, providing a measure of how similarly the triplet pairs (σ_i, τ_j) and (σ_k, τ_l) are oriented. This allows clustering of triplet fragments using the single-linkage algorithm [43] as follows. If for two triplet pairs, (σ_i, τ_j) and (σ_k, τ_l), $D_{kl}^{ij} < D_0$, $i \neq k$ and $j \neq l$, then the triplets are clustered. Let G_1 and G_2 be two clusters, with the first containing (σ_i, τ_j) and (σ_k, τ_l) and the second containing $(\sigma_{i'}, \tau_{j'})$ and $(\sigma_{k'}, \tau_{l'})$. If $D_{k'l'}^{ij} < D_0$, then G_1 and G_2 are merged to form a new cluster G_3, where $G_3 = G_1 \cup G_2$. These procedures are performed iteratively until no new clusters can be formed. For each final cluster G_μ, we can obtain the transformation matrix $M_{k,l}^{\mu}$ and aligned substructure pair $S_\mu = \bigcup_{\sigma_k \in G_\mu} \sigma_k$ and $T_\mu = \bigcup_{\tau_l \in G_\mu} \tau_l$, where G_μ has the minimum Cartesian distance when using $M_{k,l}^{\mu}$.

4.5. Scoring Function. For each residue i, the binding score C_i is defined as

$$C_i = \underset{\sigma_i \in G_\mu}{\mathrm{MAX}} \left(\varepsilon_\mu \times C_\mu^R \times C_\mu^B \times C_\mu^D \right), \quad (2)$$

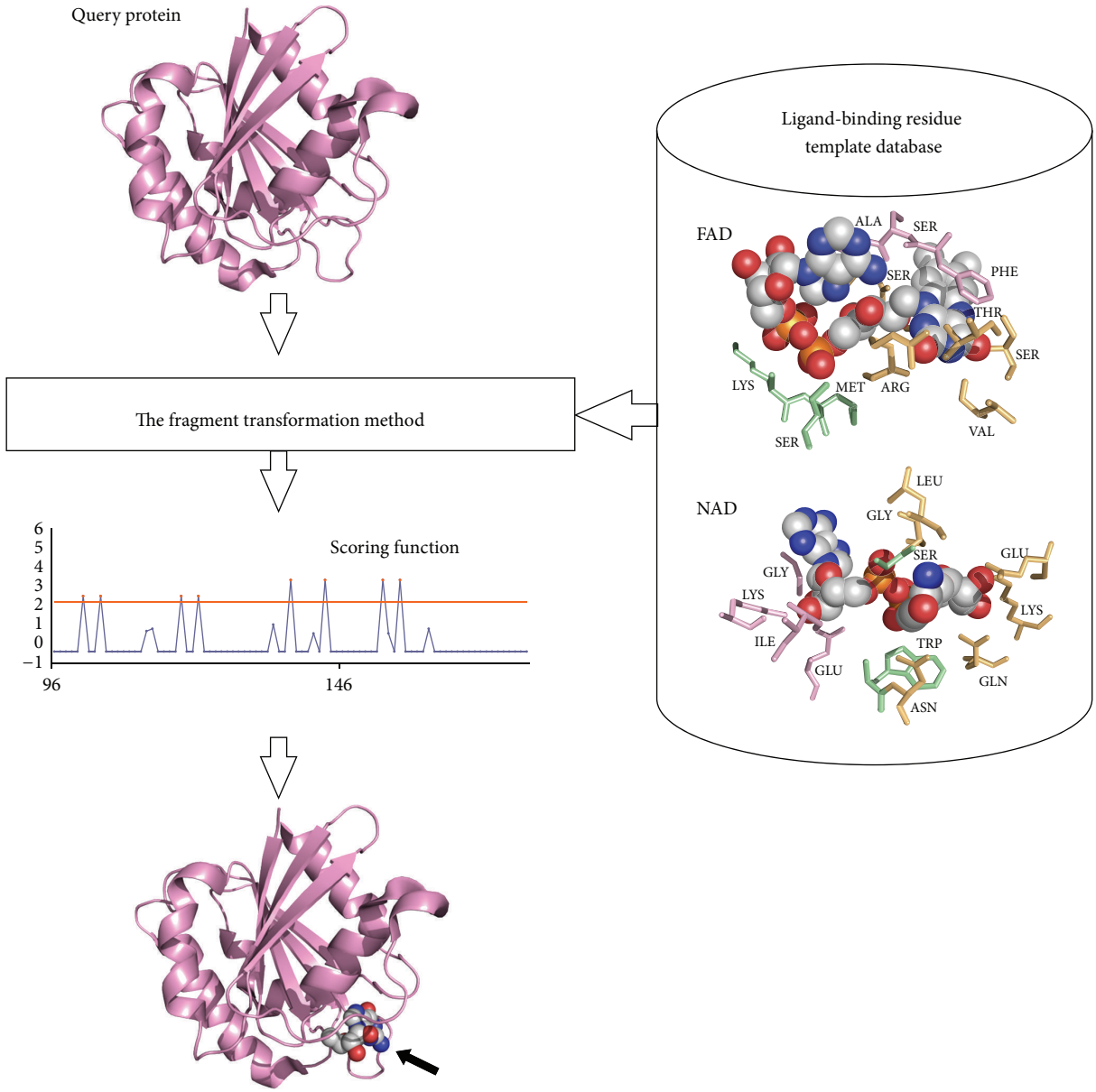


Figure 8: Schematic of the method for predicting NAD-/FAD-binding sites.

where ε_μ is the number of triplets of S_μ (i.e., the aligned residues of the query structure). The alignment scores C_μ^R, C_μ^B, and C_μ^D are defined as

$$C_\mu^R = \frac{1}{1 + \mathrm{RMSD}\left(S_\mu, T_\mu\right)},$$
$$C_\mu^B = \frac{\mathrm{BLOSUM}\left(S_\mu, T_\mu\right)}{\mathrm{BLOSUM}\left(T_\mu, T_\mu\right)} + 1 \quad (3)$$
$$C_\mu^D = \frac{\mathrm{DSSP}\left(S_\mu, T_\mu\right)}{\mathrm{DSSP}\left(T_\mu, T_\mu\right)} + 1,$$

where $\mathrm{RMSD}(S_\mu, T_\mu)$ is the RMSD of all C_α atoms between S_μ and T_μ, $\mathrm{BLOSUM}(S_\mu, T_\mu)$ is the sequence alignment score between S_μ and T_μ calculated using the BLOSUM62 [41] substitution matrix, $\mathrm{BLOSUM}(T_\mu, T_\mu)$ is the maximum sequence alignment score of T_μ, $\mathrm{DSSP}(S_\mu, T_\mu)$ represents the secondary structure alignment score based on a construction substitution matrix [22] using the definition of DSSP [42] between S_μ and T_μ, and $\mathrm{DSSP}(T_\mu, T_\mu)$ is the maximum secondary structure alignment score of T_μ. The value of $\mathrm{RMSD}(S_\mu, T_\mu)$ should be <3 Å.

For each residue i, we predict a geometric center Θ_i^ω of the ligand by $\Theta_i^\omega = M_{k,l}^{\mu\ -1} L_\omega$, where L_ω is the geometric center of the binding template type ω in template T. ω represents the three moieties of NAD/FAD: nicotinamide, adenosine,

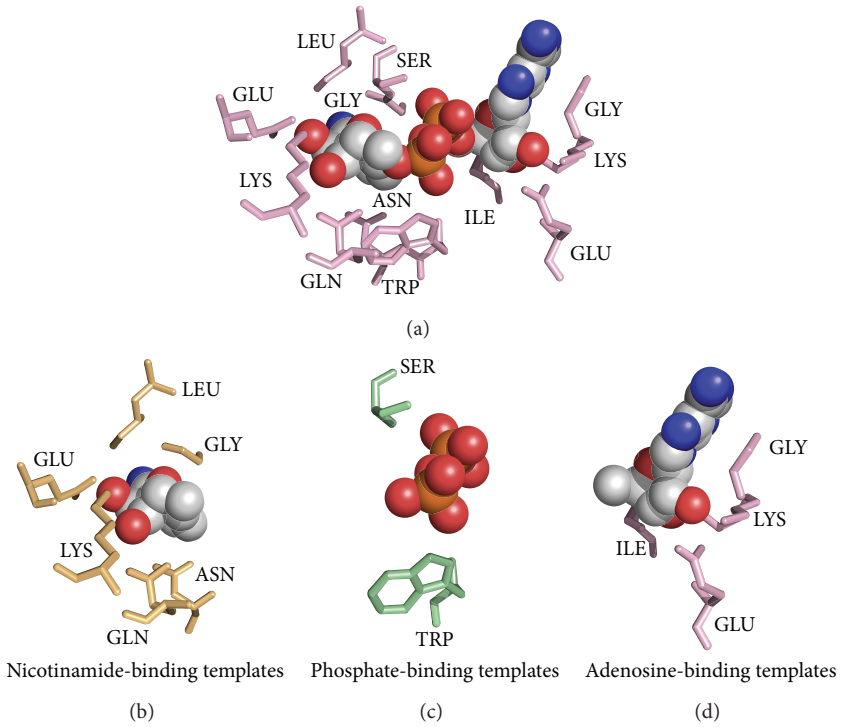


Figure 9: NAD-binding residue templates. (a) The entire NAD-binding template. (b) Nicotinamide-binding templates. (c) Phosphate-binding templates. (d) Adenosine-binding templates.

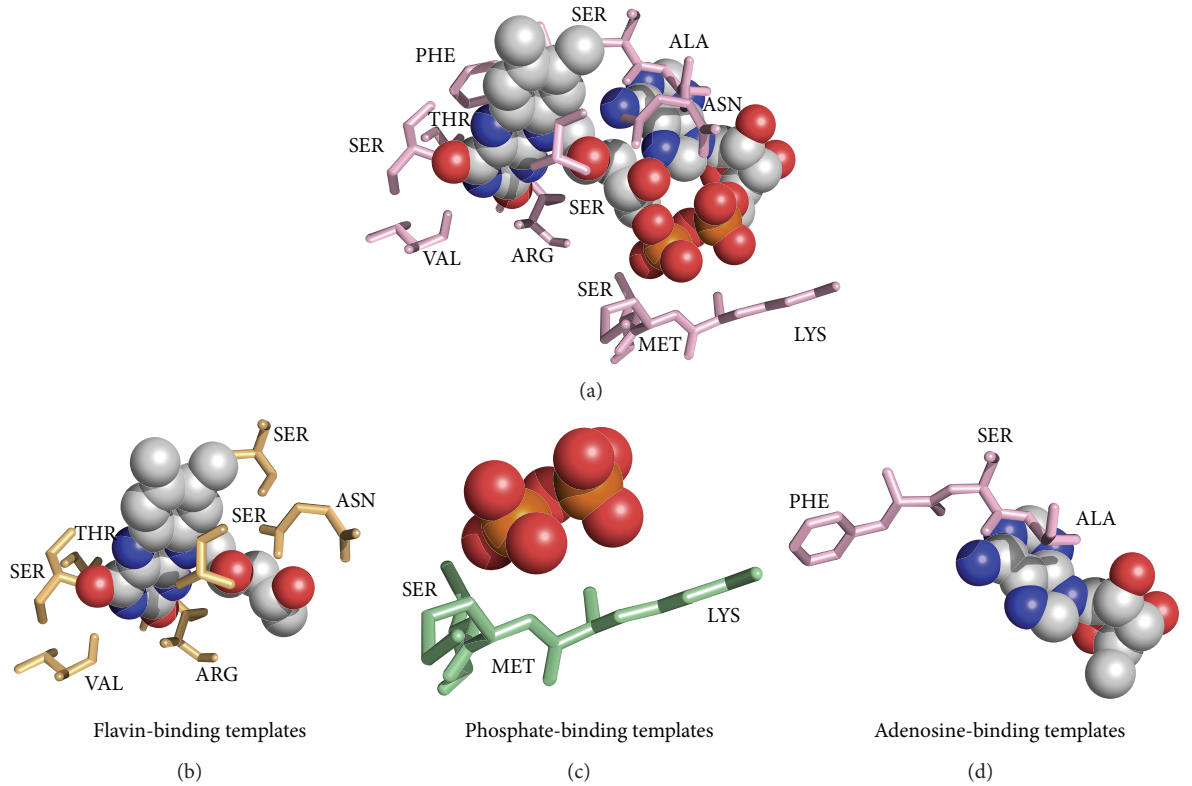


Figure 10: FAD-binding residue templates. (a) The entire FAD-binding template. (b) Flavin-binding templates. (c) Phosphate-binding templates. (d) Adenosine-binding templates.

and phosphate for NAD; flavin, adenosine, and phosphate for FAD. The binding score C_k is added to C_i if the distance between Θ_i^{ω} and $\Theta_k^{\omega'}$ is between 3 and 9 Å, and $\omega \neq \omega'$. Finally, the normalized binding score Z_i^C is calculated as

$$Z_i^C = \frac{C_i - \overline{C}}{SD_C}, \tag{4}$$

where $\overline{C}$ and SD_C denote the mean and standard deviation, respectively, of the binding score C_i.

4.6. Performance Assessment. The accuracy of predicting NAD-/FAD-binding sites wasdefined as the number of true positives and true negatives and was evaluated using a leave-one-out approach. Accuracy (ACC), the true positive rate (TPR), and the false positive rate (FPR) were calculated using true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values as follows:

$$\begin{aligned} \text{ACC} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ \text{TPR} = \text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{FPR} = 1 - \text{Specificity} &= \frac{\text{FP}}{\text{FP} + \text{TN}}. \end{aligned} \tag{5}$$

Conflict of Interests

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

Authors' Contribution

Chih-Hao Lu and Chin-Sheng Yu developed and implemented the methods; Chih-Hao Lu, Chin-Sheng Yu, Yu-Feng Lin, and Jin-Yi Chen carried out the analysis; Chih-Hao Lu and Yu-Feng Lin drafted the paper; Chih-Hao Lu supervised the work. All the authors have read and approved the content of the final paper. Chih-Hao Lu and Chin-Sheng Yu contributed equally to this work.

Acknowledgment

This work was supported by Grants from the National Science Council (NSC 99-2113-M-039-002-MY2) and China Medical University (CMU99-N2-02), Taiwan, to Chih-Hao Lu. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper. The authors are grateful to Yeong-Shin Lin (National Chiao Tung University, Taiwan) for invaluable comments.

References

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

[2] A. Wilkinson, J. Day, and R. Bowater, "Bacterial DNA ligases," *Molecular Microbiology*, vol. 40, no. 6, pp. 1241–1248, 2001.

[3] A. Bürkle, "Physiology and pathophysiology of poly(ADP-ribosyl)ation," *BioEssays*, vol. 23, no. 9, pp. 795–806, 2001.

[4] Q. Zhang, D. W. Piston, and R. H. Goodman, "Regulation of corepressor function by nuclear NADH," *Science*, vol. 295, no. 5561, pp. 1895–1897, 2002.

[5] J. S. Smith and J. D. Boeke, "An unusual form of transcriptional silencing in yeast ribosomal DNA," *Genes and Development*, vol. 11, no. 2, pp. 241–254, 1997.

[6] R. M. Anderson, K. J. Bitterman, J. G. Wood et al., "Manipulation of a nuclear NAD^+ salvage pathway delays aging without altering steady-state NAD^+ levels," *The Journal of Biological Chemistry*, vol. 277, no. 21, pp. 18881–18890, 2002.

[7] J. Rutter, M. Reick, L. C. Wu, and S. L. McKnight, "Regulation of crock and NPAS2 DNA binding by the redox state of NAD cofactors," *Science*, vol. 293, no. 5529, pp. 510–514, 2001.

[8] K. Chen, M. J. Mizianty, and L. Kurgan, "Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors," *Bioinformatics*, vol. 28, no. 3, pp. 331–341, 2012.

[9] M. Saito, M. Go, and T. Shirai, "An empirical approach for detecting nucleotide-binding sites on proteins," *Protein Engineering, Design and Selection*, vol. 19, no. 2, pp. 67–75, 2006.

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

[11] N. K. Mishra and G. P. S. Raghava, "Prediction of FAD interacting residues in a protein from its primary sequence using evolutionary information," *BMC Bioinformatics*, vol. 11, article S48, no. 1, 2010.

[12] Z.-P. Liu, L.-Y. Wu, Y. Wang, X.-S. Zhang, and L. Chen, "Prediction of protein-RNA binding sites by a random forest method with combined features," *Bioinformatics*, vol. 26, no. 13, pp. 1616–1622, 2010.

[13] L. Wang, Z. P. Liu, X. S. Zhang, and L. Chen, "Prediction of hot spots in protein interfaces using a random forest model with hybrid features," *Protein Engineering, Design and Selection*, vol. 25, no. 3, pp. 119–126, 2012.

[14] J. S. Chauhan, N. K. Mishra, and G. P. S. Raghava, "Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information," *BMC Bioinformatics*, vol. 11, article 301, 2010.

[15] A. Roy and Y. Zhang, "Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement," *Structure*, vol. 20, no. 6, pp. 987–997, 2012.

[16] L. Xie and P. E. Bourne, "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 14, pp. 5441–5446, 2008.

[17] J. Yang, A. Roy, and Y. Zhang, "Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment," *Bioinformatics*, vol. 29, no. 20, pp. 2588–2595, 2013.

[18] Y. T. Yan and W.-H. Li, "Identification of protein functional surfaces by the concept of a split pocket," *Proteins: Structure, Function and Bioinformatics*, vol. 76, no. 4, pp. 959–976, 2009.

[19] K. A. Dill, "Dominant forces in protein folding," *Biochemistry*, vol. 29, no. 31, pp. 7133–7155, 1990.

[20] S. Govindarajan and R. A. Goldstein, "Evolution of model proteins on a foldability landscape," *Proteins*, vol. 29, no. 4, pp. 461–466, 1997.

[21] G. Parisi and J. Echave, "Structural constraints and emergence of sequence patterns in protein evolution," *Molecular Biology and Evolution*, vol. 18, no. 5, pp. 750–756, 2001.

[22] C. H. Lu, Y. S. Lin, Y. C. Chen, C. S. Yu, S. Y. Chang, and J. K. Hwang, "The fragment transformation method to detect the protein structural motifs," *Proteins: Structure, Function and Genetics*, vol. 63, no. 3, pp. 636–643, 2006.

[23] L. Schrodinger, *The PyMOL Molecular Graphics System, Version 1.3r1*, 2010.

[24] U. Dengler, K. Niefind, M. Kieß, and D. Schomburg, "Crystal structure of a ternary complex of D-2-hydroxy-isocaproate dehydrogenase from Lactobacillus casei, NAD^+ and 2-oxoisocaproate at 1.9 Å resolution," *Journal of Molecular Biology*, vol. 267, no. 3, pp. 640–660, 1997.

[25] E. Gross, C. S. Sevier, A. Vala, C. A. Kaiser, and D. Fass, "A new FAD-binding fold and intersubunit disulfide shuttle in the thiol oxidase Erv2p," *Nature Structural Biology*, vol. 9, no. 1, pp. 61–67, 2002.

[26] S. V. Antonyuk, R. W. Strange, M. J. Ellis et al., "Structure of d-lactate dehydrogenase from *Aquifex aeolicus* complexed with NAD^+ and lactic acid (or pyruvate)," *Acta Crystallographica F: Structural Biology and Crystallization Communications*, vol. 65, part 12, pp. 1209–1213, 2009.

[27] C. K. Wu, T. A. Dailey, H. A. Dailey, B. C. Wang, and J. P. Rose, "The crystal structure of augmenter of liver regeneration: a mammalian FAD-dependent sulfhydryl oxidase," *Protein Science*, vol. 12, no. 5, pp. 1109–1118, 2003.

[28] J. R. Thompson, J. K. Bell, J. Bratt, G. A. Grant, and L. J. Banaszak, "Vmax regulation through domain and subunit changes. The active form of phosphoglycerate dehydrogenase," *Biochemistry*, vol. 44, no. 15, pp. 5763–5773, 2005.

[29] M. Nardini, S. Spanò, C. Cericola et al., "CtBP/BARS: a dual-function protein involved in transcription co-repression and Golgi membrane fission," *EMBO Journal*, vol. 22, no. 12, pp. 3122–3130, 2003.

[30] R. Kort, H. Komori, S. I. Adachi, K. Miki, and A. Eker, "DNA apophotolyase from Anacystis nidulans: 1.8 Å structure, 8-HDF reconstitution and X-ray-induced FAD reduction," *Acta Crystallographica Section D: Biological Crystallography*, vol. 60, no. 7, pp. 1205–1213, 2004.

[31] Y. Huang, R. Baxter, B. S. Smith, C. L. Partch, C. L. Colbert, and J. Deisenhofer, "Crystal structure of cryptochrome 3 from Arabidopsis thaliana and its implications for photolyase activity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 47, pp. 17701–17706, 2006.

[32] J. B. Thoden, T. M. Wohlers, J. L. Fridovich-Keil, and H. M. Holden, "Molecular basis for severe epimerase deficiency galactosemia. X-ray structure of the human V94M-substituted UDP-galactose 4-epimerase," *The Journal of Biological Chemistry*, vol. 276, no. 23, pp. 20617–20623, 2001.

[33] C. A. Brautigam, B. S. Smith, Z. Ma et al., "Structure of the photolyase-like domain of cryptochrome 1 from Arabidopsis thaliana," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 33, pp. 12142–12147, 2004.

[34] H. Komori, R. Masui, S. Kuramitsu et al., "Crystal structure of thermostable DNA photolyase: pyrimidine-dimer recognition mechanism," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13560–13565, 2001.

[35] F. Todone, M. A. Vanoni, A. Mozzarelli et al., "Active site plasticity in D-amino acid oxidase: a crystallographic analysis," *Biochemistry*, vol. 36, no. 19, pp. 5853–5860, 1997.

[36] I. F. Sevrioukova, H. Li, and T. L. Poulos, "Crystal structure of putidaredoxin reductase from Pseudomonas putida, the final structural component of the cytochrome P450cam monooxygenase," *Journal of Molecular Biology*, vol. 336, no. 4, pp. 889–902, 2004.

[37] S. Y. Song, Y. B. Xu, Z. J. Lin, and C. L. Tsou, "Structure of active site carboxymethylated D-glyceraldehyde-3-phosphate dehydrogenase from Palinurus versicolor," *Journal of Molecular Biology*, vol. 287, no. 4, pp. 719–725, 1999.

[38] L. Pollegioni, K. Diederichs, G. Molla et al., "Yeast D-amino acid oxidase: structural basis of its catalytic properties," *Journal of Molecular Biology*, vol. 324, no. 3, pp. 535–546, 2002.

[39] E. C. Settembre, P. C. Dorrestein, J. H. Park, A. M. Augustine, T. P. Begley, and S. E. Ealick, "Structural and mechanistic studies on thiO, a glycine oxidase essential for thiamin biosynthesis in Bacillus subtilis," *Biochemistry*, vol. 42, no. 10, pp. 2971–2981, 2003.

[40] J. Yang, A. Roy, and Y. Zhang, "BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions," *Nucleic Acids Research*, vol. 41, no. 1, pp. D1096–D1103, 2013.

[41] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 22, pp. 10915–10919, 1992.

