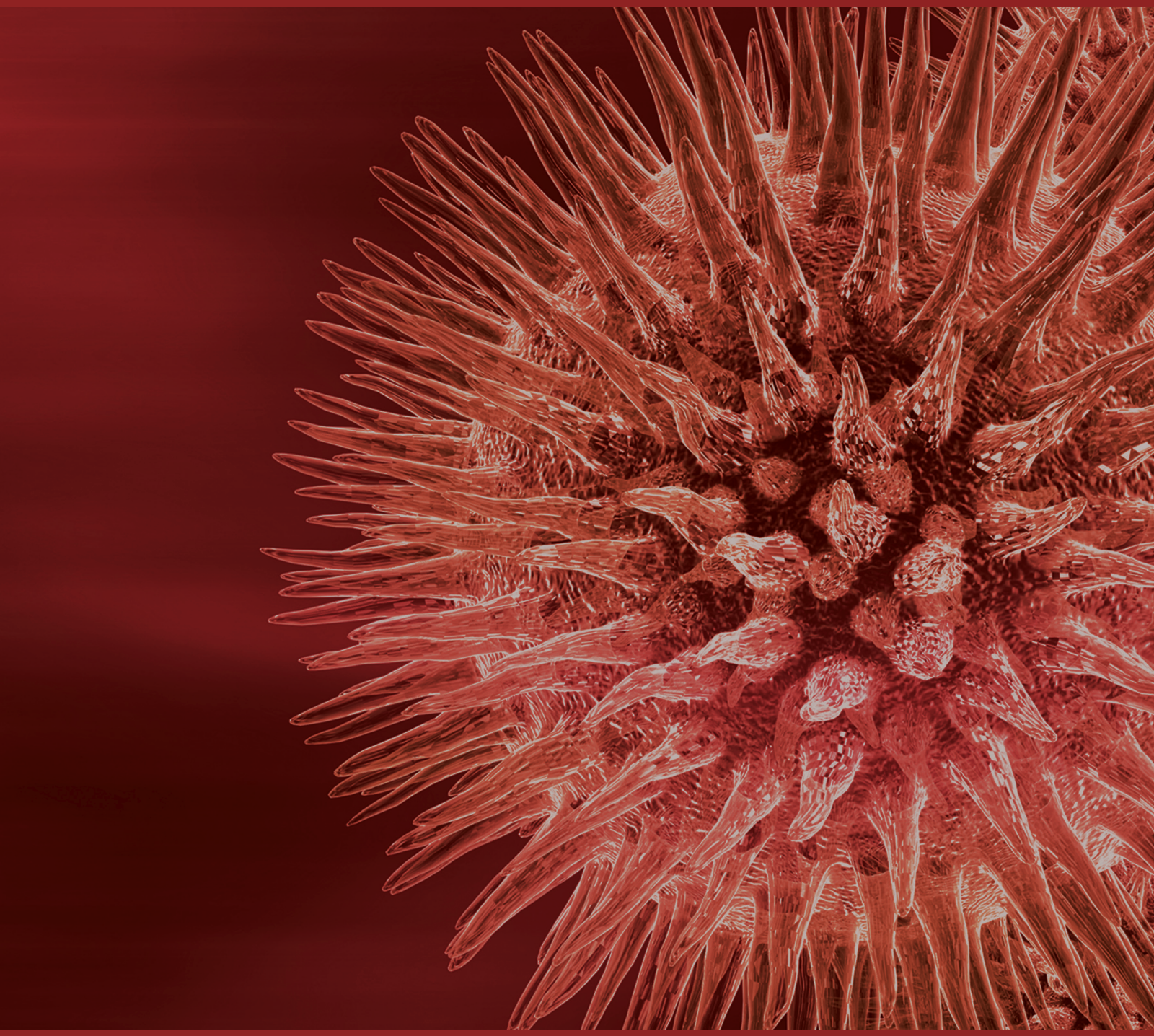
BioMed Research International

Data Mining in Translational Bioinformatics

Guest Editors: Xing-Ming Zhao, Jean X. Gao, and Jose C. Nacher



Data Mining in Translational Bioinformatics

BioMed Research International

Data Mining in Translational Bioinformatics

Guest Editors: Xing-Ming Zhao, Jean X. Gao,
and Jose C. Nacher

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 656519, 2 pages
http://dx.doi.org/10.1155/2014/656519

Editorial

Data Mining in Translational Bioinformatics

Xing-Ming Zhao,[1] Jean X. Gao,[2] and Jose C. Nacher[3]

[1] *School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China*
[2] *Department of Computer Science & Engineering, University of Texas, Arlington, TX 76019, USA*
[3] *Department of Information Science, Faculty of Science, Toho University, Chiba 274-8510, Japan*

Correspondence should be addressed to Xing-Ming Zhao; xm_zhao@tongji.edu.cn

Received 28 May 2014; Accepted 28 May 2014; Published 12 June 2014

Translational bioinformatics is an emerging field that aims to exploit various kinds of biological data for useful knowledge to be translated into clinical practice. However, the flooding of the huge amount of omics data makes it a big challenge to analyze and to interpret these data. Therefore, it is highly demanded to develop new efficient computational methodologies, especially data mining approaches, for translational bioinformatics. Under these circumstances, this special issue aims to present the recent progress on data mining techniques that have been developed for handling the huge amount of biological data arising in translational bioinformatics field.

In data mining, one of the most important problems is how to represent the data so that the computational approaches could handle these data appropriately. In this special issue, B. Gan et al. utilized the latent low-rank representation to extract useful signals from noisy gene expression data and then classified tumors with sparse representation classifier and obtained promising results on benchmark datasets. C. Zhao et al. proposed a new feature representation of facial complexion for diagnosis in traditional Chinese medicine and achieved high recognition accuracy. G. Zhang et al. formulated the skin biopsy image annotation as a multi-instance multilabel (MLML) problem and automatically annotated the skin biopsy images with a sparse Bayesian MLML algorithm based on region structures and texture features. Except for feature extraction, feature selection is also very important in data mining. Z. Ji et al. proposed a particle swarm optimization-based feature selection approach to predict syndromes for hepatocellular carcinoma and improved diagnosis accuracy. With the accumulation of various data in translational bioinformatics, it is becoming a challenging task for traditional intelligent approaches to handle and interpret these data; S. Li et al. presented a survey on the recent progress about the hybrid intelligences and their applications in bioinformatics, where the hybrid intelligence is more powerful and robust compared with traditional intelligent approaches.

The rapid accumulation of various kinds of biological data requires more powerful statistical approaches to extract useful signals from the huge amount of noisy data. L. Sun et al. built a new pipeline to investigate the DNA methylation profiles in male and female nonagenarians/centenarians and identified some differentially methylated probes between male and female nonagenarians/centenarians, which provide insights into the mechanism of longevity gender gap of human beings. Z. Teng et al. developed a new algorithm to predict protein function based on weighted mapping of domains and GO terms, which outperforms other popular approaches on benchmark datasets. J.-L. Huang et al. presented an online cross-species comparative system to identify conserved and exclusive simple sequence repeats within model species, which can facilitate both evolutionary studies and understanding of gene functions. L. Guo et al. proposed a new approach to identify microRNAs (miRNAs) associated with breast cancer and found that miRNA gene clusters demonstrate consistent deregulation patterns despite their different expression levels, which may provide insights into the regulatory roles of miRNAs in tumors.

Recently, network biology is becoming a promising research field by organizing different kinds of data into

a network representation. T. Jacquemin et al. proposed a new approach to identify disease associated protein complexes based on a heterogeneous network that consists of a disease similarity network and a tissue-specific protein-protein interactions network and successfully found disease associated complexes. X. Li et al. proposed a new pipeline to detect symptom-gene associations by integrating multiple data sources and found some potential disease genes. It is known that DNA mutations will affect gene expression. However, it is difficult to know which mutations will affect the gene expression and how the genes are regulated within the biological system. D. Kim et al. developed a novel approach that can both identify the Quantitative Trait Loci and infer the gene regulation network and successfully identified the genes associated with psychiatric disorder. R. Zhang et al. presented a new approach to identify the pathways linking TGF β to ovarian carcinoma immunoreactive antigen-like protein 2 (OCIAD2) by exploring the pathway bridge, and the resultant pathway explained how TGF β affects the expression of OCIAD2 in cancer microenvironment.

Xing-Ming Zhao
Jean X. Gao
Jose C. Nacher

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 629697, 9 pages
http://dx.doi.org/10.1155/2014/629697

Research Article

Inference of SNP-Gene Regulatory Networks by Integrating Gene Expressions and Genetic Perturbations

Dong-Chul Kim,[1] Jiao Wang,[2] Chunyu Liu,[3] and Jean Gao[1]

[1] *Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA*
[2] *Beijing Genomics Institution at Wuhan, Wuhan 430075, China*
[3] *Department of Psychiatry, University of Illinois at Chicago, Chicago, IL 66012, USA*

Correspondence should be addressed to Jean Gao; gao@uta.edu

Received 28 January 2014; Accepted 9 May 2014; Published 9 June 2014

Academic Editor: Xing-Ming Zhao

In order to elucidate the overall relationships between gene expressions and genetic perturbations, we propose a network inference method to infer gene regulatory network where single nucleotide polymorphism (SNP) is involved as a regulator of genes. In the most of the network inferences named as SNP-gene regulatory network (SGRN) inference, pairs of SNP-gene are given by separately performing expression quantitative trait loci (eQTL) mappings. In this paper, we propose a SGRN inference method without predefined eQTL information assuming a gene is regulated by a single SNP at most. To evaluate the performance, the proposed method was applied to random data generated from synthetic networks and parameters. There are three main contributions. First, the proposed method provides both the gene regulatory inference and the eQTL identification. Second, the experimental results demonstrated that integration of multiple methods can produce competitive performances. Lastly, the proposed method was also applied to psychiatric disorder data in order to explore how the method works with real data.