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

[43] J. C. Gower and G. J. S. Ross, "Minimum spanning trees and single-linkage cluster analysis," *Journal of the Royal Statistical Society*, vol. 18, no. 1, article 11, 1969.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 974316, 22 pages
http://dx.doi.org/10.1155/2014/974316

Research Article

Computational Biophysical, Biochemical, and Evolutionary Signature of Human R-Spondin Family Proteins, the Member of Canonical Wnt/β-Catenin Signaling Pathway

Ashish Ranjan Sharma,[1,2] Chiranjib Chakraborty,[1,3] Sang-Soo Lee,[1] Garima Sharma,[1] Jeong Kyo Yoon,[4] C. George Priya Doss,[5] Dong-Keun Song,[1] and Ju-Suk Nam[1]

[1] *Institute for Skeletal Aging & Orthopedic Surgery, Hallym University-Chuncheon Sacred Heart Hospital, Chuncheon 200704, Republic of Korea*
[2] *Institute for Skeletal Aging & Orthopedic Surgery, Hallym University Hospital, College of Medicine, Chuncheon-si, Gangwon-do 200-704, Republic of Korea*
[3] *Department of Bioinformatics, School of Computer Sciences, Galgotias University, Greater Noida 203201, India*
[4] *Center for Molecular Medicine, Maine Medial Center Research Institute, 81 Research Drive, Scarborough, ME 04074, USA*
[5] *Medical Biotechnology Division, School of Biosciences and Technology, VIT University, Vellore, Tamil Nadu 632014, India*

Correspondence should be addressed to Chiranjib Chakraborty; drchiranjib@yahoo.com and Ju-Suk Nam; jsnam88@hallym.ac.kr

Received 4 April 2014; Revised 12 July 2014; Accepted 12 July 2014; Published 8 September 2014

Academic Editor: Tun-Wen Pai

In human, Wnt/β-catenin signaling pathway plays a significant role in cell growth, cell development, and disease pathogenesis. Four human (Rspo)s are known to activate canonical Wnt/β-catenin signaling pathway. Presently, (Rspo)s serve as therapeutic target for several human diseases. Henceforth, basic understanding about the molecular properties of (Rspo)s is essential. We approached this issue by interpreting the biochemical and biophysical properties along with molecular evolution of (Rspo)s thorough computational algorithm methods. Our analysis shows that signal peptide length is roughly similar in (Rspo)s family along with similarity in aa distribution pattern. In Rspo3, four N-glycosylation sites were noted. All members are hydrophilic in nature and showed alike GRAVY values, approximately. Conversely, Rspo3 contains the maximum positively charged residues while Rspo4 includes the lowest. Four highly aligned blocks were recorded through Gblocks. Phylogenetic analysis shows Rspo4 is being rooted with Rspo2 and similarly Rspo3 and Rspo1 have the common point of origin. Through phylogenomics study, we developed a phylogenetic tree of sixty proteins ($n = 60$) with the orthologs and paralogs seed sequences. Protein-protein network was also illustrated. Results demonstrated in our study may help the future researchers to unfold significant physiological and therapeutic properties of (Rspo)s in various disease models.

1. Introduction

R-spondins (Rspo)s are a recently discovered family of genes that encodes cysteine-rich secretory proteins containing a thrombospondin type 1 domain/repeat-1 [1]. The (Rspo)s family includes four conserved proteins (Rspo1, Rspo2, Rspo3, and Rspo4), showing overall similarity of 40–60% sequence homology and domain organization [2]. Besides the existence of TSR-1 domain, all four (Rspo)s can be recognized by the existence of a carboxy-terminal region with positively charged amino acids and two furin-like cysteine-rich repeats adjacent to the amino terminus of the mature protein. Numerous studies have implicated (Rspo)s for acting synergistically with extracellular components of the Wnt signaling pathway (Figure 1) [3–5]. Studies showed close or overlapped gene expression of Wnt and (Rspo)s during developmental events, implying a possible coupling of the (Rspo)s with Wnt signaling [6–8]. Consistent with this, a significant reduction in mRNA expression of Rspo1 was observed in a Wnt1/3a double knockout mouse [1]. Rspo1

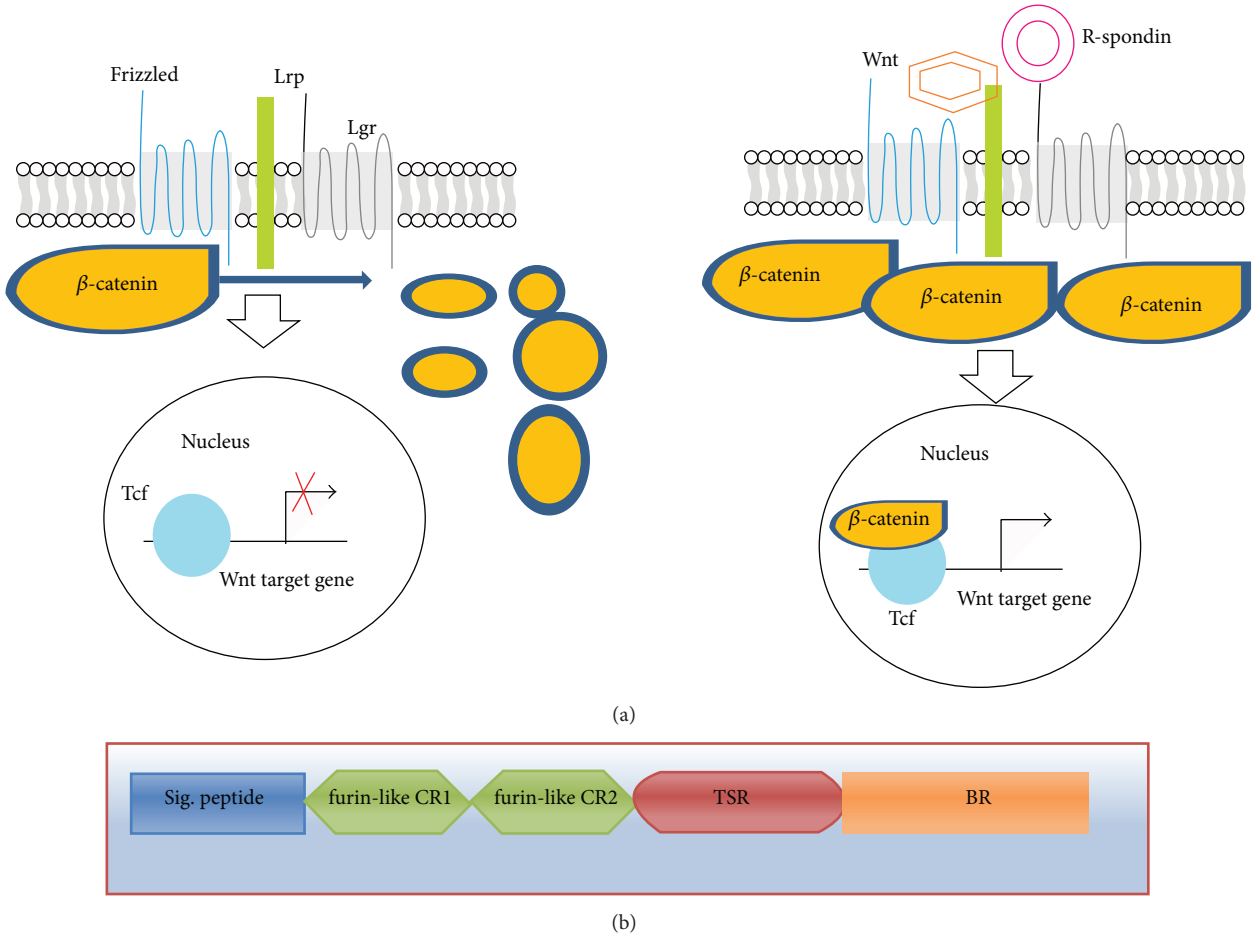


FIGURE 1: The role of (Rspo)s in canonical Wnt signaling pathway and the general architecture of (Rspo)s. (a) Schematic diagram of Wnt and R-spondin signaling models. In absence of Wnt ligand, constitutively synthesized cytoplasmic β-catenin is destroyed by the β-catenin destruction complex causing no β-catenin complex formation with T-cell transcription factor (Tcf)/Lef transcription factors for an active transcriptional response. Canonical Wnt signaling is instigated by the binding of Wnt ligands to the frizzled/LRP receptor complex which in turn deactivates the β-catenin destruction complex increasing its concentration in cytoplasm. The Wnt-frizzled/LRP complex-induced cytoplasmic buildup of β-catenin leads to its import into the nucleus and binding to (Tcf)/Lef transcription factors initiating transcription of Wnt targeted genes. (Rspo)s also act in similar manner but induce this unique property of enhancing Wnt activity by binding to recently discovered seven transmembrane G protein coupled receptors, Lgr (4, 5, and 6). (b) Schematic diagram shows the general domain architecture of human (Rspo)s. The architecture shows (i) signal peptide at N terminal end, (ii) two cysteine-rich furin-like repeats/domains, (iii) a single thrombospondin domain, and (iv) a basic amino-acid-rich domain at C-terminal basic region.

has been shown to augment Wnt signaling by interacting with the low-density lipoprotein receptor related protein 5 or 6 (LRP5/6) coreceptor and inhibiting Dickkopf-1 (Dkk-1) mediated receptor internalization [9]. Rspo2 deficient mice show death at early stages and have limb patterning defects associated with altered Wnt signaling [10, 11]. Rspo3 interacts with Frizzled 8 and LRP-6 and enhances Wnt ligand signaling [3, 4]. In addition to interaction with Wnt/β-catenin signaling, (Rspo)s can also regulate noncanonical Wnt signaling [12]. It was found that furin domain repeats are essential and sufficient for (Rspo)s to mediate Wnt-potentiating effects [13, 14]. Most recently, several studies conclusively determined that the (Rspo)s are the ligands for the leucine-rich repeat containing G protein-coupled receptor 4/5/6 (LGR4/5/6 receptors) [15–18].

Wnt signaling plays a fundamental role during fate determination steps of embryonic development and has been shown to govern process like cell differentiation, cell proliferation, and stem cell maintenance [19, 20]. Due to (Rspo)s ability to function as regulators of Wnt signaling pathways, various potential roles of (Rspo)s have been proposed and have been suggested as novel therapeutic targets [17, 21]. Rspo1 has been shown to control sex phenotypes between individuals. A study by Parma et al. [22] observed sex reversal due to the homozygous Rspo1 gene mutations in affected individuals. In addition, palmoplantar hyperkeratosis and predisposition to squamous cell carcinoma of the skin were also observed in these individuals. Rspo1 has also been recognized as a potent and specific mitogen for the gastrointestinal epithelium [13, 23]. Various studies have also

implicated the importance of Rspo1 in skeletal biology. Rspo1 has been shown to synergize with Wnt3a to promote the process of osteoblast differentiation and inhibit the process of osteoclastogenesis by inducing expression of osteoprotegerin (OPG) [24–26]. Expression of Rspo2 has been shown to promote myogenesis via the Wnt/β-catenin signaling pathway in Xenopus [6]. A study with Rspo2 gene-targeted mutant mice observed that Rspo2 is requisite for normal development of several tissues, including craniofacial structures, lung, kidney, and limbs [27]. Moreover, study reported that Rspo2 is required for the maintenance of apical ectoderm ridge in the hind limbs of the mice. In other studies on Rspo2 mutant mice, hypoplasia and branching defects within the lungs were also being reported [11, 28]. It was observed that Lrp6-mediated Rspo2 signaling via the canonical Wnt pathway is essential for normal morphogenesis of the respiratory tract and for limbs as well [11]. Investigation into the genes responsible for coat features in domestic dogs revealed that Rspo2 is also supported in the Wnt-mediated hair follicle growth [29]. More recently, the role of recurrent Rspo2 gene fusion exclusively with APC mutations has been linked to the activation of Wnt signaling and colon tumorigenesis [30]. Like Rspo2 gene, recurrent Rspo3 gene fusions were also found to be associated with human colon tumors [30]. In recent time, it was proposed that Rspo3 gene may function along with Rspo2 gene in hind limb development, since the knockout of both Rspo2 and Rspo3 in limb mesenchymal cells caused more severe hind limb defects than those of Rspo2 mutant mice [31]. Rspo2 and Rspo3 genes were also identified for their oncogenic potential in mouse mammary tumor virus associated with mammary tumorigenesis in mice [32, 33]. Expression of Rspo4 has been shown to play a key role during nail development and mutations in Rspo4 gene results into absence of the nails in humans termed as anonychia/hyponychia congenita [34].

Given the diverse role of (Rspo)s in dynamic processes of life, like embryogenesis, tumor progression, angiogenesis, myogenesis, development of skeletal system, and so forth, we can expect (Rspo)s as vital therapeutic targets for a number of disabilities. Therefore, we tried to decipher biochemical, biophysical, molecular evolution, and protein-protein interaction characteristics of (Rspo)s by a series of computer based analysis. It may help us to understand the basic molecular properties of these molecules and thus their participation in critical events regulating essential life processes.

2. Materials and Methods

2.1. Data Mining for Human R-Spondin Protein Family Sequences and Their Feature of the Different Regions. We gathered the information on the sequences of human (Rspo)s family members based on searches in the National Centre for Biotechnology Information database (http://www.ncbi.nlm.nih.gov/protein) [35] and UniProt (http://www.uniprot.org/) [36, 37]. The FASTA formats of the sequences were further retrieved for analysis. To investigate the features of the primary structure such as the signal peptide in the protein chain and the chain other than the signal peptide portion, we used UniProt server (http://www.uniprot.org/), a database for information on proteins [36–38]. To understand signal peptide with "C-score" (predicted cleavage site value), "S-score" (the predicted signal peptide value), and "Y-score" (a combination of C- and S-scores), SignalP 4.0 server was used [39]. In addition, different repeats and domain in the R-spondin family members have been analysed using UniProt server.

2.2. Investigation of Amino Acid Distribution, Amino Acid Composition, and Some Parameters Related to the Primary Structure Such as Charge Distribution Analysis, Repetitive Structures, Cysteine Positions, and Disulphide Bonds of Human R-Spondin Family Proteins. To understand the amino acid distribution in the investigated proteins, we used protein calculator (http://spin.niddk.nih.gov/clore/Software/A205.html) [40]. In order to examine the amino acid prototype and protein sequence properties, such as amino acid composition percentage, high scoring hydrophobic segments, and tandem and periodic repeats of structure data of the human (Rspo)s, we used the statistical analysis of protein sequences (SAPS) [41], which is one of the most significant tools to bring out the details about protein sequence properties.

For the study of the secondary structural aspect of R-spondin family members such as cysteine positions and disulphide bond topology prediction, we used "SCRATCH protein predictor" for cysteine positions [42] as well as UniProt (http://www.uniprot.org/) [36, 37] server.

2.3. Structural Prediction of Thrombospondin-1 Domain Type 1 (TSP1) Repeats and Its Molecular Dynamics and Geometry. To understand the thrombospondin-1 domain type 1 repeats, we used the PDB file (1LSL.pdb) extracted from the protein data bank (http://www.rcsb.org); for further analysis see [43, 44]. The structure was visualized using Jmol Applet. We used InterPro, a database for protein families, domains, and functional sites, to understand domain structure [45]. The geometry of thrombospondin-1 domain type 1 repeats such as B factor plot, Omega plot, and FDS (fold deviation score) plot was developed using PDB server. Furthermore, we also developed Ramachandran plot for thrombospondin-1 domain type 1 repeats using PROCHECK server (http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/).

2.4. Prediction of Glycosylation Sites. Analyses of the sequence location of the posttranslational modifications assist to determine the functional characteristics of the proteins. Glycosylation is a type of posttranslational modification (PTM) that assists in protein structural folding, transport, and different types of functions. We predicted the two kinds of glycosylation such as O-glycosylation and N-glycosylation sites by using the NetNGlyc and NetOGlyc servers of the four human (Rspo)s [46–48].

2.5. Prediction of R-Spondin Family Proteins Instability Index, Grand Average of Hydrophobicity (GRAVY), Aliphatic Index,

and Total Number of Positively/Negatively Charged Residues. A comparison of the various biophysical and biochemical parameters of the proteins coded by the human (Rspo)s was carried out using the ProtParam tool from the ExPASy portal (http://web.expasy.org/protparam/) [49]. The different computed parameters for the (Rspo)s includes instability index, aliphatic index, grand average of hydrophobicity (GRAVY), total number of negative charged residues (Asp, Glu), and the total number of positive charged residues (Arg, Lys).

2.6. Prediction of Globularity in the R-Spondin Protein Family. Globular (globe-like) domain of the protein is having spherical domain. The ability to discover the functional sites of domains in proteins is becoming increasingly important. GlobPlot was used to predict the globularity in the domains [17]. The algorithm was as follows:

$$\Omega(a_i) = \sum_{j=1}^{i=1} \Omega(a_j) + \ln(i+1) \cdot P(a_i) \quad \text{for } i = 1, \ldots, L. \tag{1}$$

For the protein sequence which is used for analysis, the length of the sequence is L; Linding et al. [50] defined the sum function Ω as $P(a_i) \in R$. $P(a_i)$ is the propensity of the ith amino acid and ln is the natural logarithm. The globularity in the domains of the regulatory subunit p85α was determined using the GlobPlot Web server.

2.7. Multiple Sequences Alignment (MSA) Analysis among R-Spondin Family Proteins. Four sequences of R-spondin family proteinswere used to understand the sequences similarity and alignment positions using MSA analysis. For that, we used clustal-omega to understand the sequence similarities and to elucidate the respective pairwise alignment scores. Clustal-omega has a graphical interface that is easy to use [51]. The clustal-omega server was organized on the basis of "progressive algorithm" [52] and the scoring system of the pairwise alignment algorithm is possibly the powerful component of the progressive algorithm. During the best alignment between N sequences, a computational complexity is found (L^N) for N sequences of length L. The basic algorithm to elucidate respective pairwise alignment scores is based on Needleman and Wunsch's algorithm [53].

Additionally, other MSA tools were used known as "multiple sequence comparison by log-expectation" (MUSCLE) to locate the conserved pattern across R-spondin protein family [54]. MUSCLE uses a function that can be described as the following log-expectation (LE) score function:

$$\mathrm{LE}^{xy} = \left(1 - f_G^x\right)\left(1 - f_G^y\right) \log \sum_i \sum_j f_i^x \frac{f_j^y p_{ij}}{p_i p_j}. \tag{2}$$

This function is a modified version of the log-average function expressed as follows:

$$\mathrm{LA}^{xy} = \log \sum_i \sum_j f_i^x \frac{f_j^y p_{ij}}{p_i p_j}, \tag{3}$$

where i and j are amino acid types; p_i is the background probability of I; p_{ij} is the joint probability of i and j being aligned to each other; f_i^x is the observed frequency of i in column x of the first profile; and f_G^x is the observed frequency of gaps in that column at position x in the family and likewise for position y in the second profile. The approximate probability α_i^x of experimental amino acid i in location x can be derived from fx. The graphical yield of MUSCLE was visualised through JalView. Finally, Gblocks server was used to observe the aligned blocks of the sequences, which describes a set of conserved blocks from an MSA according to a set of simple requirements [55].

2.8. Multiple Sequences Alignment (MSA) Analysis of R-Spondin Family Proteins with Other Species. To understand the sequence similarity of human four (Rspo)s with other species, we used PhylomeDB server [56, 57]. This server performed homology searches by means of the Smith-Waterman algorithm [58] and ultimately filtered the sequences according to specific e-value and overlap cut-offs.

2.9. Analysis of Molecular Phylogenetics of Human R-Spondin Family Proteins. For the molecular phylogenetics, we used three servers to develop two phylogenetic trees. First we used accessible computer software and constructed the phylogenetic tree using Phylogeny.fr and performed computational biology [59]. This software uses several kinds of software for the workflow such as MUSCLE multiple alignment, Gblocks for the alignment curation, PhyML for the construction of the phylogenetic tree, and TreeDyn for the visualisation of phylogenetic tree. We have developed two types of the phylogenetic tree, namely, phylogram and cladogram (without branch distance). The phylogram depicted distances among protein sequences within the(Rspo)s. Then, another tree known as the "circular alpha phylogenetic tree" has been developed using MAFFT (version 7) [60]. Again using the four family sequences, we used clustal-omega to develop another phylogenetic tree [51]. The servers implemented either a neighbour-joining method or the bottom-up clustering method developed by Saitou and Nei [61] and the algorithm used a distance matrix to specify the distance between each pair of taxa. In this case, the matrix had a magnitude which is $N \times N$. In this case, N is the number of points or nodes.

2.10. Prediction of Phylogenomics of Human R-Spondin Family Proteins Using Molecular Phylogenetics to Understand the Framework Topology of Other Related Species. To understand the phylogenomics of four human (Rspo)s and framework topology of other related species, we developed another phylogenetic tree using the sequence similarity of four human (Rspo)s with other species. For this analysis, we use PhylomeDB server, one of the largest phylogenetic repository [56, 57]. This server performed homology searches by means of the Smith-Waterman algorithm [58] and ultimately filtered the sequences according to specific e-value and overlap cut-offs. The server is a resulting collection of trees which characterize the full complement of evolutionary histories

of all genes determined in a given genome. This has been entitled with the term phylome [59]. For phylogenomics analysis, the method used in this study is more closely a gene-centered method. And it is computationally more extensive compared to developing a family-based approach.

2.11. Understanding the Protein-Protein Interaction Network of R-Spondin Family Proteins. We have developed protein-protein interaction network using STRING server to understand the possible protein interactions with (Rspo)s [62, 63]. We developed four interaction networks, one for each (Rspo). Finally, we also developed scores to understand the interaction among possible interacting proteins with (Rspo)s.

3. Results

3.1. Searched Data for Corresponding Proteins and Their Features Such as Signal Peptide, Repeats, and Domains. Supplementary Table S1 (available online at http://dx.doi.org/10.1155/2014/974316) shows the protein sequence information related to the human (Rspo)s analysed in this study, while the genes and proteins information related to human (Rspo)s have been displayed in Supplementary Table S2. The sequence lengths of Rspo1, Rspo2, Rspo3, and Rspo4 have been plotted in Figure 2(a). The figure shows that Rspo3 contains the highest sequence length of amino acids (aa) (272 aa), while Rspo4 contains the lowest sequence length (234 aa). Next, we plotted the sequence of amino acid number in the scattered distribution ($R^2 = 0.1824$) (Figure 2(b)). The features of the primary structure such as the signal peptide in the protein chain and the chain other than the signal peptide and information of different regions such as repeat and domain of human R-spondin family members were analysed. We depicted the position of regions, length, and graphical view of such regions in Figures 2(c), 2(d), 2(e), and 2(f). Thereafter, we compared the amino acid length of the signal peptide in the protein chains and the chain other than the signal peptide of these four proteins (Figure 2(g)). We observed that the length of the signal peptide portions is more or less similar (19 to 21 aa length) among the four proteins. Conversely, differences in the amino acid length have been noted in the chain other than the signal peptide portion where Rspo3 comprises the highest sequence length (251 aa) while Rspo4 contains the lowest sequence length (215 aa). Furthermore, we have analyzed the signal peptides of four human (Rspo)s and depicted their "C-score" (predicted cleavage site value), "S-score" (the predicted signal peptide value), and "Y-score" (a combination of C- and S-scores) (Figure 3).

3.2. Investigation of Amino Acid Distribution, Amino Acid Composition, and Some Parameters Related to the Primary Structure Such as Charge Distribution Analysis, Repetitive Structures, Cysteine Positions, and Disulphide Bonds of Human R-Spondin Protein Family. Amino acid distributions of human R-spondin protein family have been reprinted in Figures 4(a), 4(b), 4(c), and 4(d). Furthermore, we exposed the four amino acid distributions at a time to understand the distribution pattern of these proteins (Figure 4(e)). The composition analysis of the amino acids of human (Rspo)s has been represented in Supplementary Table S3 and Supplementary Figure S1. From the calculated distribution of amino acids as well as the composition of the amino acids of human (Rspo)s, we found the following data: Rspo1 with highest Arg 27 (10.3%) and lowest Trp 3 (1.1%) and Tyr 3 (1.1%) both, Rspo2 with highest Arg 38 (11.1%) and lowest Trp 4 (1.6%), Rspo3 with highest Cys 22 (8.1%) and Ser 22 (8.1%) both and lowest Trp 3 (1.1%), and Rspo4 with highest Gly 27 (11.5%) and Arg 27 (11.5%) both and lowest Trp 3 (1.3%), respectively. From the distribution and composition of amino acid, it was noted that the highest of number of Arg residue was noted in the three proteins (Rspo1, Rspo2, and Rspo4), and the lowest number of residue was Trp in all of the R-spondin family proteins. The charge distribution analysis, repetitive structures, and cysteine positions of human (Rspo)s has been noted in Supplementary Table S3. Total numbers of cysteine and disulphide bonds present among (Rspo)s have been illustrated in Figure 4(f) showing maximum number of cysteine residues in Rspo2 protein (twenty-four). However, disulphide bonds are same in all of the human (Rspo)s (eleven).

3.3. Structural Prediction of Thrombospondin-1 Type 1 (TSP1) Repeats and Its Geometry. Human R-spondin family proteinscontain a thrombospondin type 1 domain type 1 repeats [1] (Figure 5(a)). The structure of monomeric assembly of the thrombospondin type 1 domain type 1 repeats has been depicted in Figure 5(b). This domain structure has been illustrated through the CATH and Pfam database and described in Figures 5(c) and 5(d). The surface structure of this domain has been developed with atomic properties described through different colours (Figure 5(e)). TSP1 domain(s) has been identified in a number of proteins, but generally in multiple copies. From this aspect, R-spondin is very unique since it has only one copy and predicted structure of this domain is hinge-like structure. This specific hinge-like structure of TSP1 domain may play a vital role in binding activity with the receptors. It has been found that TSR motifs especially the WSGWSSCSVSCG sequence are most significant for different neuronal responses such as neurite extension, neuronal survival, neuronal aggregation, and so forth [64].

B factors plot signifies the convolution of static and dynamic disorder in the crystal structure. While, dynamic disorder present in a crystal can be recognized through the local motions of individual atoms. Conversely, static disorder signifies the different atomic positions in a particular protein molecule [65]. Omega plot is helpful to understand the proper residue. Fold Deviation Score (FDS) plot is important to understand the structural geometry of the protein [32]. Ramachandran plot is also significant to comprehend residues in a generously allowed region [66]. Therefore, we developed the geometry of the thrombospondin type 1 domain type 1 repeats and the associated different geometry of these domain, such as B factor Plot, Omega plot, Fold Deviation Score (FDS) plot and Ramachandran plot and

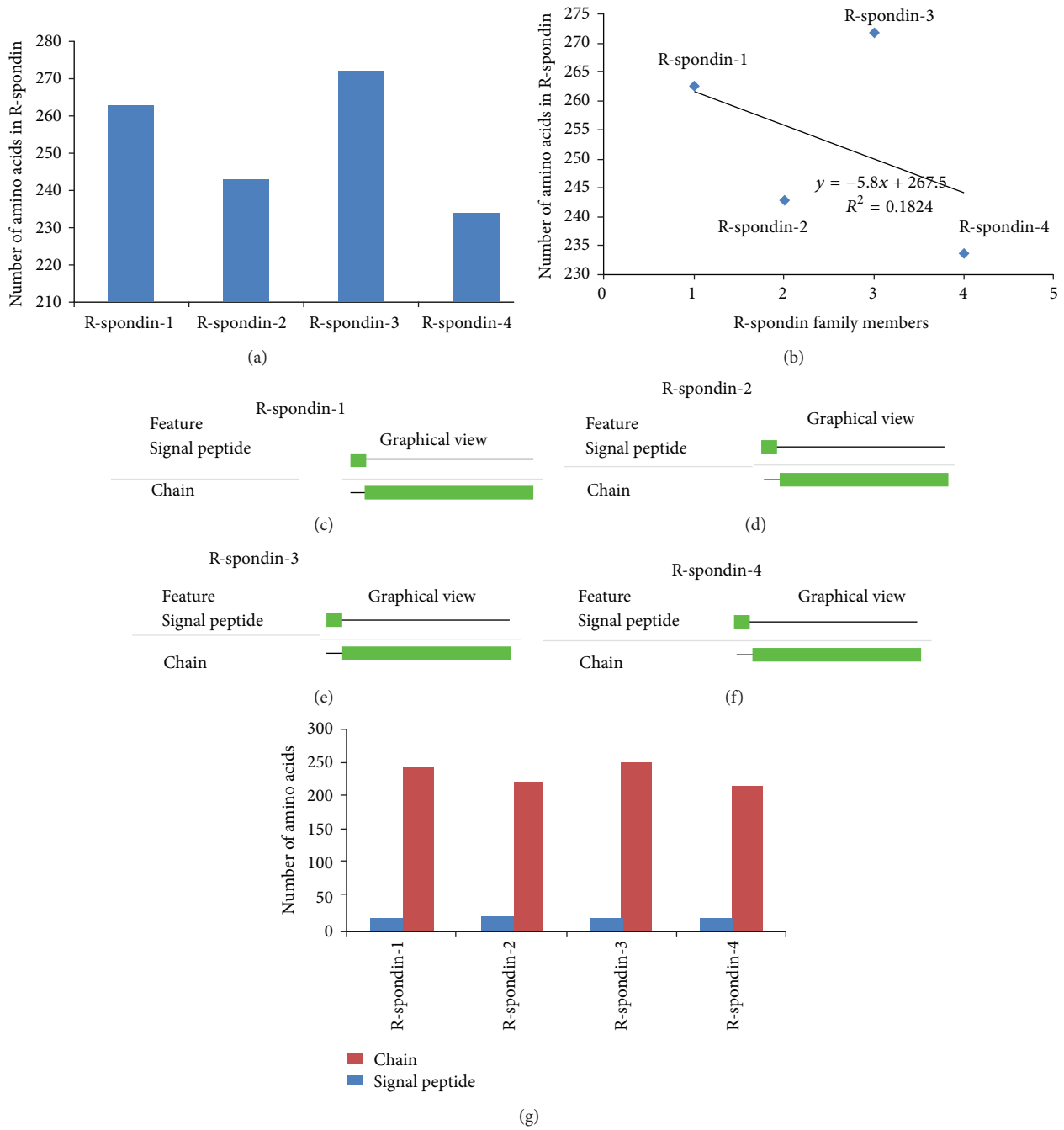


Figure 2: General architecture of human (Rspo)s in respect to amino acid sequence. (a) Comparison between the numbers of amino acids in all of the R-spondin family proteins. (b) Plot showing the scattered distribution of amino acid numbers along with the (Rspo)s and their correlations. ((c), (d), (e), (f)) Graphical overview of signal peptide and other parts of the amino acid chain-Rspo1, Rspo2, Rspo3, and Rspo4. (g) Comparison between number of amino acids in signal peptide and other parts of the protein in four human (Rspo)s.

recorded in Supplementary Figures S2(a), S2(B), S2(C), and S2(D), respectively.

3.4. Prediction of Glycosylation Sites. Similar to phosphorylation, in some eukaryotic proteins, glycosylation plays a significant role in protein function and interaction during the signalling process [67]. In biophysical and biochemical point of view, N-glycosylation sites and O-glycosylation sites are important for functionality of the protein. In reviewing the presence of N-glycosylation sites (Supplementary Table S5) among (Rspo)s, we found the following: Rspo1 with 1 site (at the residue position of 137), Rspo2 with 1 site (at the residue position of 160), Rspo3 with 4 sites (at the residue position of 23, 36, 137 and 194) and Rspo4 with 1 site. The results of Rspo3

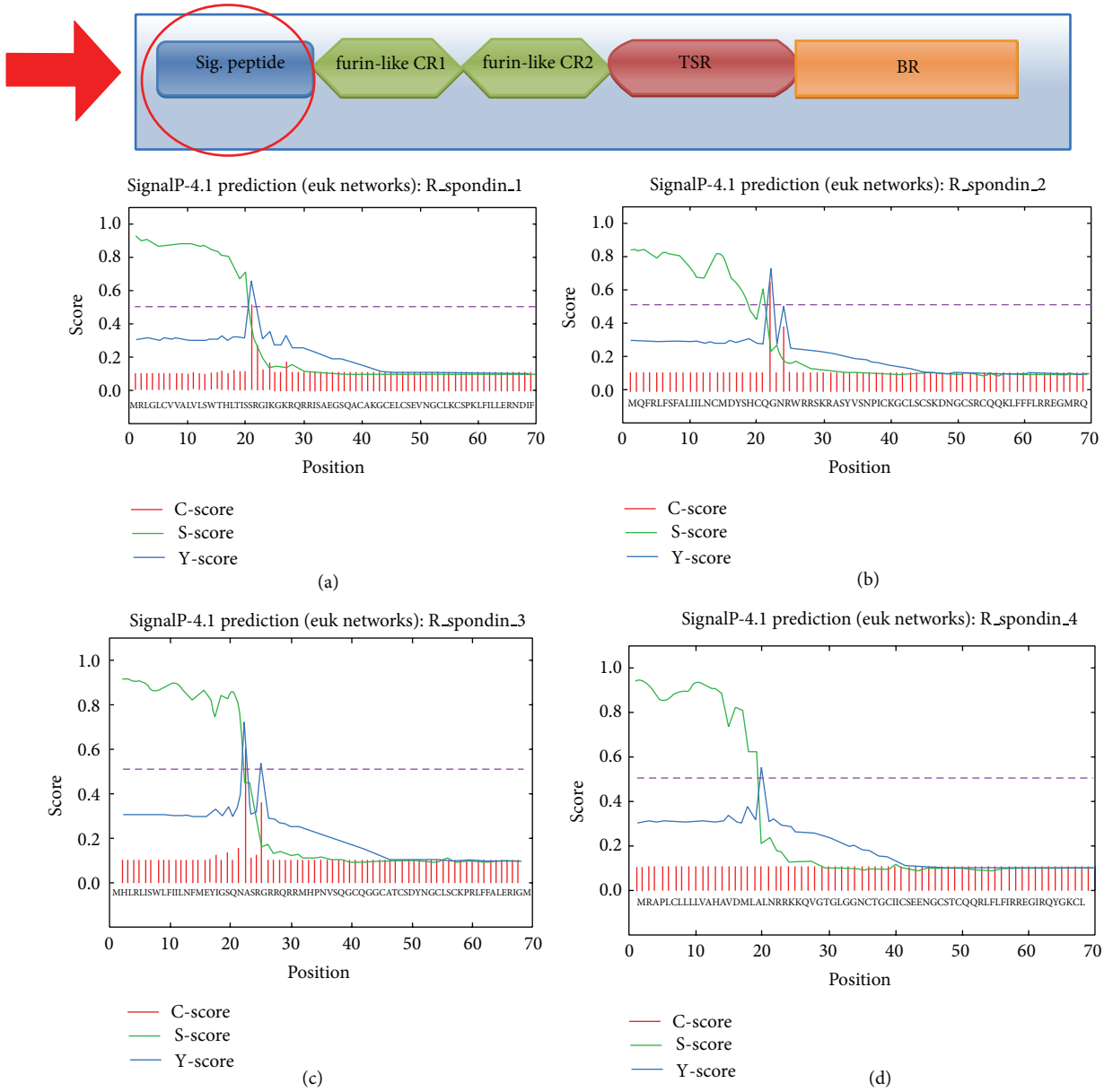


Figure 3: Predicted architecture of signal peptide of different human (Rspo)s with "C-score," "S-score," and "Y-score" (C-score represents predicted cleavage site value; S-score represents the predicted signal peptide value; Y-score represents a combination of C- and S-scores). (a) Rspo1, (b) Rspo2, (c) Rspo3, and (d) Rspo4 (the schematic diagram shows the location of signal peptide in the general domain architecture of human (Rspo)s and our region of analysis).

showed highest N-glycosylation sites. While reviewing the O-glycosylation potentiality and location (Supplementary Table S6), we found that only Rspo1 has one site. No other (Rspo)s have O-glycosylation sites. However, several O-glycosylation sites potentialities were recorded among (Rspo)s; although, the values of these sites were below the threshold limit (Figure 6).

3.5. Prediction of R-Spondin Family Proteins Instability Index, Grand Average of Hydrophobicity (GRAVY), Aliphatic Index, and Total Number of Positively/Negatively Charged Residues. The protein stability is associated with different structural properties and functionality of the proteins such as metabolic stability [68], protein-protein interactions [69], and so forth. An instability index provides the knowledge about a protein's stability, in particular in an *in vitro* environment. The instability index value greater than 40 designates an unstable protein, and one less than 40 designates a stable protein. Several factors such as the arrangement of amino acids in a sequence and some peptide bonds make *in vivo* proteins stable [70]. The results of our instability index analysis of the R-spondin family proteins are shown in Figure 7(a). The Rspo1 was found to have the highest instability index, whereas

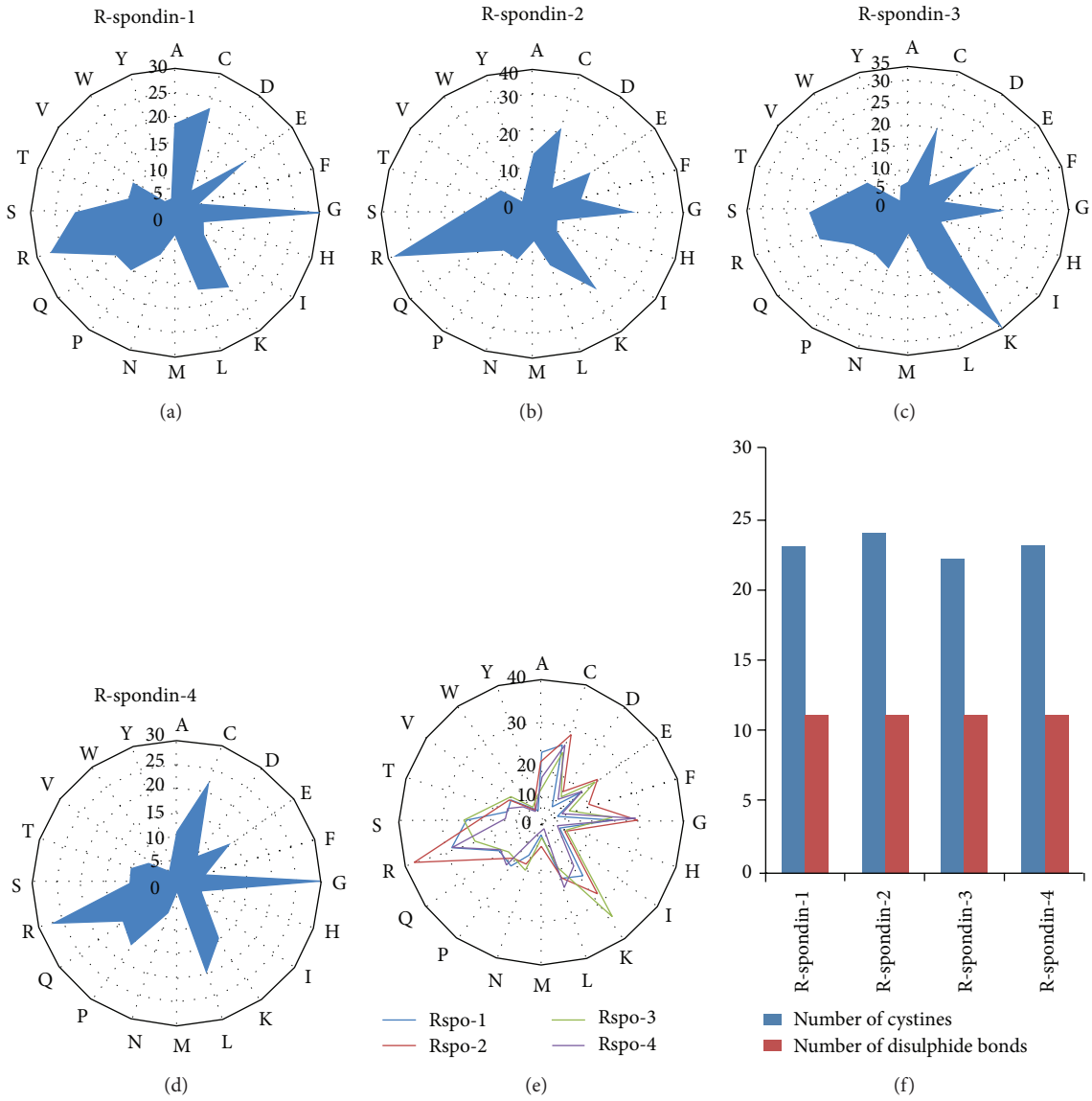


Figure 4: General amino acid distribution of amino acids in human (Rspo)s. (a) Rspo1, (b) Rspo2, (c) Rspo3, (d) Rspo4, (e) a general trend of amino acid distribution of amino acids for all human (Rspo)s where we have exposed the four protein's amino acid distribution at a time, and (f) comparison between number of cystine residue and disulphide bond in four human (Rspo)s.

Rspo3 was found to have the lowest. Every R-spondin protein was found to be unstable as per their instability index, since the values are greater than 40. The changes in amino acid composition and hydrophobicity may have caused the observed distinct stability of the protein.

Kyte and Doolittle have formulated the scale of hydropathy in which the hydrophilic and hydrophobic possessions of amino acid chain are assessed in a protein [71]. Grand average of hydrophobicity (GRAVY) score can be computed as the sum of the hydropathy values for all the amino acids in a protein that can be divided by the total number of residues in the protein. Grand average of hydrophobicity (GRAVY) is associated with protein solubility. It has been noted that the positive GRAVY value is positively associated with hydrophobicity and negatively associated with the hydrophilicity. Because a more hydrophilic protein forms a larger amount of hydrogen bonds with water, therefore, the solubility is more. A ProtParam GRAVY study predicted grand average of hydrophobicity in the (Rspo)s (Figure 7(b)). Our analysis revealed that all (Rspo)s were hydrophilic in nature, Rspo3 being the most hydrophilic. The GRAVY value shows approximate similar values for the Rspo1, Rspo2, and Rspo4 (−0.717, −0.769, and −0.701, resp.).

The aliphatic index (AI) is very significant for understanding a protein, as it describes the relative volume occupied by aliphatic side chains such as alanine, valine, isoleucine and leucine. Aliphatic hydrophobicity is amplified with a rise in temperature and is, therefore, a positive factor enhancing the thermal stability of globular proteins [72]. Our analyses (Figure 7(c)) showed that, Rspo4 have the highest aliphatic

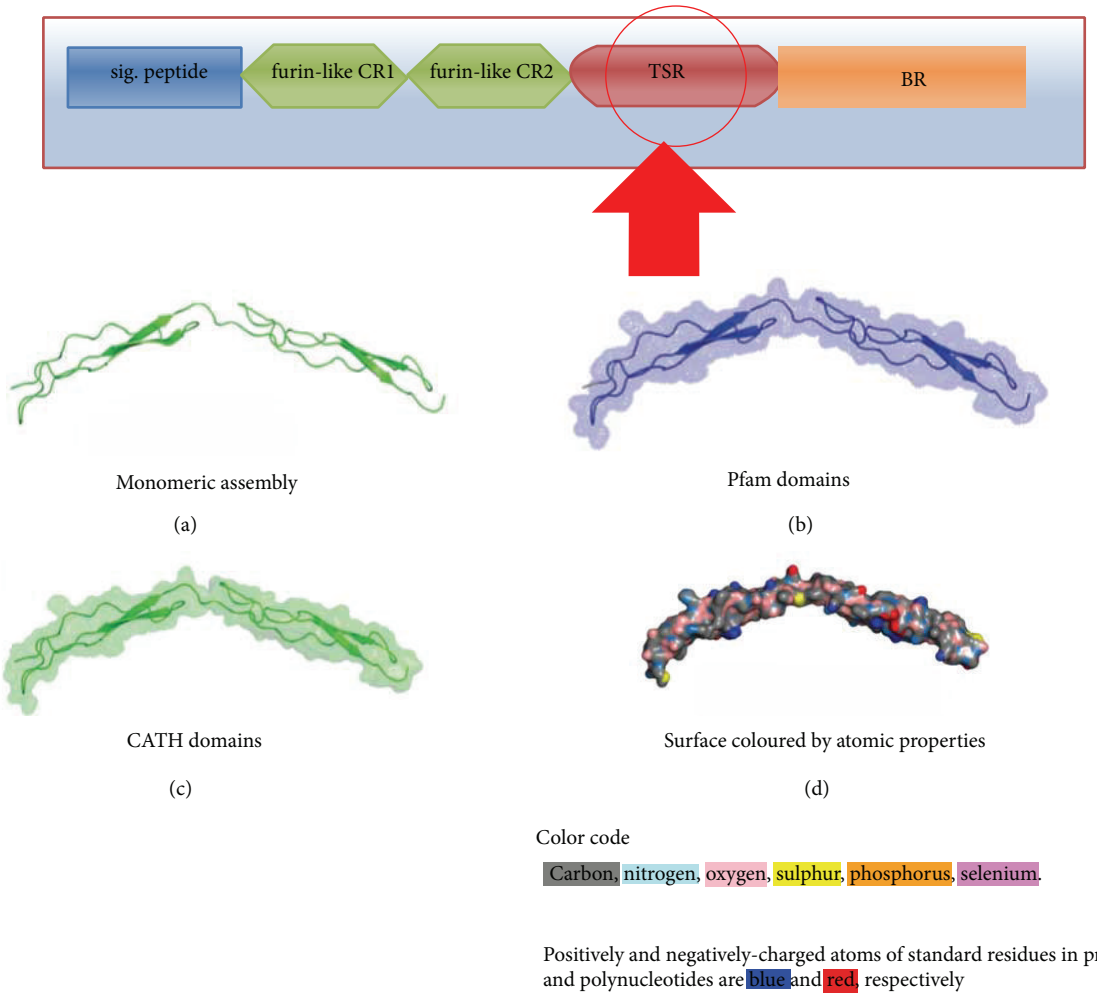


Figure 5: Unique backbone structure of thrombospondin-1 type 1 repeats/domain (TSR). (a) Monomeric assembly structure of TSR, (b) structure of TSR domain generated through Pfam domain database, (c) structure of TSR domain generated through CATH domain database, and (d) surface structure of TSR domain shows the atomic properties (the schematic diagram shows the location of thrombospondin-1 type 1 repeats/domain peptide in the general domain architecture of human (Rspo)s).

index among (Rspo)s and the Rspo2 have the lowest. The AI value of Rspo1 (54.94) was approximately closer to the value of Rspo3 (51.58).

It has been reported that AI value is directly proportional to the structural stability of the protein. The procedure is generally used to calculate the AI of a protein [72, 73], which is as follows:

$$\mathrm{AI} = X_A + aX_V + b\left(X_I + X_L\right), \tag{4}$$

where, X_A, X_V, X_I, and X_L represent the mole percentage of the four residues in a protein which are Ala, Val, Ile, and Leu, respectively. The notation "a" and "b" are coefficients representing the relative volumes of aliphatic side chains and the values are ($a = 2.9 \pm 0.1$ and $b = 3.9 \pm 0.1$), calculated from the volume occupied by the aliphatic amino acids in a protein.

Positively charged residues (PCR) and negatively charged residues (NCR) control several cell properties such as PCR controlled ribosomal velocity [74], NCR controlled K+ channels [75]. These two parameters are helpful to determine the topology of protein [76, 77]. A sum of Arg and Lys are calculated for the presence of the total number of positively charged residues in a protein. Conversely, totality of Asp and Glu are used to calculate the total number of negatively charged residues. Our analysis revealed that, Rspo3 contains the maximum number of positively charged residues while Rspo4 consisted of lowest number. Similarly, Rspo3 consisted of the highest number of negatively charged residues while Rspo4 had the lowest number (Figure 7(d)). The results signify that total numbers of positively charged residues are more than the total number of negatively charged residues for all (Rspo)s.