1. Introduction

In order to understand more accurate causal relationships between a complex disease and genetic variations, we need to consider how the genotypic perturbations affect expression phenotypes that are potentially associated with a target disease. In other words, it is more crucial to look at the overall mechanisms considering a series of three factors, which include genetic variations, altering gene regulations, and caused diseases rather than partial mappings between them. Therefore it is important to evaluate how genetic perturbations affect genes on regulatory networks that are associated with a target disease phenotype. In practice, when biological networks are inferred with high throughput data, we have to consider not only the relationships among genes but also how genetic factors such as single nucleotide polymorphism (SNP) and copy number variation (CNV) can affect genes in gene regulatory network (GRN). Over the last decade, research for mapping genotype to expression phenotype or disease phenotype such as expression quantitative trait loci (eQTL) study and genome wide association study have been actively performed [1]. However, we are now required to do a network-based analysis with genotype data and gene expression because it is more effective in discovering underlying biological process from genotype to phenotype. In doing so, the analysis of SNP-gene regulatory networks (SGRN) will provide more definite relationships of genotypic causes and phenotypic effects so that it will facilitate prognosis and drug designs for therapies.

In this paper we propose a SGRN inference method. In order to identify regulatory interactions among genes, quite a number of network inference methods have been developed by using gene expression data such as gene microarray. Those methods can be generally classified into different theoretical categories: Boolean networks [2, 3], mutual information [4, 5], Bayesian networks (BN) [6, 7], and regression [8, 9]. As each method has its own advantages and limitations under different assumptions and network models such as acyclic or cyclic network and directed or undirected network, there should be trade-offs in inferences given different target network structure and applications [10]. For example, the MI-based approach is very simple and fast so that it can

build a large scale network (e.g., genome wide scale) but it cannot estimate direction of edges. It produces worse performance than other approaches in detecting linear cascading structures [10]. The BN-based inference is limited to imply only acyclic network with high computational cost while the regression-based approach supports both directed and cyclic network, which are assumed in SGRN. In addition to directed network model, it should be considered that SGRN is different from conventional GRN inference. In SGRN inference, a gene can be regulated by SNPs as well as other genes, but SNPs are assumed to not be regulated by other SNPs. That is, a SNP cannot be a child node in the network.

Recently, a number of approaches have been suggested to infer SGRNs integrating genetic variation and gene expression data. Kim et al. [11] considered genetic perturbations, gene expression, and disease phenotypes together to find the causal genes to a disease. The electric circuit approach and heuristic search were used to infer SGRN where causal genes are mapped to SNP in the preliminary step before network inference. Keurentjes et al. [12] built a SNP-gene network associated with a particular phenotype, but this method also performed eQTL mapping (SNP-gene) to define the candidate regulator genes before genetic network construction. In addition, Kim and Xing [13] used lasso regression considering the case that a SNP is weakly associated with highly correlated multiple traits rather than a single trait. Chen et al. [14] focused on identifying which pathway among those already known pathways was more likely to be affected by changes of genotype and gene expression rather than inferring a new pathway. The related works we especially noted are the methods that are based on structural equation modeling (SEM) [15–18]. SEM allows us to not only incorporate eQTL information to gene expression in a single model but also identify eQTL simultaneously. However, Logsdon and Mezey [17] assumed that every gene has at least one eQTL, and eQTL mapping was performed by preprocessing but not in a network inference step. Cai et al. [18] introduced sparsity-aware maximum likelihood (SML), which can be potentially extended for eQTL identification. However, SNP-gene pairs were still given in evaluations and implementations of the SML algorithm.

In this paper, we proposed a novel method to infer SGRN where both eQTL identification and SGRN inference are performed simultaneously given a set of gene expression and genotype data without assuming eQTLs are known. The proposed method is based on SEM and multiple steps of edge filtering such as elastic net regression and iterative adaptive lasso. Basically SEM is a regression-based model which is likely to select as many variables causing an overfitting, so the sparsity is enforced by lasso (l_1-regularized least square estimation) considering the sparsity of biological network. Initial weights of edges are estimated by ridge regression [19] and elastic net regression [20], and then the second step is to identify final eQTLs from candidate SNPs selected in the first steps. In the last step, the final network is constructed by iterative adaptive lasso. The first two steps are to fix SNPs before selecting genes. In the third step, edges are selected by iteratively giving more penalties to the edge whose weight is relatively low until network structure is converged.

To evaluate the method, we explore the performance with a simulated data set, that is, generated from random networks with different number of samples and nodes and expected number of edges per node. The result shows that the method can achieve a high detection rate of true edges with low false discovery rate without eQTL information. In addition, to explore the performance in real expression phenotype and SNP data, the method was applied to the psychiatric disorder data. After genes and SNPs were selected from related Genome-Wide Association Study (GWAS), it was tested how the method identify true positive edges between genes and SNPs without eQTL information.

2. Method

2.1. Problem Definitions. We define the problem and notations here. Let $Y \in \mathbb{R}^{M_g \times N}$ denote the matrix of gene expression levels of M_g genes and N samples where a row vector $\mathbf{y}_i = \{y_{i1}, \ldots, y_{iN}\}$ is observed expression level of ith gene. X is $M_s \times N$ matrix to denote genotypes of individuals, where $x_{ij} \in \{1, 2, 3\}$ represents the number of minor alleles of ith SNP of jth sample as an element of matrix X supposing that the number of minor alleles should be zero, one, or two in real data. So, x_{ij} represents a relative quantity of minor alleles of samples. As a gene can be regulated by other genes and genetic variations (SNPs), we define SEM as

$$\mathbf{y}_i = \mathbf{b}_i Y + \mathbf{f}_i X + \mu_i + \varepsilon_i, \tag{1}$$

where $\mathbf{b}_i$ denotes ith row vector of square matrix $B \in \mathbb{R}^{M_g \times M_g}$; $\mathbf{f}_i$ denotes ith row vector of square matrix $F \in \mathbb{R}^{M_g \times M_s}$; μ_i is a model bias; and ε_i is a residual modeled as zero-mean Gaussian with a variance σ^2. As we assume there is no self-regulation (self-loop edge), $b_{ii} = 0, \forall i = 1, \ldots, M_g$, where b_{ii} denotes ith element of $\mathbf{b}_i$. The parameters of $\mathbf{b}_i$ and $\mathbf{f}_i$ decide the network structure defining the weight of regulation from every possible gene and SNP to a target gene i. For example, if there is no regulation relationship (directed edge) from jth gene to ith gene, b_{ij} is set to zero. Similarly f_{ij} has nonzero value as a weight of regulation from jth SNP to ith gene if jth SNP is identified as an eQTL for ith gene. It is assumed that each gene has at least one eQTL but it is unknown which SNP among a given set of SNPs is an eQTL for a target gene. Our goal in this model is to find B and F that best fit to observed gene expression and genotype data. To make the problem simpler, we remove μ_i from (1) by applying mean centering for row vectors $\mathbf{y}_i$ and $\mathbf{x}_i$ to have zero mean. The goal is to find $\mathbf{b}_i$ and $\mathbf{f}_i$ that minimize a residual ε_i, so (1) can be expressed in a least square minimization problem as

$$\arg\min_{\mathbf{b}_i, \mathbf{f}_i} \left\| \mathbf{y}_i - \mathbf{b}_i Y - \mathbf{f}_i X \right\|_2^2. \tag{2}$$

However, regression tends to select as many genes and SNPs as possible to explain the expression level of target gene i. To avoid the overfitting, sparse regression methods such as ridge regression, elastic net, and lasso are used.

2.2. The Algorithm. The method we propose is based on l_1-regularized linear regression known as lasso [21] that yields

(1) **procedure** ELASTIC($Y, X, \widehat{\lambda}_1, \widehat{\lambda}_2, i, \varepsilon$) ▷ $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ are optimal parameters estimated by cross validation
(2) **while** err > ε **do**
(3) $\mathbf{b}_i^{\text{old}} = \mathbf{b}_i, \mathbf{f}_i^{\text{old}} = \mathbf{f}_i$
(4) **for** $j \leftarrow 1, M_s$ **do**
(5) Update f_{ij} via (12)
(6) **end for**
(7) Update $\mathbf{b}_i$ via (5)
(8) err = $\left\|\mathbf{b}_i^{\text{old}} - \mathbf{b}_i\right\|_2 + \left\|\mathbf{f}_i^{\text{old}} - \mathbf{f}_i\right\|_2$
(9) **end while**
(10) **return** $\mathbf{b}_i$ and $\mathbf{f}_i$
(11) **end procedure**

ALGORITHM 1: Optimization for *elastic net* in Step 1-2.

a sparsity of variable selection. The algorithm consists of 3 steps, (i) *elastic net*, (ii) *lasso*, and (iii) *iterative adaptive lasso*. The first two steps are to decide F where SNPs are selected but their coefficients can be changed in the third step. Then, B is finalized by iterative adaptive lasso in the last step.