3.6. Prediction of Globularity in the R-Spondin Family Proteins. From globular domains, several conventional concepts of protein science were initially developed and it challenge by essentially disordered domains [78]. It is frequently analysed

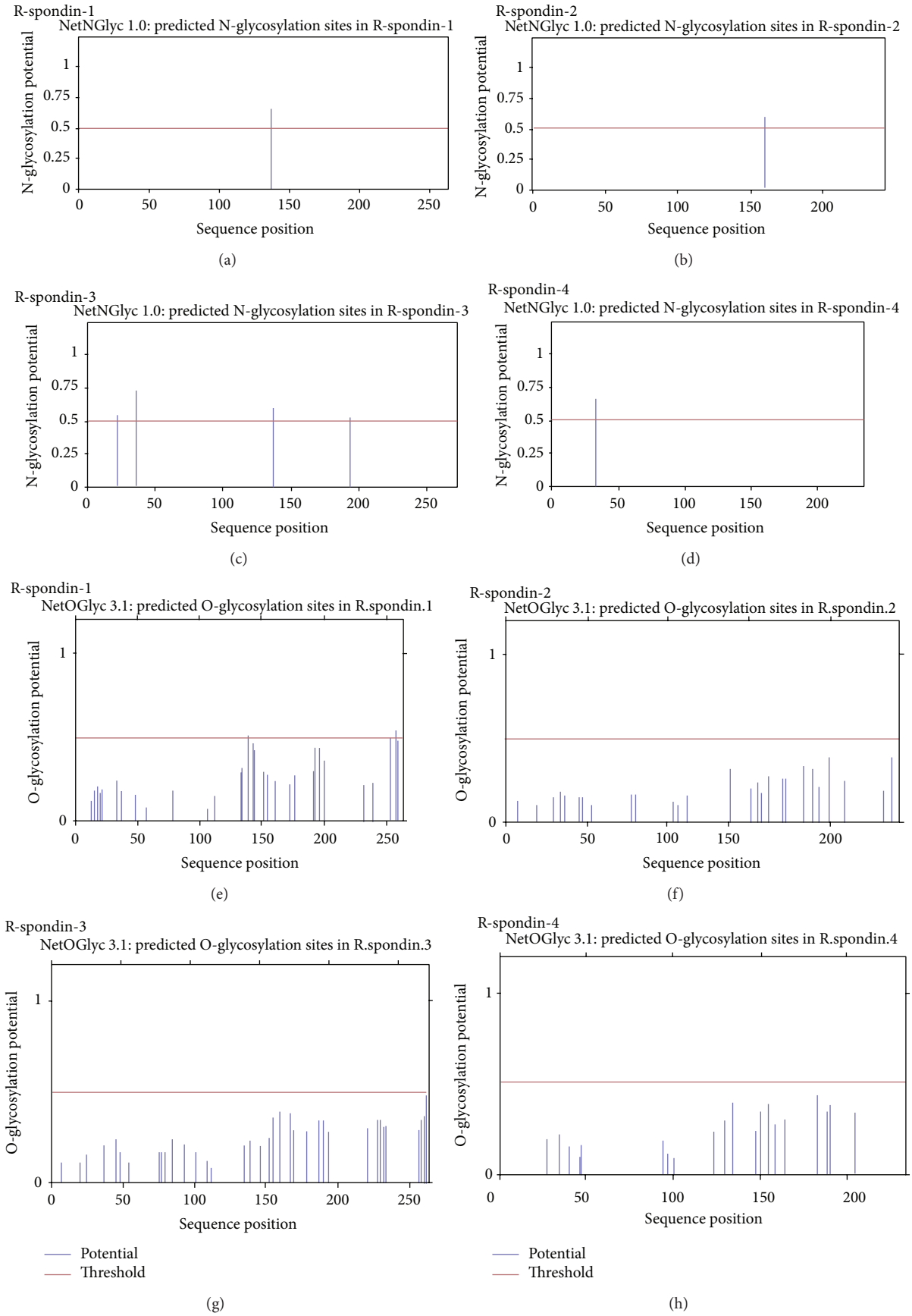
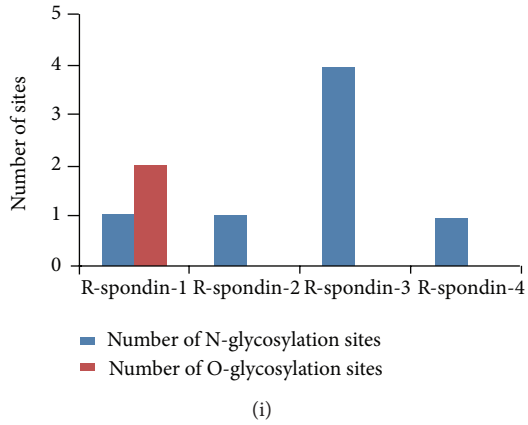


Figure 6: Continued.



(i)

Figure 6: Predicted N-glycosylation and O-glycosylation potentialities and their positions in the different human (Rspo)s. (a) N-glycosylation potentialities of Rspo1, (b) N-glycosylation potentialities of Rspo2, (c) N-glycosylation potentialities of Rspo3, (d) N-glycosylation potentialities of Rspo4, (e) O-glycosylation potentialities of Rspo1, (f) O-glycosylation potentialities of Rspo2, (g) O-glycosylation potentialities of Rspo3, (h) O-glycosylation potentialities of Rspo4, and (i) comparison of predicted N-glycosylation and O-glycosylation sites for four human (Rspo)s.

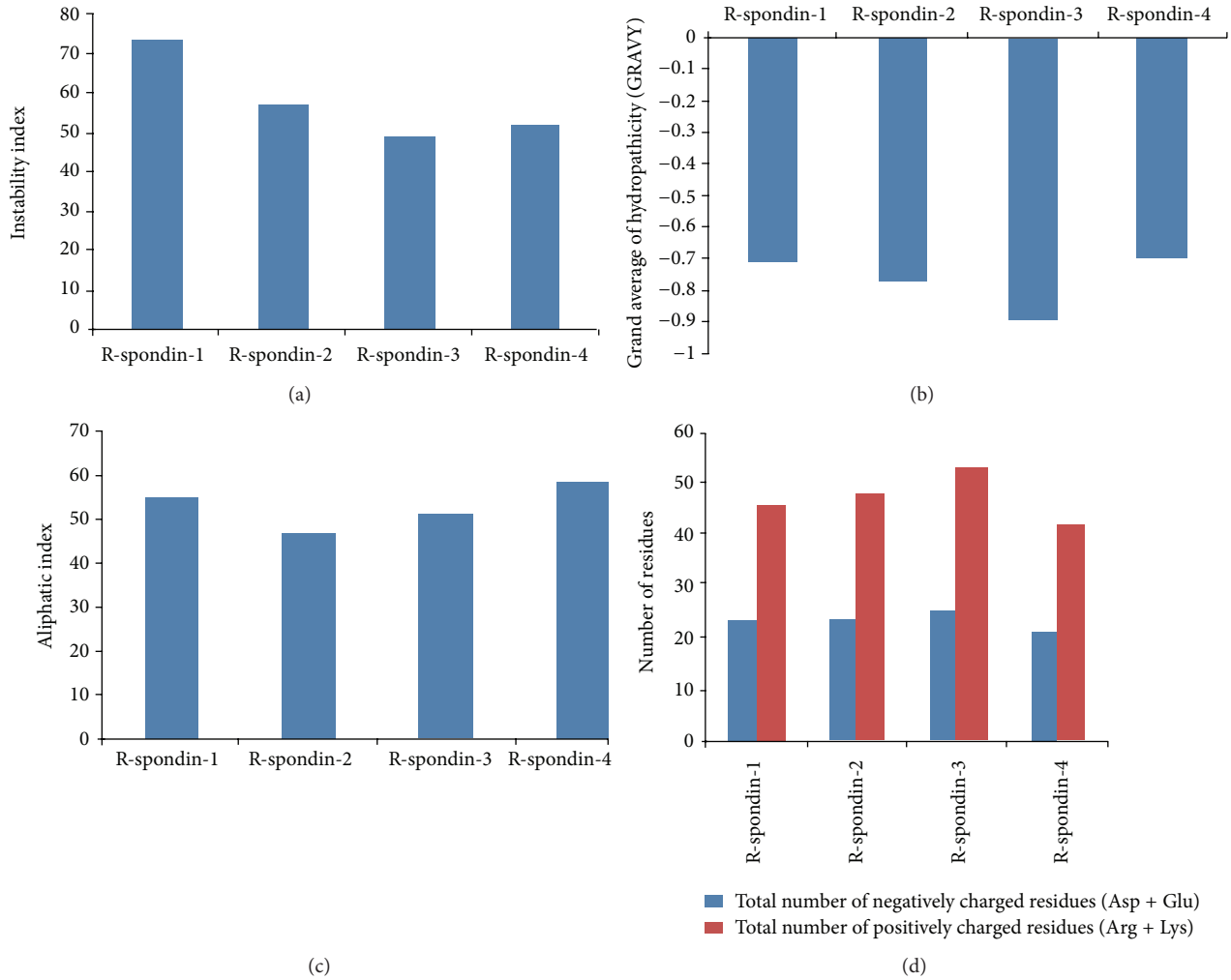


Figure 7: Comparison of biophysical and biochemical properties of four human (Rspo)s. (a) Comparison of instability index, (b) comparison of grand average of hydrophobicity (GRAVY), (c) comparison of aliphatic index, and (d) comparison of total number of positively/negatively charged residues.

to understand thestructure-function relationships, because the structure is having one or numerous catalytic or binding sites on its surface [79]. The globular domains which we analysed are shown in Figure 8. The amino acid sequence alignment in the upper portion of the figure illustrates the differences between the domains. All the proteins were found to contain disordered regions on its surfaces which are as following: Rspo1 (5), Rspo2 (6), Rspo3 (7) and Rspo4 (6). Although globular domain analysis found that Rspo1, Rspo2 and Rspo3 contains globular domain, but no globular domain was observed in Rspo4.

3.7. Multiple Sequences Alignment (MSA) Analysis among R-Spondin Protein Family. The alignment of the (Rspo)s sequences using Clustal Omega is illustrated in supplementary Figure S3. The MUSCLE output was visualised through JalView and is shown in Figure 9(a). As mentioned, 37 small and large aligned divisions were found. We observed best aligned parts between the Rspo4 and Rspo2 sequences, as well as between Rspo4 and Rspo2. We also analysed the highly aligned blocks through Gblocks. The alignment results of Gblocks are shown in Figure 9(b) which shows four highly aligned blocks. From this result, we found highly conserved amino acids such as Leu, Arg, Ser, Gly, Cys, Asn, and Phe.

3.8. Multiple Sequences Alignment (MSA) Analysis of R-Spondin Family Proteins with Other Species. Thereafter, we performed MSA analysis of R-spondin family proteins with other species ($n = 53$). The MSA result is shown in Figure 10. The maximum conservation found was up to 270 sequence and some amino acids such as glycine, cysteine, valine, serine, proline, histidine, leucine and tyrosine were found highly conserved between the sequences.

3.9. Analysis of Molecular Phylogenetics of Human R-Spondin Family Proteins. Phylogram, cladogram and binary tree (equivalent to cladogram) have been depicted (Figures 11(a), 11(b), and 11(c)) and it demonstrates a significant relationship among the proteins of R-spondin family. A molecular phylogenetic analysis of R-spondin members would represent a significant feature of (Rspo)s evolution. In the constructed phylogenetic tree, the distance of branches was illustrated through the likelihood ratio mapping for evolutionary relationships among distinct members of R-spondin family. During the analysis of the tree algorithm, another figure have been described (Figure 11(c)) from the cladogram (Figure 11(b)), that clearly shows the phylogenetic tree rooted with ideal binary numbers (Figure 11(c)). The rooted tree contains two internal nodes and each internal node is further divided into two children nodes, highlighting proteins at their tips. We observed that the altitude of the binary tree was 2 stage. To cross check, other phylogenetic tree called "circular alpha phylogenetic tree", was developed (Figure 11(d)) using MAFFT server. Developed tree resembled the first phylogenetic tree when compared (Figure 11(a)). Both the tree shows, Rspo4 being rooted with Rspo2 and likewise, Rspo3 and Rspo1 have the common point of origin. Again using the four sequences, we developed another phylogenetic tree using Clustal Omega (Figure 11(e)). This tree also showed that Rspo4 and Rspo2 have the common point of origin. We plotted the branch length from the tree in (Figure 11(f)), where Rspo4 showed longest branch length while Rspo3 had the shortest branch length.

3.10. Prediction of Phylogenomics of Human R-Spondin Proteins Using Molecular Phylogenetics to Understand the Framework Topology of Other Related Species. Presently, the phylogenomics, the study of genomes from an evolutionary perspective, is one of the most significant branches to understand the molecular phylogenetics [80, 81]. Phylogenomics provides an understanding about the framework topology of other related species containing orthologous and paralogous genes. The phylogenomics and the framework topology may provide an understanding about the speciation event or duplication event [82, 83]. The phylogenomics (molecular phylogenetics) of human four (Rspo)s with other species have been depicted in Figure 12. Here, phylogenetic tree has been developed using sixty proteins ($n = 60$) and it is an interactive tree with the orthologs and paralogs of the seed sequences. Form Phylogenetic tree, it is very clear that R-spondin family is only distributed among vertebrate species. Our interactive tree shows the origin and evolution of R-spondin among vertebrate family members and it illustrated that non-vertebrate members (Drosophila and C. Elegans) are not having domains similar to that of R-spondins. We also specified the tree legend containing different color codes of the different tree nodes. In front of the figure, the domain and sequence panel have been illustrated showing PFAM motifs. The motifs are represented by different shapes. Inter-domain coding regions have been demonstrated as the standard amino acid colour codes and the gap regions are pointed up as a flat line.

Our tree not only describes the phylogenomics of the R-spondin family but also offers an ideal framework topology based on the biological knowledge of R-spondin family and other related sequences. Our result shows the state-of-the-art evolutionary patterns of R-spondin family and the related gene families.

3.11. Understanding the Protein-Protein Interaction Network of R-Spondin Protein Family. Complete knowledge about the protein-protein interaction networks offers direct and indirect interactions between proteins in a cell, helping us to depict a comprehensive description of cellular mechanisms and functions [84, 85].

The protein-protein network of R-spondin protein family is illustrated in Figure 13. The input file for the development of protein-protein network of (Rspo)s has been is noted in the Supplementary Figure S4. Four different developed protein-protein interaction networks exemplified the different interactive proteins with the four members of R-spondin family. Rspo1 shows interaction network between FURIN, DKK1, ZNRF3, LRP6, FZR8, SRY, FOXL2, SOX9, MYF5, and STRA8 (Figure 13(a)). The interaction network is more condensed among DKK1, LRP6 and FZR8. Rspo2 shows interaction network between SP8, KRT71, FGF5,

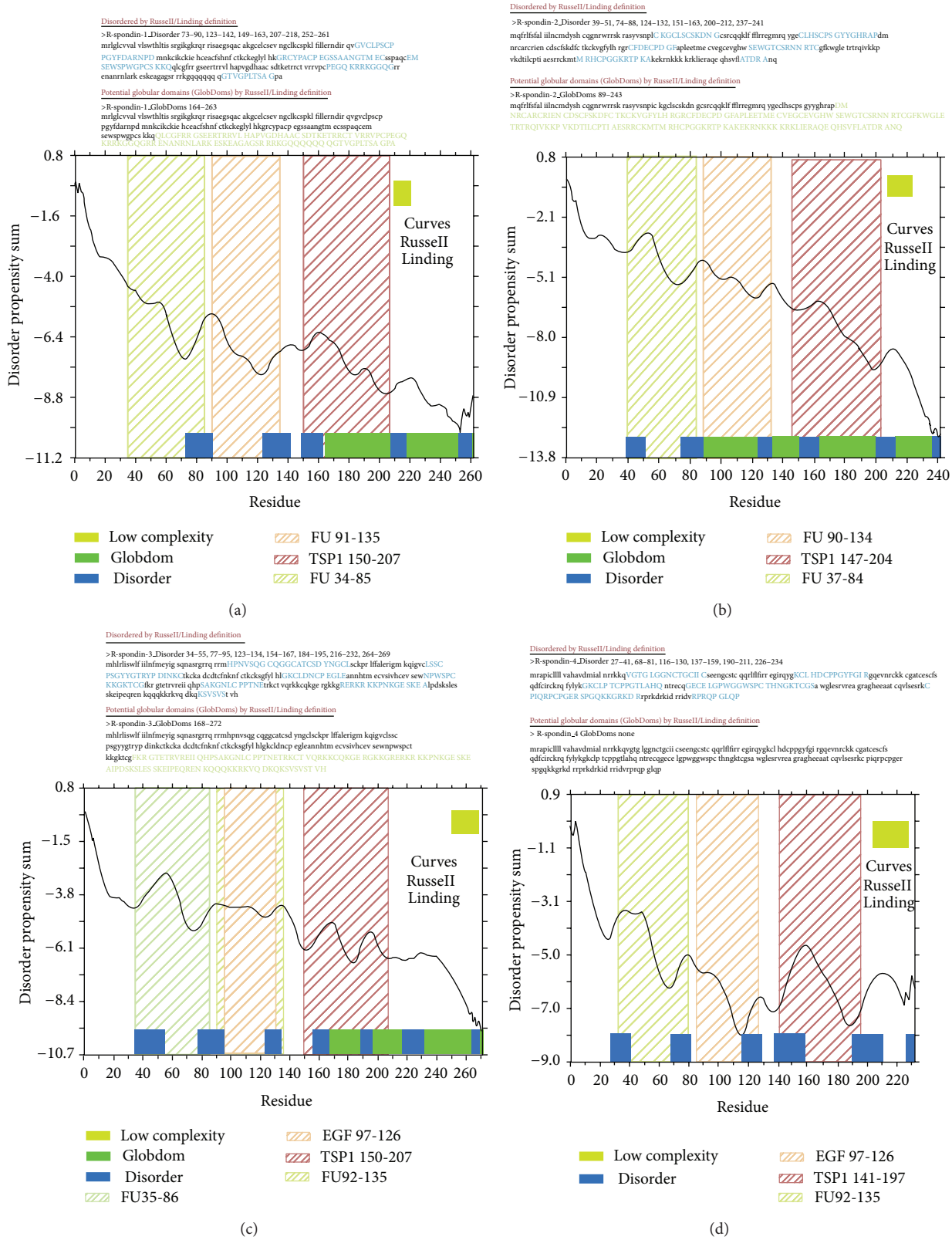


FIGURE 8: Globular domain gain/loss as a function of the variation between the four human (Rspo)s. The disorder propensity of the protein stretch was calculated using GlobPlot analyses to identify the disorder region (blue). The upper portion in the figure illustrates the differences between the amino acid sequence alignments among the domains. The tool uses a simple peak-finder algorithm to select the putative globular and disorder segments. (a) Rspo1, (b) Rspo2, (c) Rspo3, and (d) Rspo4.

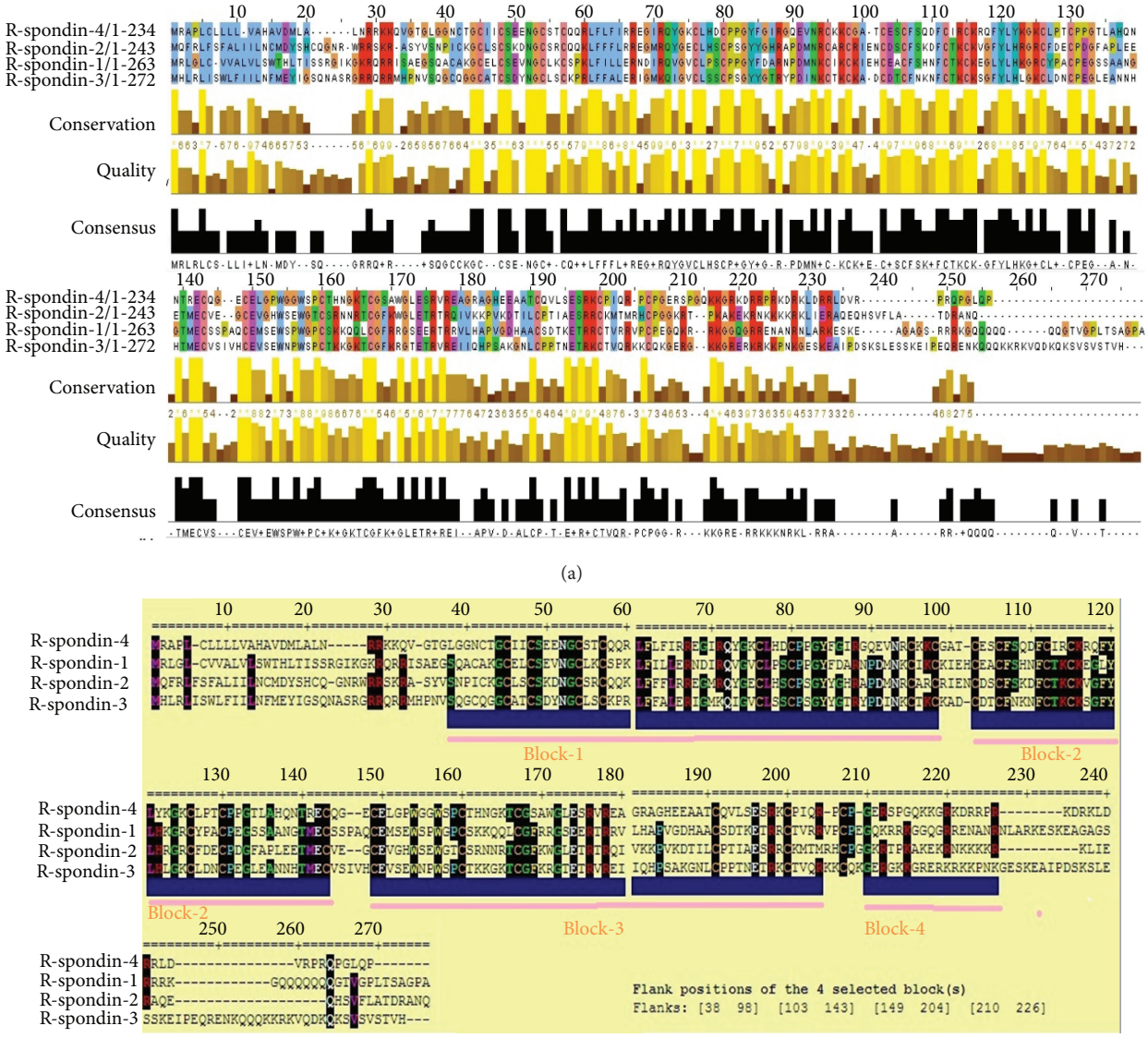


FIGURE 9: Multiple sequence alignment (MSA) of the different human (Rspo)s. (a) MSA output visualised through JalView and (b) the Gblocks results of human (Rspo)s show blocks from the alignments. The results show highly aligned four blocks.

GORAB, PTPRK, KIAA1804, PDIK1L, GUCY2F, MYLK2, and WNT3A (Figure 13(b)). In this network, no condensed part was found. Rspo3 shows interaction network between FZD8, SDC4, MYF5, FURIN, FAM70A, WNT1, LRP6, KREMEN2, DVL1, and CTNNB1 (Figure 13(c)). The interaction network is more condensed among the proteins which are located in the upper portion of the network such as FZD8, SDC4, WNT1, LRP6, KREMEN2, DVL1, and CTNNB1. Rspo4 shows network between only one protein that is, FURIN (Figure 13(d)) and it is the shortest network among R-spondin protein family.

4. Discussion

R-spondin protein family is an immensely important protein family, which acts as a key regulator factor during vertebrate development and several signalling pathways, especially as agonists for the canonical Wnt/β-catenin signalling pathway [17]. Association with different diseases has been found with R-spondin family proteins. (Rspo)s are associated with various developmental stages as an essential regulator. For example, Rspo1 has been found to be associated with sex determination and skin differentiation [22]; Rspo2 is a crucial protein for development of limbs; lungs and hair follicles [11, 27, 86]; Rspo3 is essential for placental development [10] and Rspo4 is a significant protein for nail deployment [17]. (Rspo)s have therapeutic potential for various diseases such as skeletal diseases [87], inflammatory bowel disease and chemotherapy-induced mucositis [23], cancer [21], and diabetes [88]. Therefore, basic understanding about the biophysical, biochemical properties of (Rspo)s may provide more understanding about their functional mechanism associated

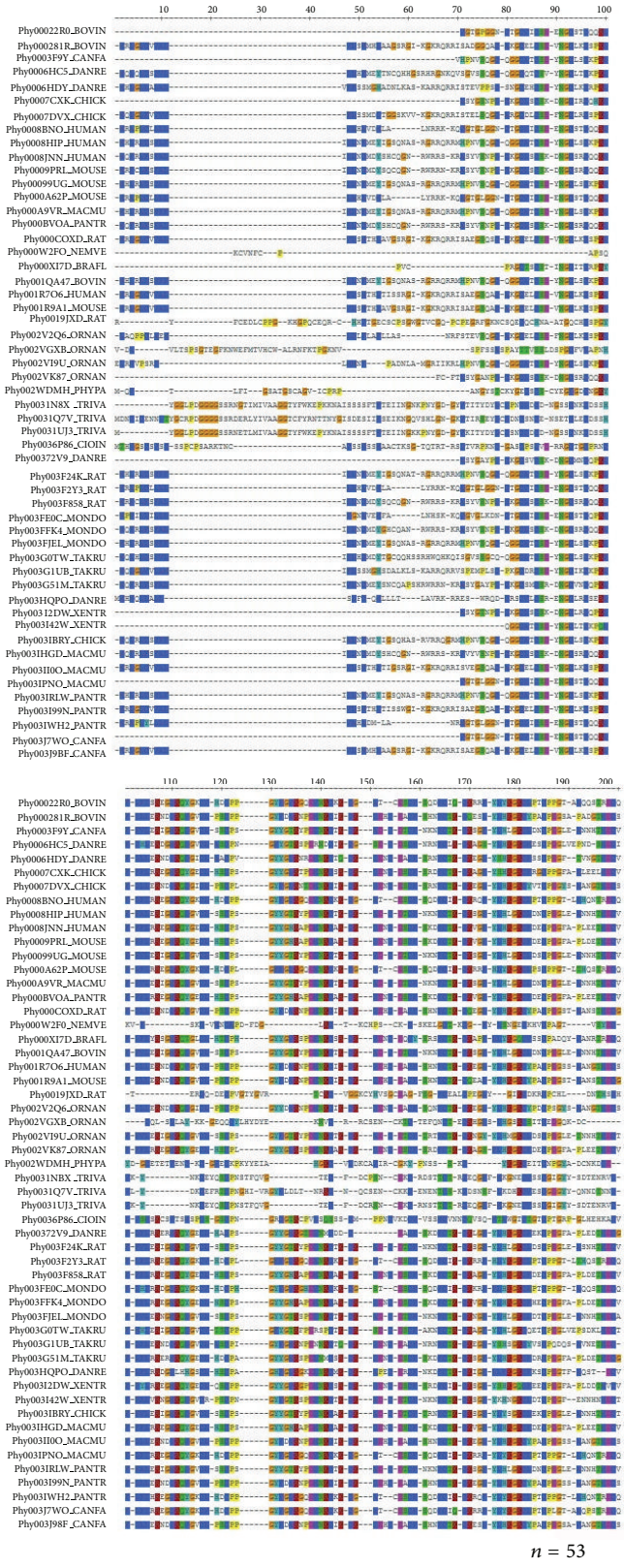


Figure 10: Multiple sequence alignment (MSA) of the different human (Rspo)s with other species ($n = 53$) which are having sequence similarity.

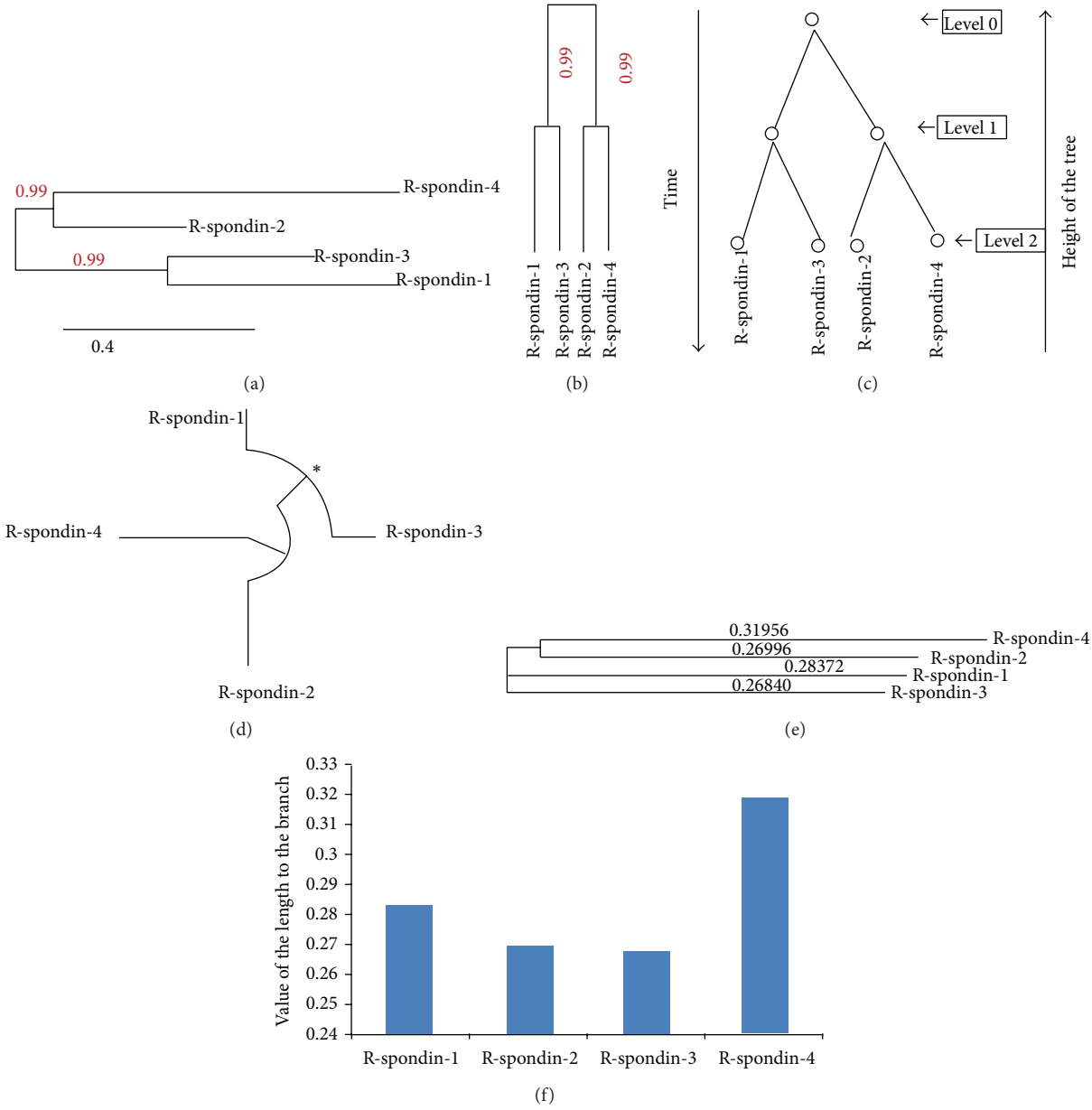


Figure 11: Phylogenetic analysis using different human (Rspo)s which shows the relationship between the family members. (a) A phylogenetic tree showing the evolution in the relationships between human (Rspo)s. Bootstrap values are pointed out at nodes. R-spondins protein family members names at the clade, (b) cladogram of protein sequences of the (Rspo)s members, (c) binary tree representation equivalent to cladogram, (d) phylogenetic tree (alpha circular type) reconstructed using the maximum likelihood method. The ∗ symbol adjacent to the node indicates the origin point of the proteins, (e) phylogenetic tree developed through clustal-omega server where the value of the branch length is mentioned immediately leading to the node, and (f) comparison between the value of the branch lengths.

with the diseases and the developmental processes. In this work, to decipher more about the biophysical, biochemical and evolutionary relationship of the R-spondin family, we carried out biophysical, biochemical, and evolutionary based computational mapping of human (Rspo)s.

In general, proteins have a small signal peptide sequence which helps them to enter into the secretory pathway. The N-terminal signal peptide sequence direct proteins to the membrane of the endoplasmic reticulum (ER) and initiate translocation into the ER lumen [89]. From our database analysis, we identified sequences similarity for signal peptides within (Rspo)s (20 to 21 sequence) (Figure 2) which corroborated to the finding of Kim et al. [2]. In addition to known findings, herein we analyzed signal peptides of human (Rspo)s in the more detail way along with their C-score, S-score and Y-score (Figure 3). Computational methods for estimating N-terminal signal peptides have been detected previously. But, our used server is an advanced tool which uses HMM-based better neural network scheme [39]. Using this tool, we have illustrated the predicted cleavage site value

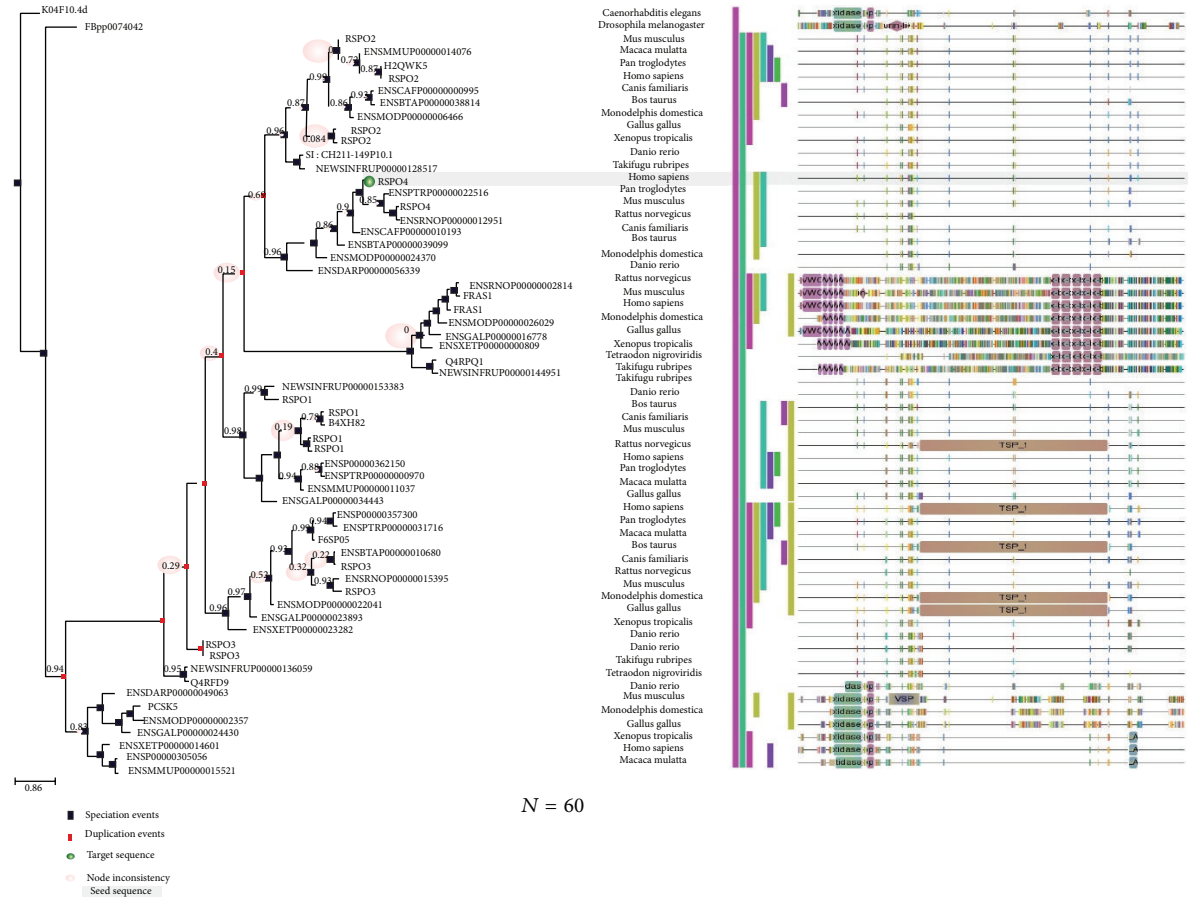


FIGURE 12: Phylogenomics of human four (Rspo)s and similar proteins from other species (n = 60) which are having sequence similarity. In front of the figure, the domain and sequence panel have been depicted which uses PFAM motifs, and the motifs are represented by different shapes.

(C-score) in the signal peptide of human (Rspo)s where possible two signals are noted in a single signal peptidase cleavage site (Rspo2 and Rspo3) (Figure 3). Hiss and Schneider [89] revealed that long signal peptides mingle two or more signals of signal peptidase cleavage site.

From the amino acids distributed pattern of human (Rspo)s especially from the exposed distribution analysis at a time (Figure 4(e)), we observed the similarity of the amino acids distributed pattern is more or less same. However, Rspo1, Rspo2 and Rspo4 showed more similarity in the distribution pattern. At the same time, our analysis revealed identical amino acid composition pattern in the Rspo1, Rspo2 and Rspo4 (Figures 4(a), 4(b), 4(c) and 4(d)). Recently, it was reported that there is an association between amino acid composition and distribution with mutation. Researchers have shown the correlation between the amino acids distribution pattern; missense mutations and genetic disorders [90]. Conversely, amino acid composition was linked with the deleterious impact of mutations [91]. Therefore, amino acid composition and distributed pattern of human (Rspo)s may help to the future researcher to understand the impact and association with genetic disorders. Further analysis with Cys residues revealed that all these four (Rspo)s are Cys rich protein. Also, the Cys architecture and the disulphide bond pattern show a common architecture and may be necessary for the stability of these proteins (Figure 4(f)). Recent *in vitro* study with mass spectrometry documented the pattern of disulfide bonds between the 15 available Cys residues present in furin domains in (Rspo)s [14]. However, they found five free cysteine residues in Rspo2.

Our analysis found some glycosylation sites for (Rspo)s which may be necessary for their functionality and signalling process (Figure 6). Previously, Kamata et al. [1] has indicated the N-linked glycosylation sites for (Rspo)s. Our previous similar kind of computational analysis shows that the N-glycosylation sites and O-glycosylation sites are vital for the functionality of the proteins in the insulin signalling pathway proteins such as IRS and GLUT4 [67, 92]. However, identified O- and N-glycosylation sites by our analysis with (Rspo)s needs to be confirmed with molecular and biochemical experiments.

Previously, Kim et al. [2] and Nam et al. [4] performed multiple sequence analysis with four (Rspo)s. We also performed MSA among four (Rspo)s as well as with several other species proteins using different computational server (Figure 10). Compared to the previous analysis, our MSA

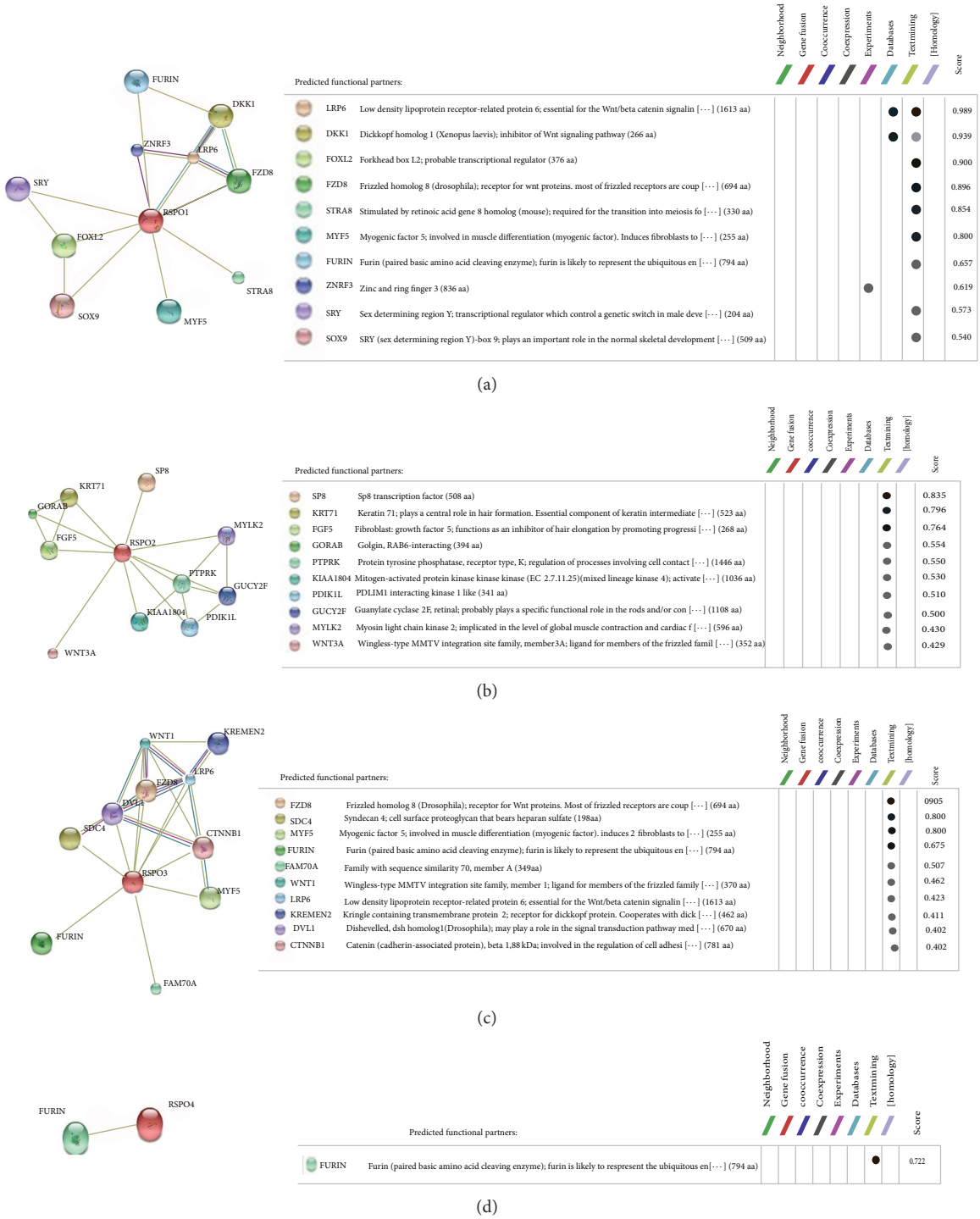


FIGURE 13: Protein-protein interaction network of R-spondin family proteins using STRING server. (a) Rspo1, (b) Rspo2, (c) Rspo3, and (d) Rspo4.

investigation provides a very clear picture about the aligned and conserved residues with different colour codes visualised through JalView. We then analysed through Gblocks server to understand the conserve blocks within the R-spondin family. Our data showed four highly conserved blocks within depicted Gblocks (Figure 9). Furthermore, another MSA analysis of (Rspo)s with other species ($n = 53$) was performed to understand more conserved residues among different species where we found several small conserved blocks and residues such as glycine, cysteine, valine, serine, proline, histidine, leucine and tyrosine (Figure 10).

The evolutionary history of R-spondin family and the phylogenetic relationships prototype can be investigated through the molecular approach involving amino acid sequencing. Utilizing similar approach, we developed phylogenetic relationships among the members of the R-spondin

family, and we found that Rspo4 and Rspo2 were siblings in 99% bootstrap replications and likewise, Rspo3 and Rspo1 were siblings in 99% bootstrap replications (Figure 11). Previously, de lau et al. [17] and our group also [4] analysed phylogenetic relationships. Here, we performed more advanced two types of phylogenetic analyses: (i) phylogenetic relationships pattern of R-spondin family (Figure 11) and (ii) phylogenetic relationships using R-spondin family using sixty species ($n = 60$) (Figure 12). Second one is the interactive tree with the orthologs and paralogs of the seed sequences which describe the phylogenomics of the R-spondin family and also determines evolutionary relationship of different species (Figure 12). This analysis directs the study towards next generation phylogenomics [93] which may be robust and alignment-free.

From our protein-protein interaction network analysis, we noted an interaction among the Rspo1 with the LRP6 and FZR8 receptor confirming them as candidate protein for Wnt signaling pathway (Figure 13(a)). Hao et al. [94] reported that LRP6 and FZD receptors are present on the membrane and these receptors permit the Wnt ligands to generate much stronger signals. The network of ZNRF3 with Rspo1 confirms that ZNRF3 is associated with Wnt receptor yield in an R-spondin sensitive manner [94]. The network also shows Rspo1 interaction with DKK1 (an antagonist of Wnt signaling). Binnerts et al. [9] reported that Rspo1 binds to the Kremen family of transmembrane proteins and it negatively regulates the LRP6 receptor through the DKK1-associated endocytosis. Due to the controlling property of individual's sex phenotype, Rspo1 networks with SRY and SOX9 protein [95]. The network of Rspo2 with FGF shows that damage Wnt signal directs to defective expression of the important apical ectodermal ridge maintenance factors, FGF4 and FGF8, which is related with the lung and limb development (Figure 13(b)). Similar to Rspo1, we observed a strong association between Rspo3 and the LRP6/FZR8 receptor as well as DVL for Wnt signaling pathway (Figure 13(c)). Rspo4 shows an interaction between FURIN proteins. It has been known that FURIN like domain is necessary for the activity of Rspo4. Blaydon et al. [96] demonstrated that mutations interrupting furin-like domains in Rspo4 may affect its signaling activity. Recent studies showed that (Rspo)s are the ligands for the leucine-rich repeat containing G protein-coupled receptor 4/5/6 (LGR4/5/6) receptors [15–18]. However, in our analysis we have not found any network between the (Rspo)s with LGR4/5/6. This might be due to the lack of updated data in server database (STRING database) containing information about the LGR4/5/6.

In summary, through computational analysis, we performed biophysical, biochemical, and evolutionary topology of human R-spondin family proteins. In this work, we have applied innovative and rapid approach to study the structural based biophysical, biochemical, and evolutionary relationship among (Rspo)s. The difficult and time-consuming nature of the experimental analysis led us to attempt to develop a cost-effective computational research of biophysical, biochemical and evolutionary topology of human R-spondin family. In this study, we have tried to highlight the possible potent sites for O- and N-glycosylation, distribution and conservation of amino acids and to predict phylogenetic and protein-protein interaction among (Rspo)s with the available data base. However, experimental biochemical and functional studies are required to further establish these finding. Our attempt to decipher the biophysical and biochemical properties of (Rspo)s may provide useful platform and a starting point for scientists to unfold significant physiological and therapeutic properties of R-spondin protein family in various disease models.