2.2.1. Ridge Regression (Step 1-1). In ridge regression, the coefficient values of irrelevant SNPs and genes to a target gene shrink to zero (but not exactly zero) while those of eQTLs and regulator genes of a target gene tend to be higher. Ridge regression of (2) is defined as

$$\arg\min_{\mathbf{b}_i,\mathbf{f}_i} \left\|\mathbf{y}_i - \mathbf{b}_i Y - \mathbf{f}_i X\right\|_2^2 + \lambda_1 \left\|\mathbf{b}_i\right\|_2^2 + \lambda_2 \left\|\mathbf{f}_i\right\|_2^2. \quad (3)$$

Given penalty weights, λ_1 and λ_2, the optimal $\mathbf{b}_i$ and $\mathbf{f}_i$ can be obtained by closed form solution given by

$$\mathbf{f}_i = \left(\mathbf{y}_i - \mathbf{b}_i Y\right) X^T \left(XX^T + \lambda_2 I\right)^{-1}, \quad (4)$$

$$\mathbf{b}_i = \left(\mathbf{y}_i - \mathbf{f}_i X\right) Y^T \left(YY^T + \lambda_1 I\right)^{-1}. \quad (5)$$

Replacing (5) for $\mathbf{b}_i$ in (4) yields

$$\mathbf{f}_i = \mathbf{y}_i S_1 \left(XS_1 + \lambda_2 I\right)^{-1}, \quad (6)$$

where

$$S_1 = X^T - Y^T \left(YY^T + \lambda_1 I\right)^{-1} YX^T. \quad (7)$$

After calculating $\mathbf{f}_i$ first in (6) then (5) can be solved. In this manner, matrices B and F are estimated by computing each $\mathbf{b}_i$ and $\mathbf{f}_i$, $i = 1, \ldots, M_g$. Parameters λ_1 and λ_2 that decide the degree of sparsity of B and F are determined by K-fold cross-validation. K is set to 5 in our experiments.

2.2.2. Elastic Net (Step 1-2). Note that zero weighted coefficient cannot be recovered back to nonzero in adaptive lasso of Step 3. Therefore, in order to carefully keep only SNPs that are more likely to be true eQTLs in $\mathbf{f}_i$, we give l_1-norm penalty to only $\mathbf{f}_i$ but not $\mathbf{b}_i$ using elastic net defined as

$$\arg\min_{\mathbf{b}_i,\mathbf{f}_i} \left\|\mathbf{y}_i - \mathbf{b}_i Y - \mathbf{f}_i X\right\|_2^2 + \lambda_1 \left\|\mathbf{b}_i\right\|_2^2 + \lambda_2 \left\|\mathbf{f}_i\right\|_1. \quad (8)$$

As the objective function is convex, which guarantees a convergence, f_{ij} can be optimized by using coordinate descent iteration given parameters, λ_1 and λ_2. To find the optimal $\mathbf{f}_i$, the derivative of (8) with respect to f_{ij} is considered as follows:

$$\mathbf{f}_i XX_j^T - \mathbf{y}_i X_j^T + \mathbf{b}_i YX_j^T + \lambda_2 \partial_{f_{ij}} \left\|\mathbf{f}_i\right\|_1. \quad (9)$$

Since the derivative of (8) with respect to $\mathbf{b}_i$ is the same as (5), $\mathbf{b}_i$ in (9) is substituted with (5), and then (9) is simplified to

$$\left(\mathbf{f}_{i(-j)} X_{(-j)} - \mathbf{y}_i\right) S_2 + f_{ij} \mathbf{x}_j S_2 - \lambda_2 \partial_{f_{ij}} \left\|\mathbf{f}_i\right\|_1, \quad (10)$$

where

$$S_2 = \left(Y^T \left(YY^T + \lambda_1 I\right)^{-1} Y - I\right) \mathbf{x}_j^T; \quad (11)$$

$\mathbf{f}_{i(-j)}$ indicates row vector $\mathbf{f}_i$ whose jth element is removed, $X_{(-j)}$ denotes matrix X whose jth row is removed, and $\mathbf{x}_j$ is jth row vector of X. After defining $C_j = (\mathbf{f}_{i(-j)} X_{(-j)} - \mathbf{y}_i) S_2$ and $a_j = \mathbf{x}_j S_2$ in (10), the update rule in the coordinate descent algorithm is written as

$$f_{ij} = \begin{cases} \dfrac{\left(-C_j - \lambda_2\right)}{a_j} & \text{if } C_j < -\lambda_2, \\ 0 & \text{if } C_j \le \left|\lambda_2\right|, \\ \dfrac{\left(-C_j + \lambda_2\right)}{a_j} & \text{if } C_j > \lambda_2. \end{cases} \quad (12)$$

Algorithm 1 describes the procedures to solve (8) in Step 2. If f_{ij} is nonzero, jth SNP is a candidate eQTL for ith gene.

2.2.3. Lasso (Step 2). In order to finalize a SNP (a single nonzero f_{ij} of $\mathbf{f}_i$) for each gene i, we apply lasso to combined matrix of Y and X as follows:

$$\left\|\mathbf{y}_i - \mathbf{h}_i Z\right\|_2^2 + \lambda \left\|\mathbf{h}_i\right\|_1, \quad (13)$$

where

$$Z^T = \left[Y_{(-i)}^T, X_{(-k_i^*)}^T\right]. \quad (14)$$

k_i^* denotes indices of low vectors where $f_{ij} = 0$, $j \in k_i^*$. So, $X_{(-k_i^*)}$ is a matrix X whose k_i^* rows are removed. If the number of rows of $X_{(-k_i^*)}$ is greater than predefined heuristic number N_k (i.e., 5 in our experiments), only top N_k highest f_{ij} of absolute values of $\mathbf{f}_i$ but not all nonzero f_{ij} are selected for $X_{(-k_i^*)}$. In Step 2, we iteratively estimate $\mathbf{h}_i$, decreasing λ from a high value that lets $\mathbf{h}_i$ have a zero vector. Regardless of elements of $\mathbf{h}_i$ for $Y_{(-i)}$, we note only which element of $\mathbf{h}_i$ for $X_{(-k_i^*)}$ has a nonzero value first assuming that the corresponding candidate SNP to h_{ij} is more likely to regulate a target gene i if h_{ij} for a row vector of $X_{(-k_i^*)}$ has nonzero value earlier than other elements of $\mathbf{h}_i$ during λ decreases.

2.2.4. Adaptive Lasso (Subroutine of Step 3). Adaptive lasso is defined as

$$\arg\min_{\mathbf{b}_i,\mathbf{f}_i} \left\|\mathbf{y}_i - \mathbf{b}_i Y - \mathbf{f}_i X\right\|_2^2 + \lambda_1 \left\|\mathbf{b}_i\right\|_{1,\mathbf{w}_i^b} + \lambda_2 \left\|\mathbf{f}_i\right\|_{1,\mathbf{w}_i^f}, \quad (15)$$

where

$$\left\|\mathbf{b}_i\right\|_{1,\mathbf{w}_i^b} = \sum_j^N \left|b_{ij} \cdot w_{ij}^b\right|, \qquad \left\|\mathbf{f}_i\right\|_{1,\mathbf{w}_i^f} = \sum_j^N \left|f_{ij} \cdot w_{ij}^f\right|. \quad (16)$$

In (16), penalty weights, vectors $\mathbf{w}_i^b$ and $\mathbf{w}_i^f$, are defined as

$$w_{ij}^b = \left(\widehat{b}_{ij}\right)^{-\alpha}, \quad w_{ij}^f = \left(\widehat{f}_{ij}\right)^{-\beta}, \quad \forall j = \left\{1, \ldots, M_g\right\}, \quad (17)$$

where $\widehat{b}_{ij}$ and $\widehat{f}_{ij}$ are estimated in Step 2 that yields a sparsity to $\mathbf{f}_i$ but not $\mathbf{b}_i$. Zero coefficient of $\widehat{\mathbf{f}}_i$ in Step 2 is not considered as an eQTL for gene i. So, zero $\widehat{f}_{ij}$ yields zero w_{ij}^f in (17), and then if w_{ij}^f is zero, f_{ij} will never have nonzero value in adaptive lasso of Step 3 (16). The parameters α and β decide how much previous estimation such as $\widehat{b}_{ij}$ or $\widehat{f}_{ij}$ is reflected to next estimation of b_{ij} or f_{ij}. Therefore, f_{ij} that has smaller penalty weight w_{ij}^f is more likely to have nonzero value. In addition, we consider a special case that α and β are set to zero supposing that (i) we do not give a penalty weight to b_{ij} or f_{ij} by setting w_{ij}^b or w_{ij}^f to 1 if $\widehat{b}_{ij}$ or $\widehat{f}_{ij}$ is nonzero and (ii) we do not estimate elements of $\mathbf{b}_i$ or $\mathbf{f}_i$ by setting w_{ij}^b or w_{ij}^f to infinity if $\widehat{b}_{ij}$ or $\widehat{f}_{ij}$ is zero. The solution is similar to Step 2 in which either $\mathbf{b}_i$ or $\mathbf{f}_i$ is optimized by coordinate descent algorithm but it is applied to solve both $\mathbf{b}_i$ and $\mathbf{f}_i$ in Step 3. Derivative of (15) with respect to b_{ij} yields

$$\begin{aligned} &\mathbf{b}_i Y \mathbf{y}_j^T - \mathbf{y}_i \mathbf{y}_j^T + \mathbf{f}_i X \mathbf{y}_j^T + \lambda_1 \partial_{\mathbf{b}_{ij}} \left\|\mathbf{b}_i\right\|_{1,\mathbf{w}_i^b} \\ &= b_{ij} \mathbf{y}_j \mathbf{y}_j^T + \left(\mathbf{b}_{i(-j)} Y_{(-j)} - \mathbf{y}_i + \mathbf{f}_i X\right) \mathbf{y}_j^T + \lambda_1 \partial_{b_{ij}} \left\|\mathbf{b}_i\right\|_{1,\mathbf{w}_i^b}, \end{aligned} \quad (18)$$

where $\mathbf{b}_{i(-j)}$ indicates row vector $\mathbf{b}_i$ whose jth element is removed and $Y_{(-j)}$ denotes matrix Y whose jth row is removed. After setting $C_j^b = (\mathbf{b}_{i(-j)} Y_{(-j)} - \mathbf{y}_i + \mathbf{f}_i X)\mathbf{y}_j^T$ and $a_j^b = \mathbf{y}_j \mathbf{y}_j^T$, the update rule for b_{ij} is as follows:

$$b_{ij} = \begin{cases} \dfrac{\left(-C_j^b - w_{ij}^b \cdot \lambda_1\right)}{a_j^b} & \text{if } C_j^b < -w_{ij}^b \cdot \lambda_1, \\ 0 & \text{if } C_j^b \le \left|w_{ij}^b \cdot \lambda_1\right|, \\ \dfrac{\left(-C_j^b + w_{ij}^b \cdot \lambda_1\right)}{a_j^b} & \text{if } C_j^b > w_{ij}^b \cdot \lambda_1. \end{cases} \quad (19)$$

We can also estimate f_{ij} in similar way. After defining $C_j^f = (\mathbf{f}_{i(-j)} X_{(-j)} - \mathbf{y}_i + \mathbf{b}_i Y)\mathbf{x}_j^T$ and $a_j^f = \mathbf{x}_j \mathbf{x}_j^T$, the update rule for f_{ij} is given as

$$f_{ij} = \begin{cases} \dfrac{\left(-C_j^f - w_{ij}^f \cdot \lambda_2\right)}{a_j^f} & \text{if } C_j^f < -w_{ij}^f \cdot \lambda_2, \\ 0 & \text{if } C_j^f \le \left|w_{ij}^f \cdot \lambda_2\right|, \\ \dfrac{\left(-C_j^f + w_{ij}^f \cdot \lambda_2\right)}{a_j^f} & \text{if } C_j^f > w_{ij}^f \cdot \lambda_2. \end{cases} \quad (20)$$

When $\mathbf{b}_i$ and $\mathbf{f}_i$ are updated, updated single element b_{ij} or f_{ij} immediately affects updating the next elements. In addition, updating order of elements can be changed since convex objective function is converged in any order of elements to update. Algorithm 2 shows the optimization procedure of adaptive lasso.

2.2.5. Iterative Adaptive Lasso (Step 3). Even if $\mathbf{b}_i$ and $\mathbf{f}_i$ are estimated in Steps 1 and 2, there should be still many false positive edges yet. The primary goal of Steps 1 and 2 is to carefully get rid of only edges that are more unlikely to be true positive edges. So, instead of simply applying adaptive lasso, we developed iterative adaptive lasso to improve the performance of naive adaptive lasso. The motivation of iterative adaptive lasso is that the coefficient value of the variable considerably depends on the value of α and β which are fixed to 1 and 0.5 in [17, 18], respectively. In iterative adaptive lasso, adaptive lasso is iteratively applied incrementally changing α and β until there is no more change in the total number of selected edges of B and F so that more coefficients of irrelevant variables can be shrunk to zero.

Algorithm 3 presents a detailed procedure of iterative adaptive lasso. $\widehat{B}$ and $\widehat{F}$ estimated in Step 2 are used as arguments. On line 2, B and F are initialized by ridge regression. Λ_1^R is a vector of optimal parameters of λ_1 for B^R in (3) but there is no penalty to F^R ($\Lambda_2^R = 0$). For F^R we estimate only nonzero elements of $\widehat{F}$ that is estimated in Step 2. Again, B and F are initialized by adaptive lasso in order that elements of B are updated by weights of B^R. In this initialization, b_{ij} that has a small value can shrink to zero. Based on updated B and F, Λ_1 (a vector of λ_1 for B on line 9) is estimated again by cross-validation of adaptive lasso before line 6 starts. Initially the second *while* loop updates B until no change in $N_e(B)$. Once the second *while* loop is terminated,

(1) **procedure** ADAPTIVE LASSO($Y, X, \widehat{\lambda}_1, \widehat{\lambda}_2, i, \alpha, \beta, \widehat{\mathbf{b}}_i, \widehat{\mathbf{f}}_i$) ▷ $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ are optimal parameters preliminary estimated by cross validation
(2) Compute $\mathbf{w}_i^b$ and $\mathbf{w}_i^f$ ($w_{ij}^b = (\widehat{b}_{ij})^{-\alpha}, w_{ij}^f = (\widehat{f}_{ij})^{-\beta}$)
(3) **while** err > ε **do**
(4) $\mathbf{b}_i^{\text{old}} = \mathbf{b}_i, \mathbf{f}_i^{\text{old}} = \mathbf{f}_i$
(5) **for** $j \leftarrow 1, M_g$ **do**
(6) Update b_{ij} via (19)
(7) **end for**
(8) **for** $j \leftarrow 1, M_s$ **do**
(9) Update f_{ij} via (20)
(10) **end for**
(11) err = $\left\|\mathbf{b}_i^{\text{old}} - \mathbf{b}_i\right\|_2 + \left\|\mathbf{f}_i^{\text{old}} - \mathbf{f}_i\right\|_2$
(12) **end while**
(13) **return** $\mathbf{b}_i$ and $\mathbf{f}_i$
(14) **end procedure**

ALGORITHM 2: Optimization for adaptive lasso as a subroutine of Step 3.

(1) **procedure** ITERATIVE ADAPTIVE LASSO($Y, X, \widehat{B}, \widehat{F}$) ▷ Ne($B$) denote the number of non-zero elements in B and F
(2) $[B^R, F^R] = Ridge(Y, X, \widehat{\Lambda}_1^R, \widehat{F})$ in (3)
(3) $\alpha = 1, \beta = 1$
(4) **for** $i \leftarrow 1, M_g$ **do**
(5) $[\mathbf{b}_i, \mathbf{f}_i] = AdaptiveLasso(Y, X, \lambda_1 = 0.001, \lambda_2 = 0, i, \alpha, \beta, \mathbf{b}_i^R, \mathbf{f}_i^R)$ in (15)
(6) **end for**
(7) **while** Ne(B) are decreased by increased α **do**
(8) **while** Ne(B) are decreased **do**
(9) **for** $i \leftarrow 1, M_g$ **do**
(10) $[\mathbf{b}_i, \mathbf{f}_i] = AdaptiveLasso(Y, X, \widehat{\lambda}_1, \lambda_2 = 0, i, \alpha, \beta, \mathbf{b}_i, \mathbf{f}_i)$ in (15)
(11) **end for**
(12) **end while**
(13) $\alpha = \alpha + 1$
(14) **end while**
(15) **return** B and F
(16) **end procedure**

ALGORITHM 3: Iterative adaptive lasso in Step 3.

α is increased, and then the second loop is performed again. If the second *while* loop is terminated without any change of $N_e(B)$, the first *while* loop is terminated.

3. Results

3.1. Simulation Studies. To evaluate the proposed method, we first perform simulations based on randomly generated acyclic networks. The simulation settings are similar to those in [17, 18]. M denotes the number of genes and SNPs and is set to 10, 20, and 30. $M \times N$ matrix B is initialized to zero matrix where N is a sample size; then elements of B are randomly selected as directed edges. The selected b_{ij} has random coefficient value uniformly distributed over 0.5~1 or −0.5~ −1. Since we consider a single eQTL per gene ($E_s = 1$), a single element (f_{ii}) is selected from each row vector ($\mathbf{f}_i$). So, F is a diagonal matrix. x_{ij} is randomly set as 1, 2, or 3 with the probabilities 0.25, 0.5, and 0.25, respectively. Y is generated by calculating $Y = (I - B)^{-1}(FX + E)$, where E_{ij} is generated from Gaussian distribution with zero mean and variance 0.01. The number of samples for each network size is $N = 100, 200, 300, 400,$ and 500. The number of edges per gene on average is set to $E_g = 1, 2,$ and 3. Given data Y and X, performances of predicting B and F are evaluated by comparing true network and inferred network.

Figure 1 displays the examples of networks, where SNP nodes are excluded. For the evaluation, true positive (TP), false positive (FP), true negative (TN), and false negative (FN) edges are counted to measure the accuracy criteria such as true positive rate (TPR) and false discovery rate (FDR) that are defined as

(i) TPR = TP/(TP + FN),

(ii) FDR = FP/(TP + FP).