Conflict of Interests

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

Authors' Contribution

Ashish Ranjan Sharma, Chiranjib Chakraborty, and Sang-Soo Lee contributed equally to this work.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A1A4A03009388 and 2011-001-4792) and by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI12C1265). The authors also take this opportunity to thank the management of VIT and Galgotias University for providing the facilities and encouragement to carry out this work.

References

[1] T. Kamata, K.-I. Katsube, M. Michikawa, M. Yamada, S. Takada, and H. Mizusawa, "R-spondin, a novel gene with thrombospondin type 1 domain, was expressed in the dorsal neural tube and affected in Wnts mutants," *Biochimica et Biophysica Acta—Gene Structure and Expression*, vol. 1676, no. 1, pp. 51–62, 2004.

[2] K.-A. Kim, J. Zhao, S. Andarmani et al., "R-spondin proteins: a novel link to β-catenin activation," *Cell Cycle*, vol. 5, no. 1, pp. 23–26, 2006.

[3] K.-A. Kim, M. Wagle, K. Tran et al., "R-Spondin family members regulate the Wnt pathway by a common mechanism," *Molecular Biology of the Cell*, vol. 19, no. 6, pp. 2588–2596, 2008.

[4] J.-S. Nam, T. J. Turcotte, P. F. Smith, S. Choi, and K. Y. Jeong, "Mouse cristin/R-spondin family proteins are novel ligands for the frizzled 8 and LRP6 receptors and activate β-catenin-dependent gene expression," *The Journal of Biological Chemistry*, vol. 281, no. 19, pp. 13247–13257, 2006.

[5] Q. Wei, C. Yokota, M. V. Semenov, B. Doble, J. Woodgett, and X. He, "R-spondin1 is a high affinity ligand for LRP6 and induces LRP6 phosphorylation and β-catenin signaling," *The Journal of Biological Chemistry*, vol. 282, no. 21, pp. 15903–15911, 2007.

[6] O. Kazanskaya, A. Glinka, I. del Barco Barrantes, P. Stannek, C. Niehrs, and W. Wu, "R-Spondin2 is a secreted activator of

Wnt/β-catenin signaling and is required for Xenopus myogenesis," *Developmental Cell*, vol. 7, no. 4, pp. 525–534, 2004.

[7] J.-S. Nam, T. J. Turcotte, and J. K. Yoon, "Dynamic expression of R-spondin family genes in mouse development," *Gene Expression Patterns*, vol. 7, no. 3, pp. 306–312, 2007.

[8] B. A. Parr, M. J. Shea, G. Vassileva, and A. P. McMahon, "Mouse Wnt genes exhibit discrete domains of expression in the early embryonic CNS and limb buds," *Development*, vol. 119, no. 1, pp. 247–261, 1993.

[9] M. E. Binnerts, K.-A. Kim, J. M. Bright et al., "R-Spondin1 regulates Wnt signaling by inhibiting internalization of LRP6," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 37, pp. 14700–14705, 2007.

[10] M. Aoki, M. Mieda, T. Ikeda, Y. Hamada, H. Nakamura, and H. Okamoto, "R-spondin3 is required for mouse placental development," *Developmental Biology*, vol. 301, no. 1, pp. 218–226, 2007.

[11] S. M. Bell, C. M. Schreiner, S. E. Wert, M. L. Mucenski, W. J. Scott, and J. A. Whitsett, "R-spondin 2 is required for normal laryngeal-tracheal, lung and limb morphogenesis," *Development*, vol. 135, no. 6, pp. 1049–1058, 2008.

[12] B. Ohkawara, A. Glinka, and C. Niehrs, "Rspo3 binds syndecan 4 and induces Wnt/PCP signaling via clathrin-mediated endocytosis to promote morphogenesis," *Developmental Cell*, vol. 20, no. 3, pp. 303–314, 2011.

[13] K.-A. Kim, M. Kakitani, J. Zhao et al., "Mitogenic influence of human R-spondin1 on the intestinal epithelium," *Science*, vol. 309, no. 5738, pp. 1256–1259, 2005.

[14] S.-J. Li, T.-Y. Yen, Y. Endo et al., "Loss-of-function point mutations and two-furin domain derivatives provide insights about R-spondin2 structure and function," *Cellular Signalling*, vol. 21, no. 6, pp. 916–925, 2009.

[15] K. S. Carmon, X. Gong, Q. Lin, A. Thomas, and Q. Liu, "R-spondins function as ligands of the orphan receptors LGR4 and LGR5 to regulate Wnt/β-catenin signaling," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 28, pp. 11452–11457, 2011.

[16] K. S. Carmon, Q. Lin, X. Gong, A. Thomas, and Q. Liu, "LGR5 interacts and cointernalizes with Wnt receptors to modulate Wnt/beta-catenin signaling," *Molecular and Cellular Biology*, vol. 32, no. 11, pp. 2054–2064, 2012.

[17] W. B. de Lau, B. Snel, and H. C. Clevers, "The R-spondin protein family," *Genome Biology*, vol. 13, no. 3, article 242, 2012.

[18] A. Glinka, C. Dolde, N. Kirsch et al., "LGR4 and LGR5 are R-spondin receptors mediating Wnt/β-catenin and Wnt/PCP signalling," *EMBO Reports*, vol. 12, no. 10, pp. 1055–1061, 2011.

[19] T. Grigoryan, P. Wend, A. Klaus, and W. Birchmeier, "Deciphering the function of canonical Wnt signals in development and disease: conditional loss- and gain-of-function mutations of β-catenin in mice," *Genes and Development*, vol. 22, no. 17, pp. 2308–2341, 2008.

[20] R. Nusse, "Wnt signaling in disease and in development," *Cell Research*, vol. 15, no. 1, pp. 28–32, 2005.

[21] Y. R. Jin and J. K. Yoon, "The R-spondin family of proteins: emerging regulators of WNT signaling," *The International Journal of Biochemistry and Cell Biology*, vol. 44, no. 12, pp. 2278–2287, 2012.

[22] P. Parma, O. Radi, V. Vidal et al., "R-spondin1 is essential in sex determination, skin differentiation and malignancy," *Nature Genetics*, vol. 38, no. 11, pp. 1304–1309, 2006.

[23] J. Zhao, J. de Vera, S. Narushima et al., "R-spondin1, a novel intestinotrophic mitogen, ameliorates experimental colitis in mice," *Gastroenterology*, vol. 132, no. 4, pp. 1331–1343, 2007.

[24] G. Krönke, S. Uderhardt, K.-A. Kim et al., "R-spondin 1 protects against inflammatory bone damage during murine arthritis by modulating the Wnt pathway," *Arthritis & Rheumatism*, vol. 62, no. 8, pp. 2303–2312, 2010.

[25] W. Lu, K.-A. Kim, J. Liu et al., "R-spondin1 synergizes with Wnt3A in inducing osteoblast differentiation and osteoprotegerin expression," *FEBS Letters*, vol. 582, no. 5, pp. 643–650, 2008.

[26] A. R. Sharma, B. S. Choi, J. M. Park et al., "Rspo 1 promotes osteoblast differentiation via Wnt signaling pathway," *Indian Journal of Biochemistry and Biophysics*, vol. 50, no. 1, pp. 19–25, 2013.

[27] J.-S. Nam, E. Park, T. J. Turcotte et al., "Mouse R-spondin2 is required for apical ectodermal ridge maintenance in the hindlimb," *Developmental Biology*, vol. 311, no. 1, pp. 124–135, 2007.

[28] W. Yamada, K. Nagao, K. Horikoshi et al., "Craniofacial malformation in R-spondin2 knockout mice," *Biochemical and Biophysical Research Communications*, vol. 381, no. 3, pp. 453–458, 2009.

[29] E. Cadieu, M. W. Neff, P. Quignon et al., "Coat variation in the domestic dog is governed by variants in three genes," *Science*, vol. 326, no. 5949, pp. 150–153, 2009.

[30] S. Seshagiri, E. W. Stawiski, S. Durinck et al., "Recurrent R-spondin fusions in colon cancer," *Nature*, vol. 488, no. 7413, pp. 660–664, 2012.

[31] S. Neufeld, J. M. Rosin, A. Ambasta et al., "A conditional allele of Rspo3 reveals redundant function of R-spondins during mouse limb development," *Genesis*, vol. 50, no. 10, pp. 741–749, 2012.

[32] M. Klauzinska, B. Baljinnyam, A. Raafat et al., "Rspo2/Int7 regulates invasiveness and tumorigenic properties of mammary epithelial cells," *Journal of Cellular Physiology*, vol. 227, no. 5, pp. 1960–1971, 2012.

[33] W. Lowther, K. Wiley, G. H. Smith, and R. Callahan, "A new common integration site, Int7, for the mouse mammary tumor virus in mouse mammary tumors identifies a gene whose product has furin-like and thrombospondin-like sequences," *Journal of Virology*, vol. 79, no. 15, pp. 10093–10096, 2005.

[34] Y. Ishii, M. Wajid, H. Bazzi et al., "Mutations in R-spondin 4 (RSPO4) underlie inherited anonychia," *Journal of Investigative Dermatology*, vol. 128, no. 4, pp. 867–870, 2008.

[35] E. W. Sayers, T. Barrett, D. A. Benson et al., "Database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 39, no. 1, pp. D38–D51, 2011.

[36] UniProt Consortium, "Update on activities at the Universal Protein Resource (UniProt) in 2013," *Nucleic Acids Research*, vol. 41, pp. D43–D47, 2013.

[37] UniProt Consortium, "The universal protein resource (UniProt) in 2010," *Nucleic Acids Research*, vol. 38, pp. D142–D148, 2010.

[38] UniProt Consortium, "Ongoing and future developments at the Universal Protein Resource," *Nucleic Acids Research*, vol. 39, pp. D214–D219, 2011.

[39] T. N. Petersen, S. Brunak, G. von Heijne, and H. Nielsen, "SignalP 4.0: discriminating signal peptides from transmembrane regions," *Nature Methods*, vol. 8, no. 10, pp. 785–786, 2011.

[40] N. J. Anthis and G. M. Clore, "Sequence-specific determination of protein and peptide concentrations by absorbance at 205 nm," *Protein Science*, vol. 22, no. 6, pp. 851–858, 2013.

[41] V. Brendel, P. Bucher, I. R. Nourbakhsh, B. E. Blaisdell, and S. Karlin, "Methods and algorithms for statistical analysis of protein sequences," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 6, pp. 2002–2006, 1992.

[42] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi, "SCRATCH: a protein structure and structural feature prediction server," *Nucleic Acids Research*, vol. 33, no. 2, pp. W72–W76, 2005.

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

[44] R. J. Read, P. D. Adams, W. B. Arendall III et al., "A new generation of crystallographic validation tools for the protein data bank," *Structure*, vol. 19, no. 10, pp. 1395–1412, 2011.

[45] S. Hunter, P. Jones, A. Mitchell et al., "InterPro in 2011: new developments in the family and domain prediction database," *Nucleic Acids Research*, vol. 40, no. 1, pp. D306–D312, 2012.

[46] S. E. Hamby and J. D. Hirst, "Prediction of glycosylation sites using random forests," *BMC Bioinformatics*, vol. 9, article 500, 2008.

[47] K. Julenius, A. Mølgaard, R. Gupta, and S. Brunak, "Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites," *Glycobiology*, vol. 15, no. 2, pp. 153–164, 2005.

[48] J. E. Hansen, O. Lund, N. Tolstrup, A. A. Gooley, K. L. Williams, and S. Brunak, "NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility," *Glycoconjugate Journal*, vol. 15, no. 2, pp. 115–130, 1998.

[49] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel, and A. Bairoch, "ExPASy: the proteomics server for in-depth protein knowledge and analysis," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3784–3788, 2003.

[50] R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson, "GlobPlot: exploring protein sequences for globularity and disorder," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3701–3708, 2003.

[51] F. Sievers, A. Wilm, D. Dineen et al., "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Molecular Systems Biology*, vol. 7, article 539, 2011.

[52] P. Hogeweg and B. Hesper, "The alignment of sets of sequences and the construction of phyletic trees: an integrated method," *Journal of Molecular Evolution*, vol. 20, no. 2, pp. 175–186, 1984.

[53] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[54] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[55] C. Chakraborty, S. S. Roy, M. J. Hsu, and G. Agoramoorthy, "Can computational biology improve the phylogenetic analysis of insulin?" *Computer Methods and Programs in Biomedicine*, vol. 108, no. 2, pp. 860–872, 2012.

[56] J. Huerta-Cepas, S. Capella-Gutierrez, L. P. Pryszcz et al., "PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions," *Nucleic Acids Research*, vol. 39, no. 1, pp. D556–D560, 2011.

[57] J. Huerta-Cepas, A. Bueno, J. Dopazo, and T. Gabaldón, "PhylomeDB: a database for genome-wide collections of gene phylogenies," *Nucleic Acids Research*, vol. 36, no. 1, pp. D491–D496, 2008.

[58] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[59] A. Dereeper, V. Guignon, G. Blanc et al., "Phylogeny.fr: robust phylogenetic analysis for the non-specialist," *Nucleic Acids Research*, vol. 36, pp. W465–W469, 2008.

[60] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[61] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[62] L. J. Jensen, M. Kuhn, M. Stark et al., "STRING —a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Research*, vol. 37, no. 1, pp. D412–D416, 2009.

[63] D. Szklarczyk, A. Franceschini, M. Kuhn et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, no. 1, pp. D561–D568, 2011.

[64] A. Meiniel, R. Meiniel, N. Gonçalves-Mendes, I. Creveaux, R. Didier, and B. Dastugue, "The thrombospondin type 1 repeat (TSR) and neuronal differentiation: roles of SCO- spondin oligopeptides on neuronal cell types and cell lines," *International Review of Cytology*, vol. 230, pp. 1–39, 2003.

[65] T. L. Rodgers, P. D. Townsend, D. Burnell et al., "Modulation of global low-frequency motions underlies allosteric regulation: demonstration in CRP/FNR family transcription factors," *PLoS Biology*, vol. 11, no. 9, Article ID e1001651, 2013.

[66] C. G. P. Doss, C. Chakraborty, B. Rajith, and N. Nagasundaram, "In silico discrimination of nsSNPs in hTERT gene by means of local DNA sequence context and regularity," *Journal of Molecular Modeling*, vol. 19, no. 9, pp. 3517–3527, 2013.

[67] C. Chakraborty, G. Agoramoorthy, and M. J. Hsu, "Exploring the evolutionary relationship of insulin receptor substrate family using computational biology," *PLoS ONE*, vol. 6, no. 2, Article ID e16580, 2011.

[68] T. Huang, X.-H. Shi, P. Wang et al., "Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks," *PLoS ONE*, vol. 5, no. 6, Article ID e10972, 2010.

[69] H. Jubb, A. P. Higueruelo, A. Winter, and T. L. Blundell, "Structural biology and drug discovery for protein-protein interactions," *Trends in Pharmacological Sciences*, vol. 33, no. 5, pp. 241–248, 2012.

[70] K. Guruprasad, B. V. B. Reddy, and M. W. Pandit, "Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence," *Protein Engineering*, vol. 4, no. 2, pp. 155–161, 1990.

[71] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–132, 1982.

[72] A. Ikai, "Thermostability and aliphatic index of globular proteins," *The Journal of Biochemistry*, vol. 88, no. 6, pp. 1895–1898, 1980.

[73] P. Argos, M. G. Rossmann, U. M. Grau, H. Zuber, G. Frank, and J. D. Tratschin, "Thermal stability and protein structure," *Biochemistry*, vol. 18, no. 25, pp. 5698–5703, 1979.

[74] C. A. Charneski and L. D. Hurst, "Positively charged residues are the major determinants of ribosomal velocity," *PLoS Biology*, vol. 11, no. 3, Article ID e1001508, 2013.

[75] B. A. Wible, M. Taglialatela, E. Ficker, and A. M. Brown, "Gating of inwardly rectifying K^+ channels localized to a single negatively charged residue," *Nature*, vol. 371, no. 6494, pp. 246–249, 1994.

[76] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *Journal of Molecular Biology*, vol. 238, no. 1, pp. 54–61, 1994.

[77] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *Journal of Molecular Biology*, vol. 305, no. 3, pp. 567–580, 2001.

[78] V. N. Uversky, "Intrinsically disordered proteins from A to Z," *The International Journal of Biochemistry and Cell Biology*, vol. 43, no. 8, pp. 1090–1103, 2011.

[79] L. B. Chemes, J. Glavina, L. G. Alonso, C. Marino-Buslje, G. de Prat-Gay, and I. E. Sánchez, "Sequence evolution of the intrinsically disordered and globular domains of a model viral oncoprotein," *PLoS ONE*, vol. 7, no. 10, Article ID e47661, 2012.

[80] J. A. Eisen and C. M. Fraser, "Phylogenomics: intersection of evolution and genomics," *Science*, vol. 300, no. 5626, pp. 1706–1707, 2003.

[81] J. Huerta-Cepas, S. Capella-Gutierrez, L. P. Pryszcz, M. Marcet-Houben, and T. Gabaldon, "PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome," *Nucleic Acids Research*, vol. 42, pp. D897–D902, 2014.

[82] B. Boeckmann, M. Robinson-rechavi, I. Xenarios, and C. Dessimoz, "Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees," *Briefings in Bioinformatics*, vol. 12, no. 5, pp. 423–435, 2011.

[83] K. Sjölander, "Phylogenomic inference of protein molecular function: advances and challenges," *Bioinformatics*, vol. 20, no. 2, pp. 170–179, 2004.

[84] A. Franceschini, D. Szklarczyk, S. Frankild et al., "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research*, vol. 41, no. 1, pp. D808–D815, 2013.

[85] A. Pottekat, S. Becker, K. Spencer et al., "Insulin biosynthetic interaction network component, TMEM24, facilitates insulin reserve pool release," *Cell Reports*, vol. 4, no. 5, pp. 921–930, 2013.

[86] R. DasGupta and E. Fuchs, "Multiple roles for activated LEF/TCF transcription complexes during hair follicle development and differentiation," *Development*, vol. 126, no. 20, pp. 4557–4568, 1999.

[87] J. M. Kerkhof, A. G. Uitterlinden, A. M. Valdes et al., "Radiographic osteoarthritis at three joint sites and FRZB, LRP5, and LRP6 polymorphisms in two population-based cohorts," *Osteoarthritis and Cartilage*, vol. 16, no. 10, pp. 1141–1149, 2008.

[88] V. S. C. Wong, A. Yeung, W. Schultz, and P. L. Brubaker, "R-spondin-1 is a novel β-cell growth factor and insulin secretagogue," *The Journal of Biological Chemistry*, vol. 285, no. 28, pp. 21292–21302, 2010.

[89] J. A. Hiss and G. Schneider, "Architecture, function and prediction of long signal peptides," *Briefings in Bioinformatics*, vol. 10, no. 5, pp. 569–578, 2009.

[90] S. Khan and M. Vihinen, "Spectrum of disease-causing mutations in protein secondary structures," *BMC Structural Biology*, vol. 7, article no. 56, 2007.

[91] S. Hormoz, "Amino acid composition of proteins reduces deleterious impact of mutations," *Scientific Reports*, vol. 3, article 2919, 2013.

[92] C. Chakraborty, S. Bandyopadhyay, U. Maulik, and G. Agoramoorthy, "Topology mapping of insulin-regulated glucose transporter GLUT4 using computational biology," *Cell Biochemistry and Biophysics*, vol. 67, no. 3, pp. 1261–1274, 2013.

[93] C. X. Chan and M. A. Ragan, "Next-generation phylogenomics," *Biology Direct*, vol. 8, no. 1, article 3, 2013.

[94] H.-X. Hao, Y. Xie, Y. Zhang et al., "ZNRF3 promotes Wnt receptor turnover in an R-spondin-sensitive manner," *Nature*, vol. 484, no. 7397, pp. 195–200, 2012.

[95] K. Kashimada and P. Koopman, "Sry: the master switch in mammalian sex determination," *Development*, vol. 137, no. 23, pp. 3921–3930, 2010.

[96] D. C. Blaydon, Y. Ishii, E. A. O'Toole, H. C. Unsworth, and M. T. Teh, "The gene encoding R-spondin 4 (*RSPO4*), a secreted protein implicated in Wnt signaling, is mutated in inherited anonychia," *Nature Genetics*, vol. 38, no. 11, pp. 1245–1247, 2006.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 398147, 12 pages
http://dx.doi.org/10.1155/2014/398147

Research Article

Detecting Epistatic Interactions in Metagenome-Wide Association Studies by metaBOOST

Mengmeng Wu and Rui Jiang

MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST, Department of Automation, Tsinghua University, Beijing 100084, China

Correspondence should be addressed to Rui Jiang; ruijiang@tsinghua.edu.cn

Received 14 June 2014; Accepted 14 July 2014; Published 24 July 2014

Academic Editor: Hao-Teng Chang

Material and Methods. We recall the definition of epistasis and extend it for metagenomic biomarkers and then we describe the overview of our method metaBOOST and provide detailed information about each step of metaBOOST. *Results.* We describe the data sources for both simulation studies and real metagenomic datasets. Then, we describe the procedure of simulation studies and provide results for it. After that, we conduct real datasets studies and report the results. *Conclusions and Discussion.* Finally, we conclude our method and discuss some possible improvements for the future.

1. Introduction

The importance of microbial communities to human health has been attracting more and more attention, with examples including the reveal of associations between intestinal microbiome and Crohn's disease [1], the study of the relationships between gut microbiome and Type II diabetes [2], and many others [3–5]. In traditional studies of microbial communities, individual microorganisms should be cultured and sequenced separately. Although this method has successfully yielded thousands of complete bacterial genomes as recorded in GenBank [6], such limitations as the requirement of culturing individual microbes have greatly restricted the scope of applications of such a traditional approach. With recent advances of high throughput sequencing techniques, the acquirement of a microbial community has become a routine, resulting in the boom of metagenomics [6–9].

Particularly, the application of metagenomics to the study of human complex diseases has yielded the so-called metagenome-wide association (MGWA) studies [2]. For example, Qin et al. developed a two-stage MGWA study to explore relationships between gut microbiome and type II diabetes [2, 10]. Such studies have also revealed associations between intestinal microbiome and Crohn's disease [1, 3, 11] and obesity [12]. Typically, the goal of a MGWA study is to identify metagenomic markers in both genetic and functional levels, and the typical procedure of a MGWA study include three main steps. First, microbial communities for both case and control individuals are sequenced, and resulting reads are assembled to obtain a microbial gene scaffold. Second, sequence reads are mapped back to the gene scaffold, and abundance levels of microbial genes are estimated. Third, microbial genes are further mapped to known microbial organisms or a gene category, and abundance levels of organisms or gene groups are obtained. Forth, statistical or machine learning methods are applied to analyze abundance levels for both the case and the control data, and candidate microbial markers (either organisms or genes) are identified. Finally, in the marker validation phase, additional samples are sequenced to confirm the discovered markers.

Although a MGWA study has supplied us with a powerful way to search metagenomic markers associated with a disease under investigation, this approach may encounter similar problems as the traditional genome-wide association (GWA) studies. For example, it is believed that a complex disease is typically caused by multiple genetic factors, their interactions, or their interactive effects with environmental factors [13]. Particularly, interactions between genetic factors,

typically referred to as epistatic interactions or epistasis, have been believed as a common pathogenic mechanism for complex diseases. However, the detection of epistatic interactions for a GWA study is extremely difficult, due to the fact of that the underlying mechanism of epistatic interaction is largely unknown and the number of possible combinations of genetic factors is typically huge. Facing this challenge, a number of computational methods have been developed. For example, Nelson et al. proposed a combinatorial partitioning method to exhaustively search for a combinatorial genotype that had the most significant contributions to a continuous phenotype [14]. Ritchie et al. proposed a multifactor-dimensionality reduction (MDR) method in which exhaustive search was performed to detect combinations of loci with the highest classification capability [15]. Jiang et al. used a machine learning method called random forest to find combinations of genotypes that contribute most to the correct classification of case against control [16]. Zhang and Liu proposed a Bayesian partition approach called BEAM to find groups of genotypes with large posterior probability [17]. Tang et al. proposed the concept of epistatic module and designed a Gibbs sampling approach to detect such modules [18]. Strategies based on high performance computing were also designed and extended to be used with graphics processing units (GPU), yielding such highly efficient method as BOOST and GBOOST [19, 20].