In order to evaluate our method, IAL is compared to SML [18]. As SML infers only B with known nonzero element

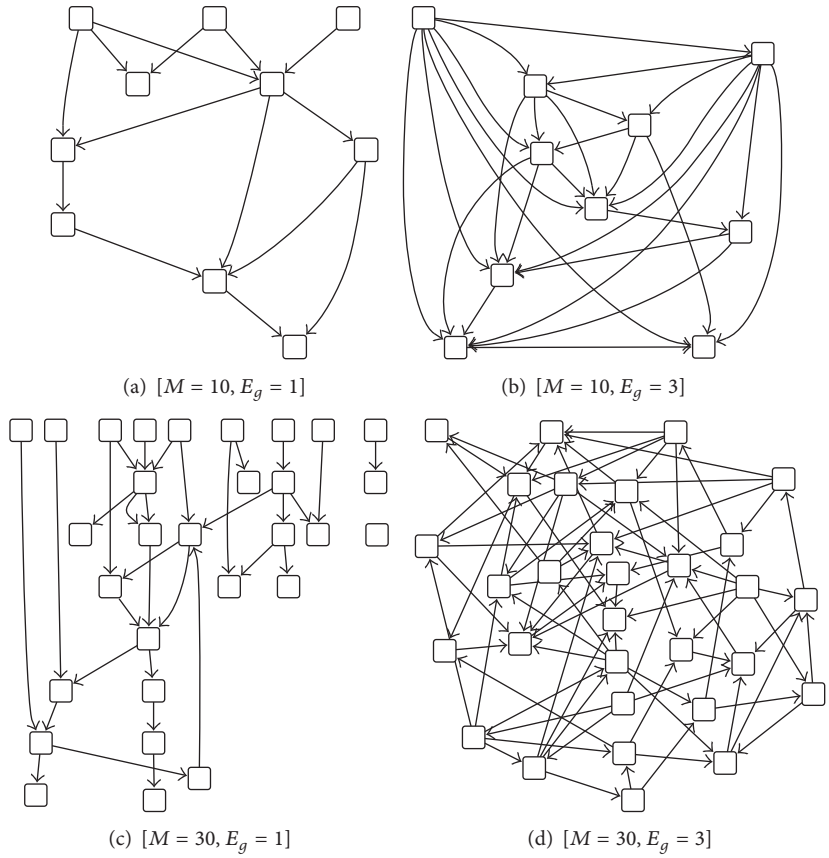


Figure 1: Example of simulated networks with different parameter settings. M and E_g indicate the number of genes and expected number of edges per node, respectively.

Table 1: TPR and FDR of SML, IAL1, and IAL2.

N	M	TPR			FDR		
		SML	IAL1	IAL2	SML	IAL1	IAL2
100	10	0.9888	1.0000	0.9742	0.0860	0	0.0104
	20	0.9980	1.0000	0.9448	0.0503	0	0.0292
	30	0.9951	1.0000	0.8936	0.0364	0	0.0754
500	10	0.9967	1.0000	1.0000	0.0704	0	0
	20	0.9850	1.0000	0.9436	0.0400	0	0.0369
	30	1.0000	1.0000	0.9128	0.0016	0	0.0562

Expected number of edges per node is $E_g = 2$ and 10 replicates of random network are used. N and M indicate the number of samples and genes, respectively.

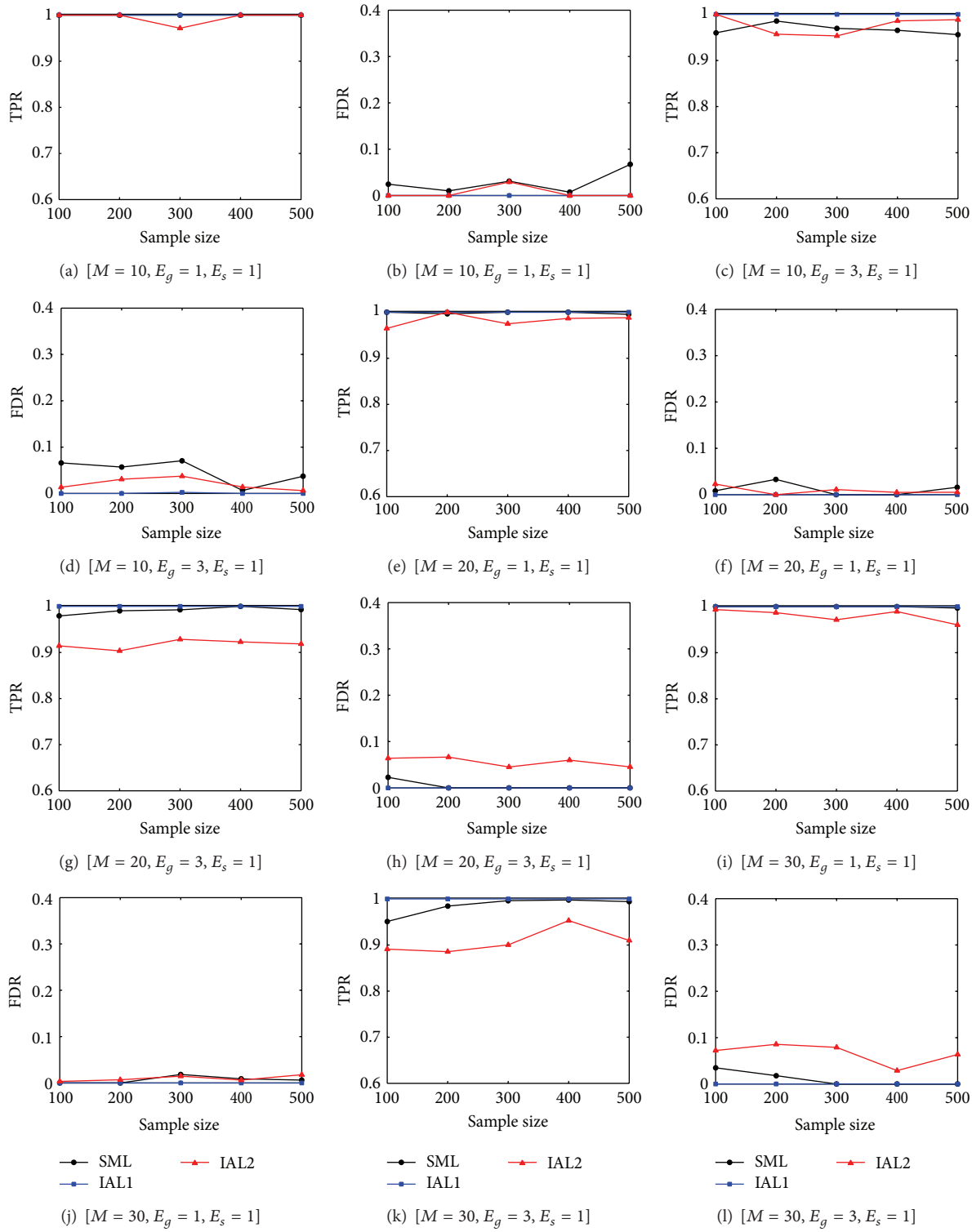
indices of F, we consider two versions of IAL, IAL without eQTL information and IAL with eQTL information, where Steps 1 and 2 are skipped and only Step 3 is performed with nonzero element index of $\mathbf{f}_i$. SML is tested by using the code the author implemented in [18]. The abbreviations of algorithms to compare in Figure 2 and Table 1 are listed below:

(i) SML: sparsity-aware maximum likelihood algorithm with eQTL information [18],

(ii) IAL1: IAL with eQTL information,

(iii) IAL2: IAL without eQTL information.

Ten replicate simulations are performed and each simulation has a different topology. The results of the different settings (M and E_g) are displayed in Figure 2. It is shown that IAL1 is superior to SML in all data sets regardless of sample size. We also note that TPR of IAL2 is higher than 0.9 and FDR is less than 0.1 on average in any sample size. It validates that the proposed IAL works very effectively when eQTL is known. In addition, the performance of IAL1 is consistent in different sample sizes while the performance of SML tends to be decreased with small sample size and complicated network ($E_g = 3$). In network inference, it is known that the performance of inference is very sensitive to the network size and density. In the inference of densely connected and large networks, the computational cost will exponentially increase and the FDR may increase because there are more possible variables that may explain a target node better than true regulators. IAL1 performed consistently in all three different network sizes while the performance of SML is affected by the network size in dense networks ($E_g = 3$). However, IAL2 shows consistent TPRs and FDRs in all three different network sizes when the network density is normal ($E_g = 1$) while TPR of IAL2 in Figures 2(g) and 2(k) is lower than Figure 2(c) and also FDR increases in



(a) $[M = 10, E_g = 1, E_s = 1]$ (b) $[M = 10, E_g = 1, E_s = 1]$ (c) $[M = 10, E_g = 3, E_s = 1]$

(d) $[M = 10, E_g = 3, E_s = 1]$ (e) $[M = 20, E_g = 1, E_s = 1]$ (f) $[M = 20, E_g = 1, E_s = 1]$

(g) $[M = 20, E_g = 3, E_s = 1]$ (h) $[M = 20, E_g = 3, E_s = 1]$ (i) $[M = 30, E_g = 1, E_s = 1]$

(j) $[M = 30, E_g = 1, E_s = 1]$ (k) $[M = 30, E_g = 3, E_s = 1]$ (l) $[M = 30, E_g = 3, E_s = 1]$

FIGURE 2: True positive rate and false discovery rate under different numbers of edges and nodes.

Table 1 when the network size increases in more dense networks ($E_g = 2$).

The result shows that the performance is better in sparse networks ($E_g = 1$) than dense networks ($E_g = 3$) because a complicated structure is more likely to cause false positive edges because of indirect regulations. For example, TPRs in Figures 2(a), 2(e), and 2(i) are much better than in Figures 2(c), 2(g), and 2(k). Similarly FDR is quite increased with $E_g = 3$ in Figures 2(d), 2(h), and 2(l) compared to the case of $E_g = 1$ in Figures 2(b), 2(f), and 2(j).

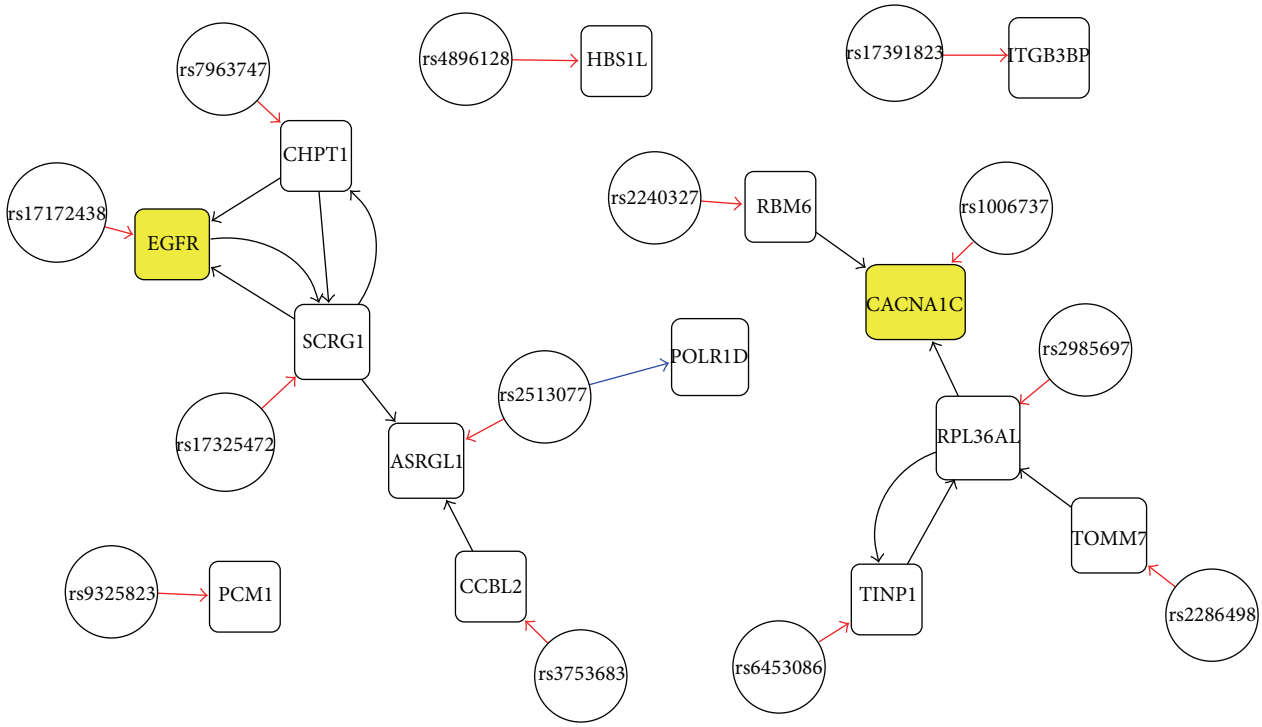


Figure 3: The inferred SGRN with 14 pairs of gene and SNP selected from [22–24].

Overall results imply that the proposed IAL1 works perfectly with known F in any network size and density. It means that the performance of IAL2 is significantly affected by false positive inference of F in steps 1 and 2 because of unknown F. More precisely $\mathbf{b}_i$ without sparsity in step 2 is more likely to have false positive nonzero elements even though a number of candidate elements of $\mathbf{b}_i$ are filtered in step 1. Therefore, the selection of nonzero element of $\mathbf{b}_i$ in IAL2 is the most critical part since IAL1 is able to correctly infer B only if F is given as eQTL information.

3.2. Experiments with Psychiatric Disorder Data. In this section, the proposed method is applied to real gene expression and genotype data for psychiatric disorder. In the application to real data, we explore the performance of GRN inferences and eQTL identifications through the inferred networks. As far as we know, the proposed method is the first solution to provide both GRN inference and eQTL identification. Thus, the performance comparison with other methods was not performed. The psychiatric disorder data consists of gene expression data of 25833 genes and 852963 SNPs for 131 samples, which were measured from human brain. Since we focus on the network inference but not gene selection, the network construction is performed with a predefined set of genes and SNPs that are selected by preliminary test of multiple sets of genes and eQTLs based on related GWAS for psychiatric disorders. The result of SGRN inference is displayed in Figure 3 where two yellow colored genes, EGFR and CACNA1C, are selected from [23, 24] and the rest of two pairs are from [22]. In applying IAL2 to the data, the weights of α and β are set to 0.5 instead of 1. Otherwise, $N_e(\mathbf{f}_i)$ tends to be zero. The reason for this is that gene variables are more correlated with their eQTLs because generally eQTLs are independently selected to other genes. In Figure 3, SNP and gene are distinguished by node shape, and a red edge indicates a correct edge from eQTL to corresponding gene. A blue edge represents false positive eQTL mapping. For eQTL identification, one false positive edge appears and thirteen true positive edges are detected (TPR = 0.9286, FDR = 0.0714).

4. Discussion

The most difficult part in network inference is to identify directions of edges. In the adjacency matrix B, both B_{ij} and B_{ji} could have a high coefficient value. In this case, regression-based methods tend to show better performance than MI-based methods because candidate edges are evaluated together in regression-based methods but each edge is independently evaluated to other edges in MI-based methods. Despite the advantage, the regression-based method needs to be integrated with other methods that can provide different information of structure. Another issue to improve in IAL is the computational cost to estimate two different λs per each row. Intuitively, a searched optimal λ per each row of B and F should provide a better result but it causes a high computation cost. Lastly, we also assumed that a gene has at least a single eQTL given a set of genes and SNPs, but multiple eQTLs should be considered and a gene may not have any eQTL in practice. Thus, the multiple eQTL of a gene is a future work in SGRN inference.

5. Conclusion

In this paper, we proposed a novel network inference method that provides both eQTL identification and network construction of both genes and SNPs. In order to understand gene regulatory mechanisms for a target disease phenotype, the regulatory network inference needs to consider effect of genetic variation and expression phenotype together but not only gene expression data. To achieve the high quality of reliable inference with better TPR and FDR, three different regression skills are integrated. Ridge regression and elastic net are used to remove more likely false positive edges and select eQTL as preliminary steps, and then the finial network is estimated by iterative adaptive lasso removing more false positive edges between genes. Through the experiments with synthetic data, it was demonstrated that IAL1 outperforms SML in SGRN inference and also IAL2 performs eQTL identification effectively. The method was also applied to psychiatric disorder data. Using the genes and eQTLs selected from GWAS of psychiatric disorder, we explored the ability of eQTL identification through inferred SGRN.

Conflict of Interests

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

References

[1] A. C. Nica and E. T. Dermitzakis, "Expression quantitative trait loci: present and future," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1620, 2013.

[2] J. Liang and J. Han, "Stochastic Boolean networks: an efficient approach to modeling gene regulatory networks," *BMC Systems Biology*, vol. 6, article 113, Article ID 20120362, 2012.

[3] B. Vasić, V. Ravanmehr, and A. R. Krishnan, "An information theoretic approach to constructing robust boolean gene regulatory networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 1, pp. 52–65, 2012.

[4] R. de Matos Simoes and F. Emmert-Streib, "Bagging statistical network inference from large-scale gene expression data," *PLoS ONE*, vol. 7, no. 3, Article ID e33624, 2012.

[5] W. Zhao, E. Serpedin, and E. R. Dougherty, "Inferring connectivity of genetic regulatory networks using information-theoretic criteria," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 2, pp. 262–274, 2008.

[6] N. Xuan, M. Chetty, R. Coppel, and P. Wangikar, "Gene regulatory network modeling via global optimization of high-order dynamic Bayesian network," *BMC Bioinformatics*, vol. 13, no. 1, article 131, 2012.

[7] D.-C. Kim, X. Wang, C.-R. Yang, and J. Gao, "Learning biological network using mutual information and conditional independence," *BMC Bioinformatics*, vol. 11, supplement 3, article S9, 2010.

[8] G. Geeven, R. E. van Kesteren, A. B. Smit, and M. C. M. de Gunst, "Identification of context-specific gene regulatory networks with GEMULA-gene expression modeling using LAsso," *Bioinformatics*, vol. 28, no. 2, pp. 214–221, 2012.

[9] M. Gustafsson, M. Hörnquist, and A. Lombardi, "Constructing and analyzing a large-scale gene-to-gene regulatory network-lasso-constrained inference and biological validation," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 3, pp. 254–261, 2005.

[10] D. Marbach, J. C. Costello, R. Küffner et al., "Wisdom of crowds for robust gene network inference," *Nature Methods*, vol. 9, no. 8, pp. 796–804, 2012.

[11] Y.-A. Kim, S. Wuchty, and T. M. Przytycka, "Identifying causal genes and dysregulated pathways in complex diseases," *PLoS Computational Biology*, vol. 7, no. 3, Article ID e1001095, 2011.

[12] J. J. B. Keurentjes, J. Fu, I. R. Terpstra et al., "Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 5, pp. 1708–1713, 2007.

[13] S. Kim and E. P. Xing, "Statistical estimation of correlated genome associations to a quantitative trait network," *PLoS Genetics*, vol. 5, no. 8, Article ID e1000587, 2009.

[14] L. Chen, L. Zhang, Y. Zhao et al., "Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways," *Bioinformatics*, vol. 25, no. 2, pp. 237–242, 2009.

[15] M. Xiong, J. Li, and X. Fang, "Identification of genetic networks," *Genetics*, vol. 166, no. 2, pp. 1037–1052, 2004.