With these understandings, we proposed in this paper the first study of epistatic interactions in MGWA studies. More specifically, we designed a method called metaBOOST to detect such epistatic interactions for metagenome-wide case-control data. Our method consists of three main steps: (1) inference of metagenomic abundance level, (2) detection of possible epistasis using statistical methods, and (3) validation and visualization of epistatic interactions patterns. We validated our method by using both simulated experiments and real datasets studies. Results not only demonstrate the effectiveness of our approach but also provide biological insights for the pathogenic mechanisms of microbial communities to human complex diseases.

2. Materials and Methods

2.1. Overview of metaBOOST. Studies in medical genetics have shown that epistasis, or epistatic interactions between two or more genes, widely exists in such human complex disease as diabetes [21], asthma [22], and many others [23]. Recent advances in statistical genetics have also resulted in the prosperity in computational methods for detecting epistatic interactions in GWA studies, with examples which include such statistical methods as BEAM [17] and epiMODE [18], such machine learning models as epiForest [16], and such high performance computing approaches as BOOST [19]. Nevertheless, the definition of epistasis is controversial. For example, Bateson first introduced in 1909 the concept of epistasis, referring to a masking effect that one locus prevents another locus from manifesting its effect [13]. Fisher further defined in 1918 the epistatic interaction of multiple alleles at different loci as the deviation from additivity when considering contributions of these alleles to a quantitative trait [13]. With this definition, an epistatic interaction can be characterized by a logistic regression model, as

$$\log \frac{P(Y=1)}{P(Y=0)} = \mu + \alpha X_1 + \beta X_2 + \gamma X_1 X_2, \quad (1)$$

where the response item at the left hand represents the log odds of the disease risk, X_1, X_2 at the right hand represent the independent effect caused by two different loci, and the multiplicative item (X_1X_2) represents the epistatic interaction. With this model, the epistatic interaction can be inferred by hypothesis testing whether the regression coefficient γ is equal to zero, which can be conducted by a likelihood ratio test. Furthermore, by enumerating all possible combinations of loci pairs and performing statistical tests, we are able to detect all epistatic interactions. However, because of the large number of loci in a genome-wide association studies, such an exhaustive search can hardly be practical, and a variable selection strategy should then be applied to reduce the search space.

On the other hand, in a metagenome-wide association study with the case-control design, one typically sequences the microbial community of a number of patients and normal individuals, obtains gene scaffold by assembling the sequencing data, mapping sequence reads to the scaffolds to obtain abundance levels of the genes, and applies statistical approaches to test whether the abundance level of such a gene is significantly different between the case and the control. In such a study, a microbial gene is used as a marker, analogous to a locus in a traditional genome-wide association study. However, there also exists significant difference between metagenome-wide and genome-wide association studies. For example, the number of markers in a metagenome-wide study can be as large as 4 million, while that in a genome-wide study is typically less than 1 million. Another more important difference is the property of markers. In a genome-wide study, markers are factors, while in a metagenome-wide study, markers are continuous variables. Obviously, the former difference requires more efficient approaches in a metagenome-wide study, while the latter difference suggests either the customization of methods in genome-wide study to facilitate the manipulation of continuous variables or the development of novel statistical methods that take the continuous nature of microbial genes into consideration.

With the above analysis, we proposed in this paper the analysis of epistatic interactions between microbial genes. More specifically, we proposed a bioinformatics approach called metaBOOST that is designed based on a highly efficient epistasis detection algorithm named BOOST [19] and includes two extra steps: the discretization of abundance levels of microbial genes and a permutation test for accessing the statistical significance of candidate epistatic interactions. In detail, as shown in Figure 1, we first mapped metagenome sequencing reads to representative sequences of both microbial genus and KEGG orthologous (KO) groups to obtain abundance levels of genus and KO groups. Then, we fitted the distribution of the abundance levels to a mixed Gaussian distribution and estimate the associated parameters using an

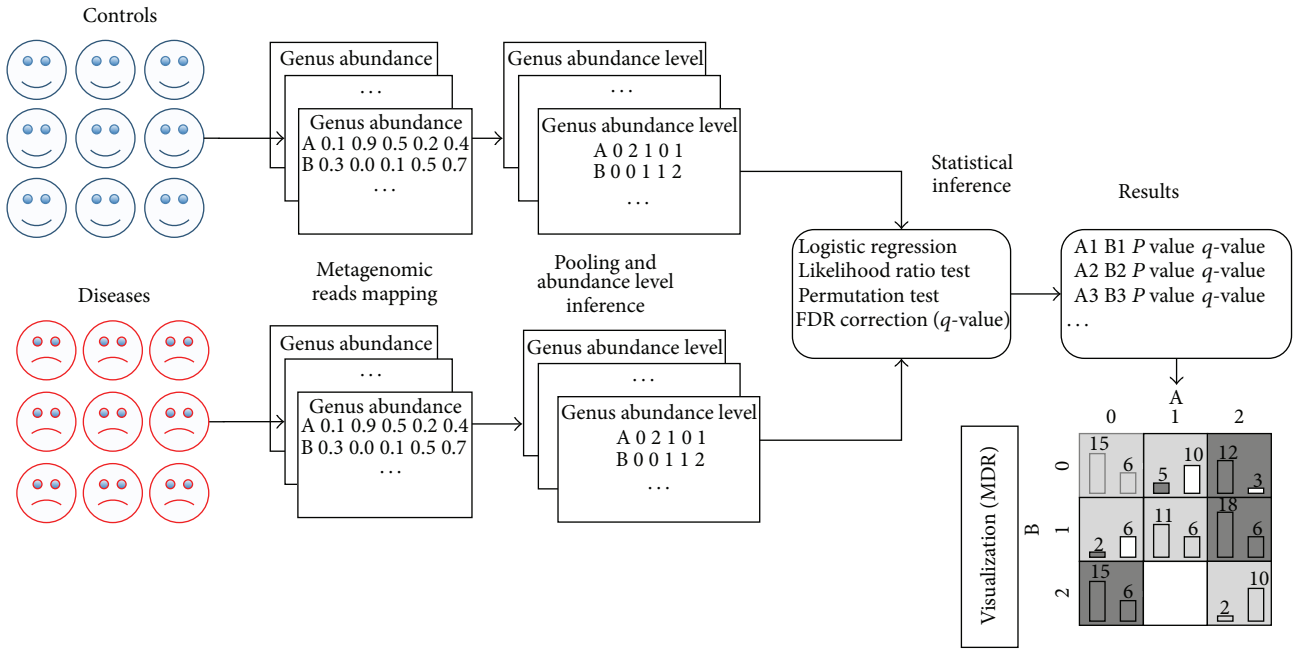


FIGURE 1: The overall procedure of metaBOOST: (1) obtain abundance through reads mapping; (2) infer abundance levels using Gaussian mixture model and EM algorithm; (3) identify possible epistatic interactions via logistic regression and permutation test; (4) validate identified epistatic interactions using MDR.

Expectation-Maximization algorithm. Next, we discretized continuous abundance to factors of at most three levels, analogous to genotypes in genome-wide association studies. Finally, we used a highly efficient algorithm in genome-wide study called BOOST to detect candidate interactions. In order to assess statistical significance of such candidate interactions, we further perform a permutation test and control the false discovery rate at a desired level.

2.2. Inference of Microbial Abundance Levels. The objective of this step is to obtain abundance levels of microbial genus or KEGG orthologous groups. In detail, given sequencing reads generated by a deep sequencing technology (typically illumina and 454), we mapped the reads to known microbial genes and summarized the number of mapped reads to obtain abundance levels of a microbial gene. The abundance levels are further normalized to illuminate the possible influence of different sequencing depths for different samples. Then, we made use of mapping between genes and genus to obtain genus abundance by summing up all of the abundance of genes corresponding to the same genus, in which gene length is considered and serves as divider since longer genes generate more reads. Similarly, we obtained KEGG orthologous (KO) abundance.

The abundance is continuous, ranging from 0 to 1 and in nature different from GWAS data. In order to utilize the methods developed for GWAS, we discretized the abundance data first into two or three levels, that is, $\{0, 1\}$ or $\{0, 1, 2\}$. Particularly, in the two-level case, 0 represents nonexistence and 1 existence. In the three-level case, 0 represents nonexistence, 1 low-level existence, and 2 high-level existence. In our method, the discretization is performed automatically by fitting abundance levels of both case and control populations to a Gaussian mixture model and estimates the parameters using an Expectation-Maximization algorithm. And there exists two strategies for determining the number of levels for abundance data. First, we can use some objective criteria such as Bayesian information criterion (BIC). Second, we can use some empirical and heuristic strategies such as discarding a level that contains mainly zeros. In our paper, we adopt the latter for the purpose of seeking for simplicity.

2.3. Detection of Microbial Epistatic Interactions. The most popular method to detect epistasis is likelihood ratio test or logistic regression, in which two models are considered, that is, model with interaction term and the model without interaction term, such as the following:

without interaction,

$$\log \frac{P(Y=1)}{P(Y=0)} = \mu + \alpha X_1 + \beta X_2, \tag{2}$$

with interaction:

$$\log \frac{P(Y=1)}{P(Y=0)} = \mu + \alpha X_1 + \beta X_2 + \gamma X_1 X_2. \tag{3}$$

After fitting the above two models, we calculated the likelihood ratio or log-likelihood difference and this score is object to chi-squared distribution. A P value can then be computed and serves as a measurement of the strength of the epistatic interaction. Considering that the theoretical P value

Permutation test for metaBOOST
Calculate the likelihood ratio using original data, denoted as *Original*.
for $i = 1 : N$
  Permute the disease-control label of samples.
  Calculate the likelihood ratio using permuted labels, denoted as *Permute*(i).
end
Final P value would be $P_{permuted} = \dfrac{\#\{Permute > Original\}}{N}$.

Procedure 1: Procedure for the permutation test.

derived from chi-squared test may not reflect the true null distribution, we further use permutation test as described briefly in Procedure 1 to obtain simulated P value and derive q-value to control false discovery rate. Here, we adopted BOOST for the calculation of P value since this method was validated for its power and speed recently [24].

2.4. Visualization and Validation. After the above two steps, we obtained a list containing potential epistasis with correspond q-values. We then adopted the strategy used in multifactor dimension reduction (MDR) [15] to supply figures for visualizing patterns of epistasis detected. Briefly, MDR include four steps. (1) Select possible associated factors or factors under investigation. (2) List all possible combinations of the factors selected before in two or higher dimensions and evaluate the relative ratio in a model. (3) In order to assess every model in more accurate manner, cross-validation (usually tenfold) is used and average classification error is computed. (4) A list of all possible interaction models with their errors is obtained.

In general, there exist two strategies for validation. First, we can validate detected epistatic interactions by using independent computational methods. Second, we can perform validation through an experimental way, such as using extra samples to check the statistical significance of our findings. As for the former, systematic comparisons between different computational methods have been carried out intensively in literature [24, 25], and thus we think it is unnecessary to repeat such work in our paper. As for the latter, considering that an experiment is costly, time-consuming, and beyond the theoretical purpose of our paper, we would like to leave this possibility to some future projects.

3. Results

3.1. Data Sources. In simulation studies, we relied on existing benchmark datasets for epistatic detection to generate artificial data for evaluating the effectiveness of our method. Briefly, Velez et al. explored 70 models with different penetrance functions and generated a total of 42,000 simulated datasets [26]. From this resource, we selected a small number of 10 models and generated two datasets of 200 and 400 samples for each model. In each dataset, a total of 1000 loci (998 random and 2 of epistatic interaction) were generated. To simulate abundance levels, we further resorted to a Gaussian mixture model as detailed in the next section. Finally, we obtain 2 datasets (200 and 400 samples) for each of the 10 epistatic models.

For real metagenome-wide association studies, we selected two real datasets published recently [2, 27]. In detail, as part of the MetaHIT project, Qin et al. sequenced faecal samples of 124 European individuals and studied the impact of gut microbes on human health [27]. Among those individuals, 99 were infected with inflammatory bowel disease and the others were not. After metagenomic sequencing on the illumina platform, 576.7 Gb of sequence reads were generated, and each individual owns 4.5 Gb of sequence reads on average. Then, those huge amounts of reads were assembled into contigs by a de Bruijn graph-based method called SOAPdenovo [28]. Then, the metagene was used to predict long ORFs (longer than 100 bp) and those nonredundant ORFs (3.3 million in total) were considered as "genes" of microbes [29]. Next, those ORFs were mapping into reference microbial genomes to obtain microbial genus abundance and KEGG orthologue groups abundance. Another real metagenome-wide association studies dataset was generated by similar procedures [2]. In this study, 368 Chinese individuals' stool samples were sequenced and 1209.2 Gb of sequence reads was generated. Those individuals contain 183 patients of type 2 diabetes and 185 normal controls. After assembling and metagenomic gene prediction, 4,267,985 predicted genes were obtained and after mapping, 6,313 KEGG orthologues were obtained.

3.2. Simulation Studies. Existing metagenomics simulator such as MetaSim [30] can simulate sequencing data for metagenomics but cannot embed possible epistatic interactions in the simulated data. Existing methods for simulating genome-wide association studies can simulate case-control data but cannot embed epistatic interaction patterns into metagenomics data. Therefore, we proposed a simple method based on Gaussian mixture model to simulate abundance levels of microbial markers such as genus and KO. In detail, we adopt a two-step procedure as described below.

We generated epistatic data in continuous case based on discrete case, as formulated below:

$$y_i = \begin{cases} 0 & x_i = 0 \\ \text{sampling from Normal}\left(\mu_1, \sigma_1\right) & x_i = 1 \\ \text{sampling from Normal}\left(\mu_2, \sigma_2\right) & x_i = 2, \end{cases} \quad (4)$$

where x_i denotes genotype data generated by simulator and y_i corresponding abundance levels.

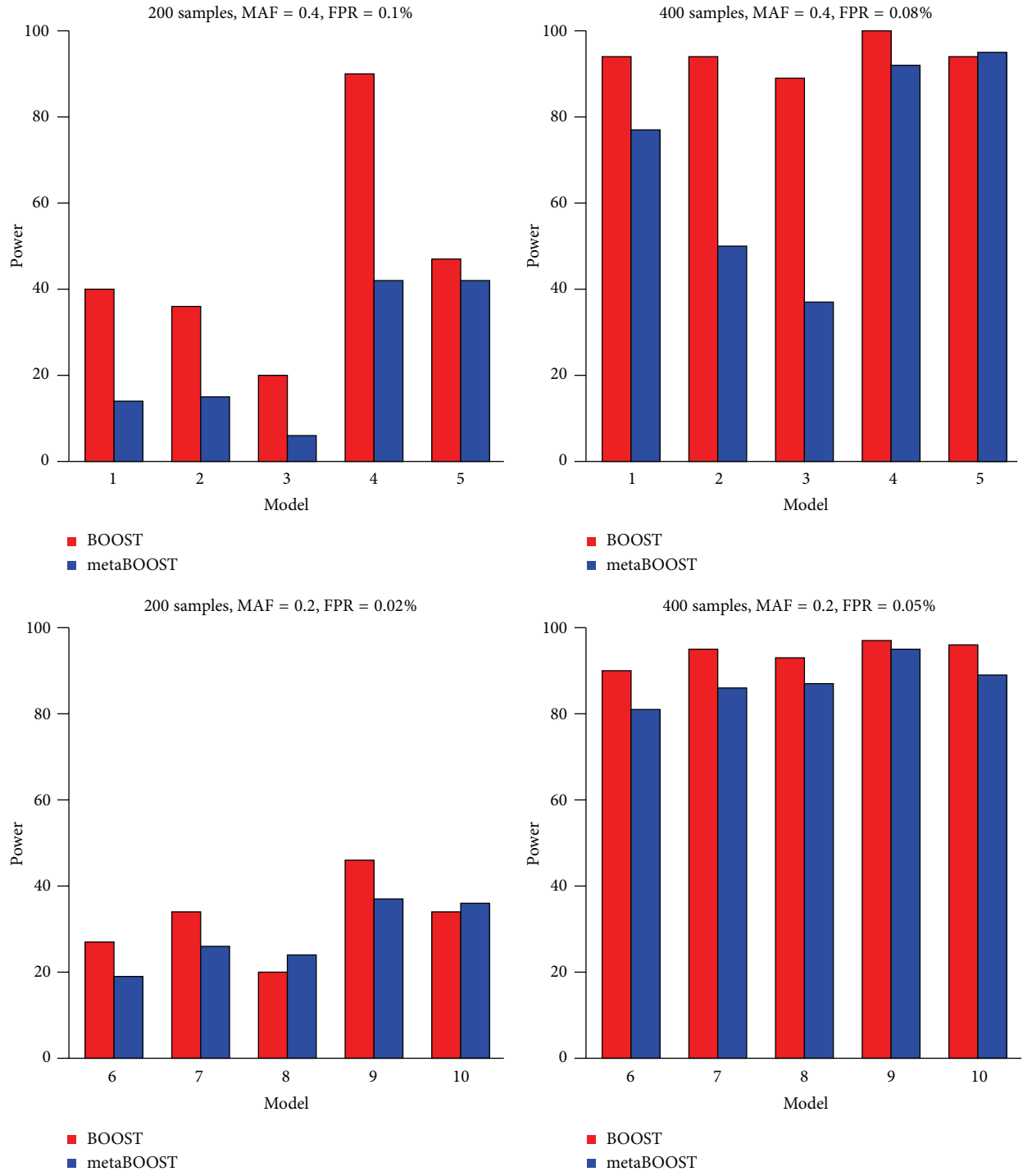


Figure 2: The results of simulation studies. We compare BOOST and metaBOOST on 10 epistatic interaction models. 100 datasets, each of 1000 markers, are generated for each model. The power is defined as proportion of identify the true epistatic interaction successfully.

After the generation of simulated data, we applied BOOST to identify embedded epistatic interactions in the original case-control data and metaBOOST to identify epistatic interactions in the simulated metagenomic case-control data, and we present results in Figure 2, in which the power of a method on a model is defined as the proportion of the ground-truth epistatic interactions being identified. From the figure, we first see clearly the effectiveness of metaBOOST in detecting epistatic interactions embedded metagenomic case-control data. For example, with 400 samples and an MAF of 0.2 and an FPR of 0.06% (subplot D), the power of metaBOOST for the five interaction models ranges from 0.80 to 0.90, suggesting the successful identification of the embedded interactions. Second, we observe that the power

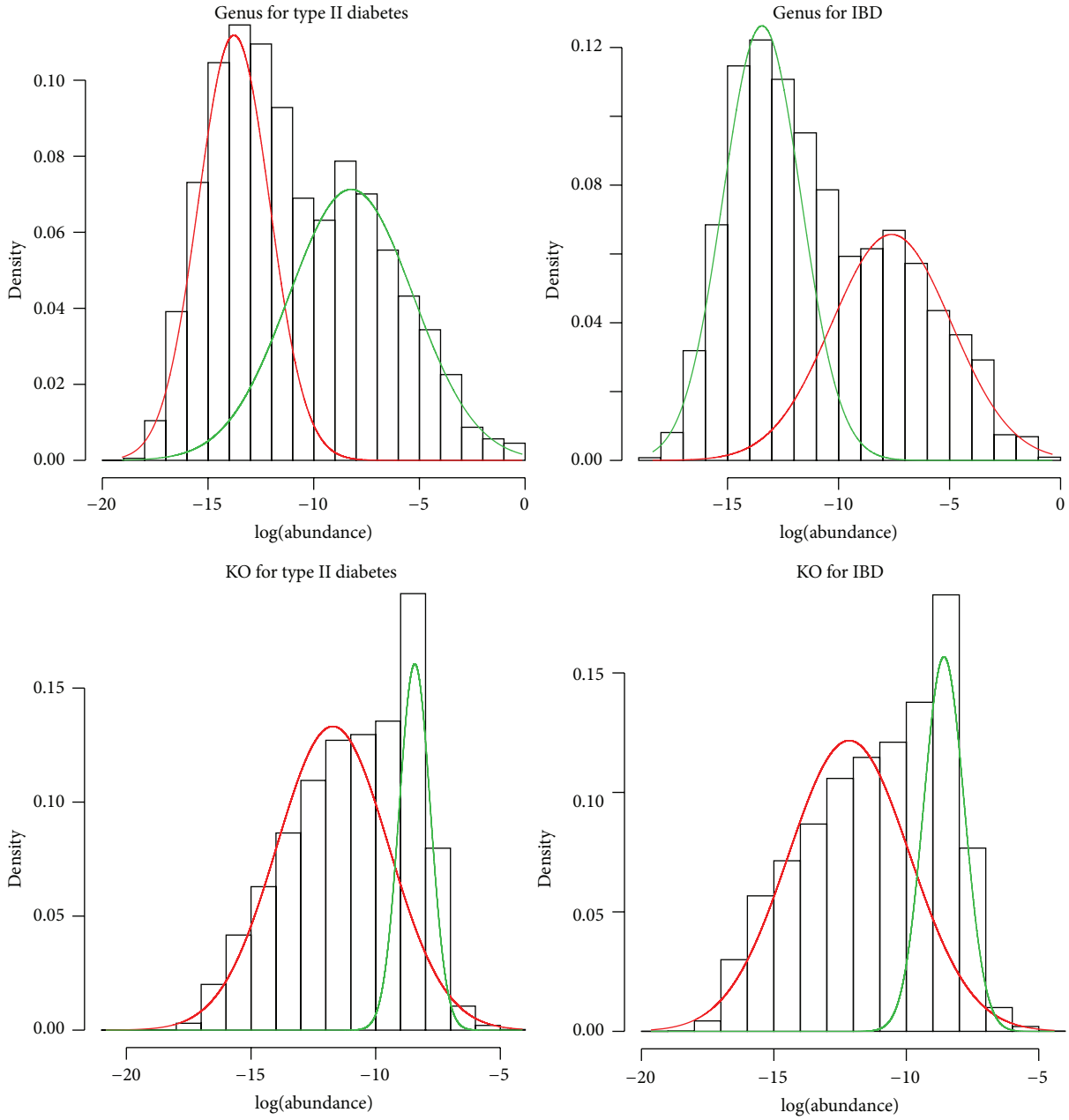


FIGURE 3: The distribution of abundance levels of genus and KOs for type 2 diabetes dataset and IBD dataset. The red and green curve are modeled by Gaussian mixture model and computed by EM algorithm.

of metaBOOST for metagenomic data is in general not as high as BOOST for genomic data. This phenomenon is easy to understand because the extra step of converting continuous abundance levels to discrete factors by fitting the Gaussian mixture model may introduce extra noises in the data, thus resulting in the loss of the detecting power. Finally, we observe that the power of metaBOOST depends heavily on the number of samples. More specifically, with 200 samples, epistatic interactions embedded in model 3 can hardly be detected (power < 10% for MAF = 0.4 and FPR = 0.1%), while with 400 samples, epistatic interactions embedded in model 3 can reach about 40% (for MAF = 0.4 and FPR = 0.1%).

3.3. Genus Epistasis for Type II Diabetes. With the power of metaBOOST being verified by the simulation studies, we applied this method to a real metagenome-wide case-control dataset of gut microbe of 368 Chinese individuals [2], including 183 patients of type II diabetes and 185 healthy individuals.