[16] B. Liu, A. de la Fuente, and I. Hoeschele, "Gene network inference via structural equation modeling in genetical genomics experiments," *Genetics*, vol. 178, no. 3, pp. 1763–1776, 2008.

[17] B. A. Logsdon and J. Mezey, "Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations," *PLoS Computational Biology*, vol. 6, no. 12, Article ID e1001014, 2010.

[18] X. Cai, J. A. Bazerque, and G. B. Giannakis, "Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations," *PLoS Computational Biology*, vol. 9, no. 5, Article ID e1003068, 2013.

[19] A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[20] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society B: Methodological*, vol. 58, no. 1, pp. 267–288, 1996.

[22] C. Liu, L. Cheng, J. A. Badner et al., "Whole-genome association mapping of gene expression in the human prefrontal cortex," *Molecular Psychiatry*, vol. 15, no. 8, pp. 779–784, 2010.

[23] P. Sklar, J. W. Smoller, J. Fan et al., "Whole-genome association study of bipolar disorder," *Molecular Psychiatry*, vol. 13, no. 6, pp. 558–569, 2008.

[24] M. A. R. Ferreira, M. C. O'Donovan, Y. A. Meng et al., "Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder," *Nature Genetics*, vol. 40, no. 9, pp. 1056–1058, 2008.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 435853, 10 pages
http://dx.doi.org/10.1155/2014/435853

Research Article

Network Based Integrated Analysis of Phenotype-Genotype Data for Prioritization of Candidate Symptom Genes

Xing Li,[1] Xuezhong Zhou,[1] Yonghong Peng,[2] Baoyan Liu,[3] Runshun Zhang,[4] Jingqing Hu,[5] Jian Yu,[1] Caiyan Jia,[1] and Changkai Sun[6]

[1] *School of Computer and Information Technology and Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China*
[2] *School of Engineering and Informatics, University of Bradford, West Yorkshire BD7 1DP, UK*
[3] *China Academy of Chinese Medical Sciences, Beijing 100700, China*
[4] *Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing 100053, China*
[5] *Institute of Basic Theory of Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China*
[6] *Liaoning Provincial Key Laboratory of Cerebral Diseases, Institute for Brain Disorders, Dalian Medical University, Dalian 116044, China*

Correspondence should be addressed to Xuezhong Zhou; lzxzhou@gmail.com and Yonghong Peng; y.h.peng@bradford.ac.uk

Received 15 January 2014; Accepted 30 April 2014; Published 2 June 2014

Academic Editor: Xing-Ming Zhao

Background. Symptoms and signs (symptoms in brief) are the essential clinical manifestations for individualized diagnosis and treatment in traditional Chinese medicine (TCM). To gain insights into the molecular mechanism of symptoms, we develop a computational approach to identify the candidate genes of symptoms. *Methods*. This paper presents a network-based approach for the integrated analysis of multiple phenotype-genotype data sources and the prediction of the prioritizing genes for the associated symptoms. The method first calculates the similarities between symptoms and diseases based on the symptom-disease relationships retrieved from the PubMed bibliographic database. Then the disease-gene associations and protein-protein interactions are utilized to construct a phenotype-genotype network. The PRINCE algorithm is finally used to rank the potential genes for the associated symptoms. *Results*. The proposed method gets reliable gene rank list with AUC (area under curve) 0.616 in classification. Some novel genes like CALCA, ESR1, and MTHFR were predicted to be associated with headache symptoms, which are not recorded in the benchmark data set, but have been reported in recent published literatures. *Conclusions*. Our study demonstrated that by integrating phenotype-genotype relationships into a complex network framework it provides an effective approach to identify candidate genes of symptoms.

1. Introduction

Traditional Chinese medicine (TCM) is an essential part of the healthcare system in China. TCM diagnosis and treatment are formed based on a comprehensive analysis of the clinical manifestations obtained through four main procedures: observation, listening, questioning, and pulse analysis [1]. Patients with different diseases would often manifest different symptoms and signs, such as anorexia and pain, which are the evidences to be considered by physicians for clinical diagnoses in TCM [2].

Although symptoms play important role in modern biomedical diagnosis and disease classification, most modern biomedical research attempts to gain understanding of the molecular mechanism of disease phenotypes [3], including investigating the genotypes of disease/disease categories. Likewise, in the TCM field, attempt has also been made to investigate the genotypes or molecular mechanisms of the diagnosis (i.e., TCM syndrome) [4, 5].

A recent research showed that there exist metabolic biomarkers of clinical manifestations like symptoms and syndromes in different types of rheumatoid arthritis (RA)

diseases [6]. However, there is no clear understanding of the underlying molecular mechanism of symptoms and the principle of TCM syndrome in TCM field.

Large-scale diagnosis and phenotype-genotype association data, including both published literature and manually curated databases, have been gathered in the last decades [7]. PubMed, which is a public-available biomedical bibliographic database, provides a significant resource for studying the associations between diseases and clinical manifestations [8]. The phenotype-genotype association database like OMIM [9] contains high-quality data on relationships between diseases and genes. In addition, large-scale molecular network data are available [10–12], such as protein-protein interaction data, metabolic pathway data, and gene regulation data. Those provide important resources to explore the molecular correlations of symptoms.

In this paper, we first extracted the symptom-disease relationships from PubMed bibliographic records. We used the cosine similarities to evaluate the association between symptoms and diseases. We then integrated the symptom-disease relationships with disease-gene associations and protein-protein interactions (PPI) to construct a new database recording the associations between symptoms and genes. We finally used the PRINCE algorithm to rank the potential genes of symptoms. We evaluate the results of the prediction by using manually curated symptom-gene data set and PubMed literature searching. The evaluation shows that the results suggest medical meaningful insight.

2. Related Work

Using network-based approaches to gain insights into human disease has found multiple potential biological and clinical applications [13]. Further understanding of the effects of cellular interconnectedness on disease progression leads to the identification of disease biomarker genes and the pathways causing the associated diseases [14], which, in turn, offer effective targets for new drug development. Many human genetic diseases are caused by multiple genes. For genes that are associated with the same or similar phenotypes, the genes are likely to be functionally related. Such relations can be exploited to aid in searching for novel disease genes. Computational approaches have recently been proposed to predict associations between genes and diseases [15–17]. Vanunu et al. developed a network-based approach, which is known as PRINCE algorithm, for predicting causal genes and protein complexes involved in a disease of interest [18]. The availability of large-scale data of phenotype-genotype associations like OMIM, CTD [19], and PharmGKB [20] provides valuable resources for studying disease-gene associations.

Recently increasing interest on the study of molecular mechanism of symptoms was found. The underlying molecular mechanisms of several symptoms, such as depression, pain, and high blood pressure, have been discussed previously [21–23]. However, no work has been done to investigate systematically the mechanism of symptoms in the literature. Until recently, Zhou et al. used large-scale biomedical literature database to construct a symptom-based human disease network and investigate the associations between clinical manifestations of diseases and the underlying molecular interactions [24]. Their results showed that symptom-based similarity of diseases correlates strongly with the number of shared genetic associations and the extent to which their associated proteins interact. This indicates that symptoms would have their underlying molecular mechanisms needed to be further explored. In this paper, we attempt to develop a new data mining framework to explore the relationships between symptoms and genes, which may provide scientific evidences to traditional Chinese medicine in individualized diagnosis and treatment because symptoms are the main clinical manifestations captured by TCM physicians for both diagnosis and treatment.

3. Methods

3.1. Phenotype-Genotype Data Integration. In order to extract the associations between symptoms and genes, we first built symptom-disease associations based on a large number of medical literatures in PubMed [25] and the Medical Subject Headings (MeSH). Using the cooccurrence of diseases and symptoms, we construct two vectors s and d to calculate the similarity of symptom and disease, in which d denotes a disease vector represented by its cooccurrence symptoms and s denotes a symptom vector represented by its cooccurrence symptoms as well. Suppose we have a dictionary with n symptom items, we would have an n-features vector for both disease and symptom. Based on the vectors of diseases and symptoms, we calculate the similarity of symptom and disease using cosine correlation:

$$T(d, s) = \frac{d \cdot s}{\|d\|^2 \times \|s\|^2}. \quad (1)$$

In this study, we integrated three public available disease-gene databases (OMIM, CTD, and PharmGKB) and five protein-protein interactions databases (HPRD, BioGrid, IntAct, MINT, and DIP) into database (Figure 1). Based on these data sets a heterogeneous network is constructed with nodes representing symptoms, diseases, and proteins, respectively, and the links representing symptom-disease relationships, disease-gene associations, and protein-protein interactions.

3.2. Network Inference for Prioritization of Symptom Candidate Genes. The network-based disease gene prediction approach, PRINCE, is used for predicting the genes with respect to symptom. The initialization of the parameters in PRINCE algorithm is the symptom-disease correlations, disease-gene associations, and protein-protein interactions. It uses a propagation-based algorithm [26] to infer a scoring function for estimating the strength of an association. A score is defined for each gene, which reflects the prior information of the genes on the related disease. The score is then used in combination with a PPI network for the identification of proteins involved in the given symptom, as shown in Figure 2.