We first fitted the distribution of abundance levels of genus of both the case and the control populations using a Gaussian mixture model and find that a two-component model fits the data very well (Figure 3) except for nonexistence. We then discretized abundance levels of genus to three categories. Because the number of genuses is not large, we

Table 1: Top 10 candidate genus epistatic interactions in the type II diabetes dataset.

Genus A	Genus B	P value	q-value
Bifidobacterium	*Actinobacillus*	0	0
Aggregatibacter	*Arcanobacterium*	$6e-6$	0.084
Pyramidobacter	*Proteus*	$8e-6$	0.084
Acidaminococcus	*Ureaplasma*	$2e-5$	0.119
Rhizobium	*Veillonella*	$2.3e-5$	0.119
Micrococcus	*Scardovia*	$2.4e-5$	0.119
Megamonas	*Selenomonas*	$2.8e-5$	0.119
Burkholderiales	*Abiotrophia*	$3e-5$	0.119
Kingella	*Desulfotomaculum*	$3.5e-5$	0.123
Catenibacterium	*Aliivibrio*	$4.2e-5$	0.133

Table 2: Top 10 candidate genus epistatic interactions in the IBD dataset.

Genus A	Genus B	P value	q-value
Lautropia	*Thauera*	$4e-6$	0.127
Slackia	*Cellulosilyticum*	$1.3e-5$	0.206
Abiotrophia	*Catenibacterium*	$3.6e-5$	0.292
Streptobacillus	*Anaerofustis*	$4e-5$	0.292
Edwardsiella	*Peptoniphilus*	$6.3e-5$	0.292
Gordonibacter	*Fusobacterium*	$6.7e-5$	0.292
Mobiluncus	*Acinetobacter*	$8e-5$	0.292
Enterobacter	*Thermoanaerobacter*	$8.3e-5$	0.292
Lautropia	*Ureaplasma*	$8.3e-5$	0.292
Corynebacterium	*Leptotrichia*	$1.03e-4$	0.305

enumerated all pairwise interactions of the genus and fitted a logistic regression model for each pair. We further performed the permutation test 1,000,000 times to estimate a P value for each pair and derive a q-value to characterize the statistical significance. Finally, we list top 10 interactions of the smallest q-values in Table 1.

In the list, the interaction between *Bifidobacterium* and *Actinobacillus* is reported with a very significant q-value ($<10^6$). By reviewing the literature, we find that this potential epistatic interaction has been supported by several existing studies. For example, in literature [10, 31, 32], *Bifidobacterium* was reported to be associated with type 2 diabetes. The increase in abundance level of this genus will improve high-fat-diet-induced diabetes in mice, and such phenomenon shows significant difference between type 2 diabetes and healthy people. Besides, in literature [33], *Actinobacillus* was suggested to be associated with type II diabetes. Therefore, it is reasonable to assume that the interaction between *Actinobacillus* and *Bifidobacterium* may further increase the risk of carrying type 2 diabetes.

We then plot four possible interaction patterns generated by MDR in Figure 4. From the figure, we can clearly see the interaction patterns between two genuses. For example, in the left bottom subfigure, the disease risk is low if both of the abundance levels of *Megamonas* and *Selenomonas* are low or high and otherwise the disease risk is high. Similar interactions pattern can also be found in other subfigures.

3.4. Genus Epistasis for IBD. We also applied metaBOOST to another real metagenome-wide case-control dataset of gut microbe of 124 European individuals [27], including 25 patients of inflammatory bowel disease and 99 individuals without this disease. We first resorted to the same procedure as the above section to obtain abundance of microbial genus. Then, we used Gaussian mixture model to fit the distribution of abundance of genus of both the case and the control and Figure 3 shows that two-component model is also a good choice for this dataset. So, we discretized the abundance into low and high levels corresponding to the two components in the GMM plus nonexistence, which leads to three categories. The same logistic regression model and permutation test are used to output P value and q-value for each pair of genuses, and we list top 10 interactions with the smallest q-values in Table 2.

In the list, the interaction between *Lautropia* and *Thauera* is reported to have the smallest P value ($4e-6$) and the smallest q-value (0.127). By reviewing literature, we find that both of *Lautropia* and *Thauera* belong to Proteobacteria, which is reported to be associated with IBD in [34] and the dysbiosis of Proteobacteria can result in IBD. So, it is reasonable to think that the interaction between *Lautropia* and *Thauera* may play an important role in the process of dysbiosis and IBD.

We then plot four potential interaction patterns generated by MDR in Figure 5. Different from above section, we do not plot all three levels of each genus but only the two of three with significant number of samples and the remaining cells were discarded because of too little samples. But this does not affect the patterns themselves and we can still discover important interaction patterns from here. For example, in the left top subfigure, the disease risk is low if both of the abundance levels of *Slackia* and *Cellulosilyticum* are high or low, and the disease risk increases significantly otherwise. And as another example at the right bottom subfigure, the disease risk is high when both of the abundance levels of *Mobiluncus* and *Acinetobacter* are low or high, and the disease risk decreases otherwise.

3.5. KEGG Orthologous Epistasis for Type II Diabetes. Besides genus, functional annotation such as KEGG orthologues groups is also important for understanding the functions of human gut microbiome [2]. After having gene reference or catalogue of the 368 samples, functional annotation using KEGG and eggNOG database was performed. Finally, 6,313 KEGG orthologues (KO) are identified, which covered 47.1% and 60.9% of gene catalogue, respectively. And corresponding read counts can be converted into the abundance levels of microbial KOs.

We then applied metaBOOST only on KEGG orthologue groups or KO since more annotation information can be retrieved. Similarly, we fitted the abundance of KO using Gaussian mixture model with EM algorithm, and Figure 6 tells us that the two-component model is a relatively good choice. Here, we have to adopt a stepwise strategy since

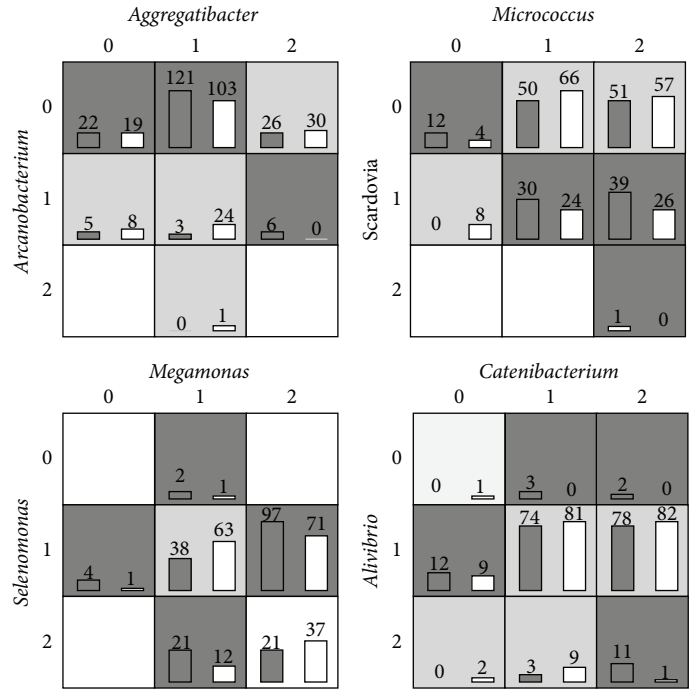


Figure 4: Four potential epistatic interactions between genuses in the type II diabetes dataset.

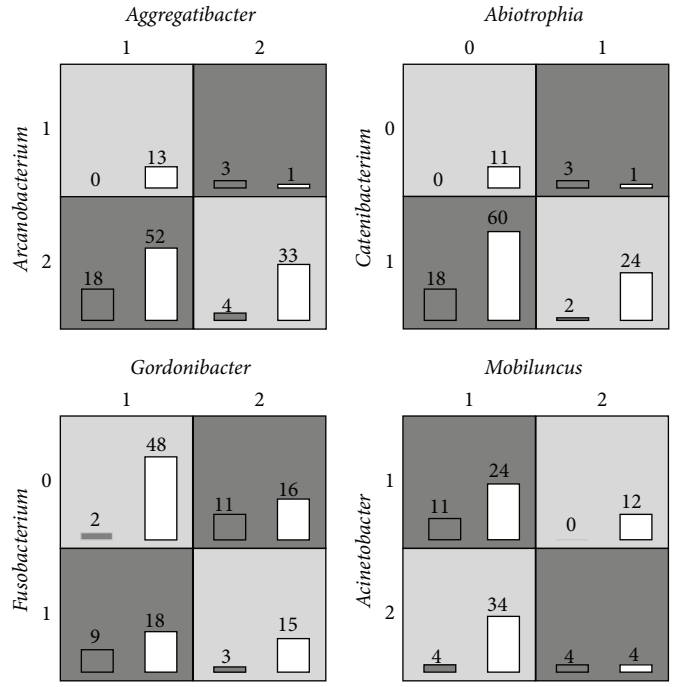


Figure 5: Four potential epistatic interactions between genuses in the IBD dataset.

the number of KOs is 6,313 and the number of combinations of KO pair would be as high as ~2 million. First, we fit logistic regression models with only one KO and compute its P value. The P value can measure the effect of one KO on disease and we only select those 265 KOs with P value < 0.1. Then we enumerate all pairwise interactions of the selected KO and fit a logistic regression model for each pair. We further perform the permutation test 100,000 times to estimate a P value for each pair and derive a q-value to characterize the statistical significance. Finally, we list top 10 interactions of the smallest q-values in Table 3.

In the list, all interactions are reported to have small P values ($\leq 2e-5$) and small q-values (≤ 0.06). But to the best of our knowledge, we find no literature to associate KOs

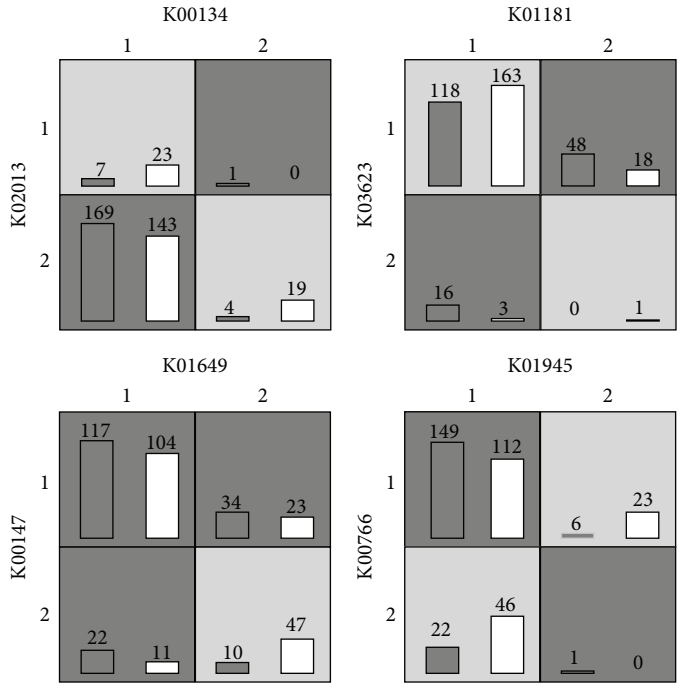


Figure 6: Four potential epistatic interactions between KOs in the type II diabetes.

Table 3: Top 10 candidate KO epistatic interactions in the Type II diabetes dataset.

KO A	KO B	P value	q-value
K01760	K00948	0	0
K00134	K02013	0	0
K00134	K01756	0	0
K01181	K03623	0	0
K07487	K00811	0	0
K01649	K00147	0	0
K01945	K00766	$1e-5$	0.04
K03427	K11928	$1e-5$	0.04
K01945	K00968	$2e-5$	0.06
K05808	K01808	$2e-5$	0.06

Table 4: Common pathways involved in KO epistatic interactions in the type II diabetes dataset.

KO A	KO B	Common pathways
K01760	K00948	ko01230, ko01110, and ko01100
K00134	K01756	ko01110, ko01100
K01649	K00147	ko01230, ko01100
K01945	K00766	ko01110, ko01100
K01945	K00968	ko01100

with type 2 diabetes. Then, we map KOs into pathways using KEGG database and find 5 of 10 epistatic interactions that can share the same pathways as Table 4. If two KOs or two genes are involved in the same pathway, it is reasonable to think they may interact with each other in some way. And those common pathways are ko01230 (biosynthesis of amino acids), ko01110 (biosynthesis of secondary metabolites), and ko01100 (metabolic pathway). Obviously, those pathways are associated with biosynthesis and metabolic activities, which are important for microbes' life.

We then plot four potential interaction patterns generated by MDR in Figure 6. Again, we only plot two of three abundance levels with enough samples and omit those cells with small samples. From the figure, we can discover some obvious interaction patterns. For example, in the left bottom subfigure, the disease risk is low only when both of the abundance levels of K01649 and K00147 are high. This may tell that K01649 and K00147 can cooperate with each other to prevent type 2 diabetes. As another example, in the tight top subfigure, omitting the cells containing only 1 sample, we find that the disease risk is low only when both of the abundance levels of K01181 and K03623 are low. This may tell that K01181 and K03623 are both harmful for type 2 diabetes and increase in one of them can result in high disease risk.

3.6. KEGG Orthologous Epistasis for IBD. We then applied the same procedure as detailed in the above sections to analyze the IBD data. We mapped genes in the gene catalogue of the 124 European samples into the KEGG database to obtain the abundance of KEGG orthologue groups. We also fitted the distribution of abundance using Gaussian mixture model and we can see that the two-component model is also a good choice from Figure 6. So we discretized abundance levels of KO to three categories. Here, we also adopted a stepwise strategy to select significant KOs using one-variable logistic regression firstly, followed by epistatic interactions

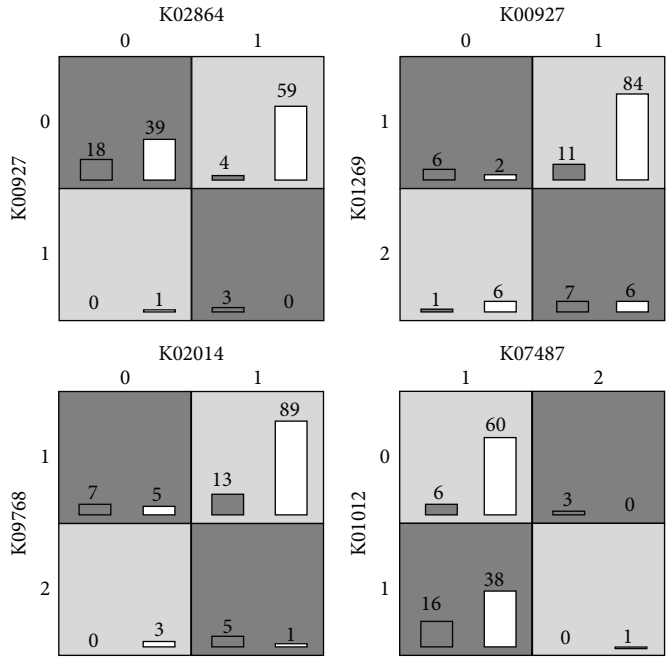


Figure 7: Four potential epistatic interactions of KOs in the IBD dataset.

Table 5: Top 10 candidate KO epistatic interactions in the IBD dataset.

KO A	KO B	P value	q-value
K02864	K00927	0	0
K07023	K06881	0	0
K02967	K11921	$1e-5$	0.136
K00927	K03760	$2e-5$	0.204
K09768	K10041	$4e-5$	0.326
K00927	K01269	$8e-5$	0.466
K07487	K03186	$8e-5$	0.466
K02014	K09768	$1.1e-4$	0.560
K07487	K01012	$1.4e-4$	0.593
K03390	K00793	$1.6e-4$	0.593

detection with two-variable logistic regression. Permutation tests (100,000 times) are also used to estimate a P value for each pair and derive a q-value to characterize the statistical significance. Finally, we list top 10 interactions of the smallest q-values in Table 5.

In the list, only the first two pairs have significant small q-value (<0.05). After mapping into pathways, we cannot find common pathways in those KOs. We then plot four potential interaction patterns generated by MDR in Figure 7. From the figure, we can see that at least two of four cells have small samples, such as only 1 and 3 samples in the left top subfigure. This tells that the epistatic interactions between KOs are weak or not significant. The possible reasons may be the following: (1) the number of samples is small and the number of IBD patients is just 25 and (2) the epistatic interactions of KOs in this IBD dataset are hard to be identified.

4. Conclusions and Discussion

In this paper, we proposed a method called metaBOOST to detect epistatic interactions in metagenome-wide association studies. We first resorted to a Gaussian mixture model to automatically discretize abundance levels of microbial genus and microbial genes to categorical values and then relied on a logistic regression model to detect epistatic interactions at the genus and KO level. Results not only show the effectiveness of our approach in simulation studies but also suggest the existence of several potential epistatic interactions between microbial biomarkers in two real datasets of human gut microbial communities. The merit feature of our approach is the automatic discretization of abundance levels of microbial genus and genes. As one of the main differences between metagenome-wide and genome-wide association studies, the continuous form of microbial biomarkers brings the main difficulty for detecting epistatic interactions between such markers and makes the discretization step the prerequisite. Resorting to the Gaussian mixture model, as we have done in this paper, is certainly an effective way.

Certainly, our method can further be extended from the following aspects. First, since the understanding of epistatic interaction is not unique, the exploration of epistatic interactions between microbial biomarkers under different definitions will be necessary. Second, although our method of discretizing continuous abundance levels of microbial biomarkers has been demonstrated to be effective, it is still worth pursuing to directly build a statistical or machine

learning method that is capable of detecting epistatic interactions between numeric-valued markers. The main difficulty is that the huge number of markers in a metagenome-wide association study prevents the exhaustive search of combinations of microbial markers, and the continuous form of such markers prevents the application of highly efficient computational tricks such as bit-wise operations that have been adopted in existing methods in genome-wide studies.

Conflict of Interests

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

Acknowledgments

This research was partially supported by the National Basic Research Program of China (2012CB316504), the National High Technology Research and Development Program of China (2012AA020401), and the National Natural Science Foundation of China (61175002).

References

[1] C. Manichanh, L. Rigottier-Gois, E. Bonnaud et al., "Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach," *Gut*, vol. 55, no. 2, pp. 205–211, 2006.

[2] J. Qin, Y. Li, Z. Cai et al., "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, vol. 490, no. 7418, pp. 55–60, 2012.

[3] S. Greenblum, P. J. Turnbaugh, and E. Borenstein, "Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 2, pp. 594–599, 2012.

[4] I. Cho and M. J. Blaser, "The human microbiome: at the interface of health and disease," *Nature Reviews Genetics*, vol. 13, no. 4, pp. 260–270, 2012.

[5] J. M. Kinross, A. W. Darzi, and J. K. Nicholson, "Gut microbiome-host interactions in health and disease," *Genome Medicine*, vol. 3, no. 3, article 14, 2011.

[6] J. C. Wooley, A. Godzik, and I. Friedberg, "A primer on metagenomics," *PLoS Computational Biology*, vol. 6, no. 2, Article ID e1000667, 2010.

[7] The Human Microbiome Project Consortium, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.

[8] S. D. Ehrlich, "MetaHIT: The European Union Project on metagenomics of the human intestinal tract," in *Metagenomics of the Human Body*, pp. 307–316, Springer, 2011.

[9] M. Arumugam, J. Raes, E. Pelletier et al., "Enterotypes of the human gut microbiome," *Nature*, vol. 473, no. 7346, pp. 174–180, 2011.

[10] K. Lê, Y. Li, X. Xu et al., "Alterations in fecal Lactobacillus and Bifidobacterium species in type 2 diabetic patients in Southern China population," *Frontiers in Physiology*, vol. 3, Article ID Article 496, 2013.

[11] N. Larsen, F. K. Vogensen, F. W. J. van den Berg et al., "Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults," *PLoS ONE*, vol. 5, no. 2, Article ID e9085, 2010.

[12] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon, "An obesity-associated gut microbiome with increased capacity for energy harvest," *Nature*, vol. 444, no. 7122, pp. 1027–1031, 2006.

[13] H. J. Cordell, "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans," *Human Molecular Genetics*, vol. 11, no. 20, pp. 2463–2468, 2002.

[14] M. Nelson, S. Kardia, R. Ferrell, and C. Sing, "A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation," *Genome Research*, vol. 11, no. 3, pp. 458–470, 2001.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] L. S. Yung, C. Yang, X. Wan, and W. Yu, "GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies," *Bioinformatics*, vol. 27, no. 9, pp. 1309–1310, 2011.

[21] Y. M. Cho, M. D. Ritchie, J. H. Moore et al., "Multifactor-dimensionality reduction shows a two-locus interaction associated with type 2 diabetes mellitus," *Diabetologia*, vol. 47, no. 3, pp. 549–554, 2004.

[22] T. D. Howard, G. H. Koppelman, J. Xu et al., "Gene-gene interaction in asthma: Il4ra and il13 in a dutch population with asthma," *American Journal of Human Genetics*, vol. 70, no. 1, pp. 230–236, 2002.

[23] Ö. Carlborg and C. S. Haley, "Epistasis: too often neglected in complex trait studies?" *Nature Reviews Genetics*, vol. 5, no. 8, pp. 618–625, 2004.

[24] Y. Wang, G. Liu, M. Feng, and L. Wong, "An empirical comparison of several recent epistatic interaction detection methods," *Bioinformatics*, vol. 27, no. 21, pp. 2936–2943, 2011.

[25] 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.

[26] D. R. Velez, B. C. White, A. A. Motsinger et al., "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," *Genetic Epidemiology*, vol. 31, no. 4, pp. 306–315, 2007.

[27] J. Qin, R. Li, J. Raes et al., "A human gut microbial gene catalogue established by metagenomic sequencing," *Nature*, vol. 464, no. 7285, pp. 59–65, 2010.

[28] R. Luo, B. Liu, Y. Xie et al., "OAPdenovo2: an empirically improved memory-efficient short-read de novo assembler," *Gigascience*, vol. 1, no. 1, article 18, 2012.

[29] H. Noguchi, J. Park, and T. Takagi, "MetaGene: prokaryotic gene finding from environmental genome shotgun sequences," *Nucleic Acids Research*, vol. 34, no. 19, pp. 5623–5630, 2006.

[30] D. C. Richter, F. Ott, A. F. Auch, R. Schmid, and D. H. Huson, "MetaSim—a sequencing simulator for genomics and metagenomics," *PLoS ONE*, vol. 3, no. 10, Article ID e3373, 2008.

[31] X. Xu, H. Hui, and D. Cai, "Differences in fecal Bifidobacterium species between patients with type 2 diabetes and healthy individuals," *Nan Fang Yi Ke Da Xue Xue Bao*, vol. 32, no. 4, pp. 531–564, 2012 (Chinese).

[32] P. D. Cani, A. M. Neyrinck, F. Fava et al., "Selective increases of bifidobacteria in gut microflora improve high-fat-diet-induced diabetes in mice through a mechanism associated with endotoxaemia," *Diabetologia*, vol. 50, no. 11, pp. 2374–2383, 2007.

[33] Y. Zheng, N. Ma, X. Hu, and L. Zhang, "Effect of actinobacillus actinomycetem on the secretion of interleukin-6 and apoptosis rate of polymorphonuclear leukocyte in type 2 diabetes patients," *West China Journal of Stomatology*, vol. 29, no. 3, pp. 286–288, 2011.

[34] I. Mukhopadhya, R. Hansen, E. M. El-Omar, and G. L. Hold, "IBD—what role do Proteobacteria play?" *Nature Reviews Gastroenterology and Hepatology*, vol. 9, no. 4, pp. 219–230, 2012.