3.3. Computing the Prioritization Function. The prioritization of genes for a query symptom (s) is performed based on

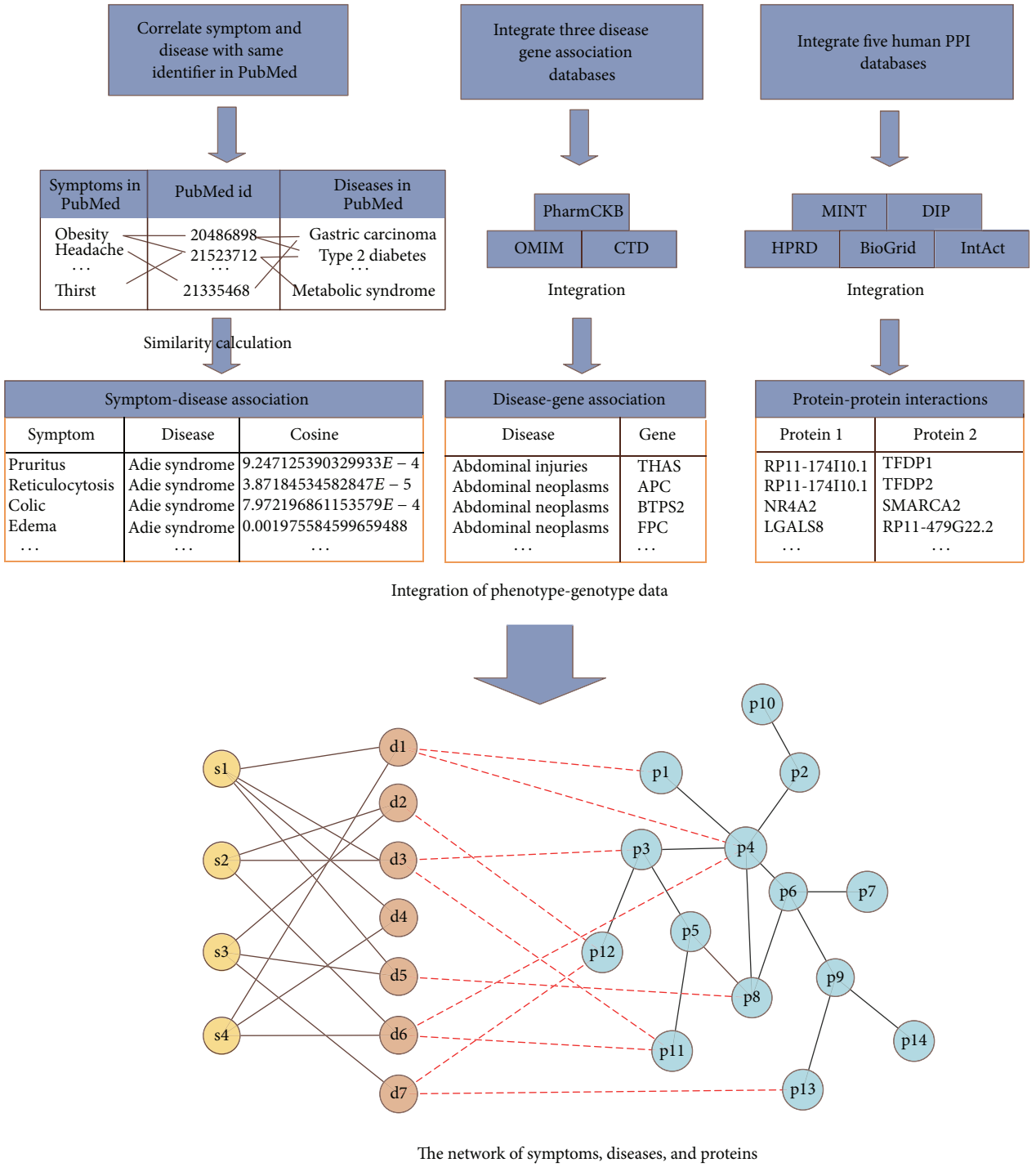


FIGURE 1: The integration of phenotype-genotype data. Symptom-disease associations are extracted based on the fact that the symptom and disease appeared in same bibliographic record (including title, abstract, and MeSH) of PubMed. Three disease gene association databases (i.e., OMIM, CTD, and PharmGKB) and five human PPI databases (i.e., HPRD, BioGrid, IntAct, MINT, and DIP) are integarted in this study. The relationships among symptoms (denoted s1–s4), diseases (denoted d1–d7), and proteins (denoted p1–p14) are then extracted.

the given symptom-disease associations (denoted by A), disease-gene associations (B), and a protein-protein interaction network $G = (V, E)$, where V is a set of proteins and E is a set of interactions between proteins. The goal of the algorithm is to prioritize all the proteins in V with respect to s.

Let $F : V \rightarrow \Re$ represent a prioritization function; F reflects the relevance of v ($v \in V$) to s. $Y : V \rightarrow [0, 1]$ represent a prior knowledge function, where 1 is assigned to a protein that is known to be related to the disease with respect to s, and 0 otherwise. In other words, Y is the vector of genes

Table 1: The result of phenotype-genotype data integration.

Number of symptoms	Number of diseases	Number of proteins
322	4,219	14,221
Number of symptom-disease associations	Number of disease-gene associations	Number of protein-protein interactions
125,226	28,336	94,536

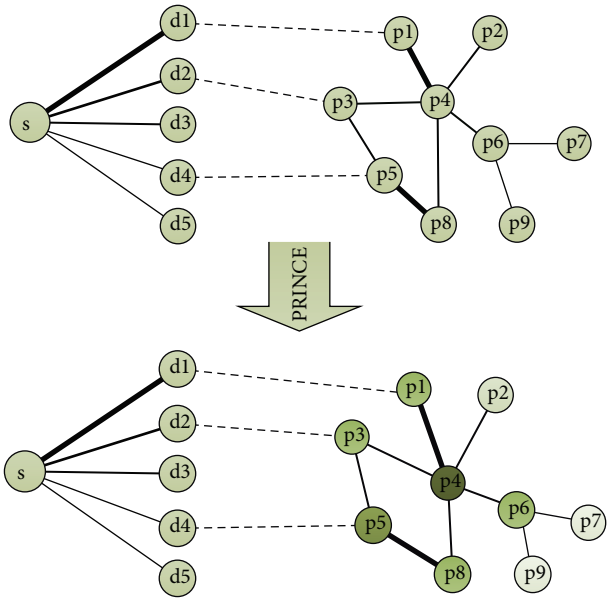


Figure 2: The approach for predicting the genes with respect to symptom using PRINCE algorithm. For a query symptom S, it has varying degrees of relationship with other diseases, denoted by d1–d5 (where the thickness of lines represents degree of correlation between symptom and diseases). p1–p9 comprise the protein set of a protein-protein interaction network, where interactions are denoted by lines with different thickness (confidence). PRINCE uses an iterative propagation method to assign a score of each protein. The protein with higher score is considered to be the causal gene candidate for symptom S.

which are known to be causal gene of diseases with respect to symptom. To obtain Y, we first analyzed the distribution of similarity between symptom and disease and found that the symptom may have high possibility of relating to a disease when their similarity is above 0.1. Here, we want to choose the diseases which have high possibility to associate with a symptom, so that we could get the related genes to build Y. The 10% top ranked disease-symptom relationships with similarities larger than 0.1 are chosen (in our experiment the threshold is 0.57). At last, we selected the ten most related diseases as the diseases corresponding to symptom and its causal genes to build Y.

By iterative procedures, the information is transferred between their neighbors, as defined by

$$F^t := \alpha W' F^{t-1} + (1 - \alpha) Y, \quad (2)$$

where $F^1 := Y \cdot W'$ is a $|V| \times |V|$ matrix which is a normalized form of W (described below) and F and Y are viewed here as vectors of size $|V|$. The details on the inference of F in PRINCE algorithm could be found [18]. The parameter $\alpha \in (0, 1)$ weighs the relative importance of these constraints with respect to one another. Here α is set to be 0.9 as suggested in the PRINCE algorithm that the appropriate values of α could be above 0.5 with fast convergence and 0.9 gets the comparative highest performance [18].

3.4. Evaluation Methods. We use Human Phenotype Ontology (HPO) [27] as the benchmark data to evaluate the results. HPO was manually curated from OMIM records and constructed with the goal of covering all phenotypic abnormalities that are commonly encountered in human monogenic diseases [28]. In this study we use the T184 (Sign or Symptom) semantic type of UMLS [29] to filter the phenotype terms and construct a subset of HPO phenotypes (349 records), after filtering the phenotype-genotype associations with focusing on symptoms results in 7,262 symptom-gene records and 1,275 related genes. To deal with the issue of HPO having different symptom terms from MeSH, we used UMLS to map HPO symptom terms to MeSH. We finally obtained 3,418 symptom-gene records with 139 symptoms and 937 genes, which were used for evaluation. Although HPO contains high-quality data on phenotype ontology and genotype-phenotype (mainly on diseases and disorders) associations, the data is rather incomplete and still lack many well-known symptom-gene associations. We evaluated the symptom-gene prediction results by three approaches: (1) compare our rank list with the genes in HPO and calculated recall and AUC [30], (2) compare our result with random case, and (3) evaluate the random chosen results by recent published literatures.

4. Results

We extracted 125,226 symptom-disease associations with 322 symptoms and 4,219 diseases from PubMed bibliographic records and calculated the cosine similarity between symptoms and diseases. We constructed 94,536 protein-protein interactions with 14,221 proteins and integrated 28,336 disease-gene associations (shown in Table 1).

The protein-protein interactions were assigned 1 if they are correlated. We used these scores to construct the adjacency matrix W. As a result, we obtained totally 4,211,956 symptom-gene associations between 290 symptoms and 14,221 genes with correlation values bigger than zero. The distribution of correlation between symptoms and genes is depicted in Figure 3. It is noted that 83% of the correlations are <0.001, and only about 0.24% are distributed on the range of bigger than 0.01. We consider that the genes with

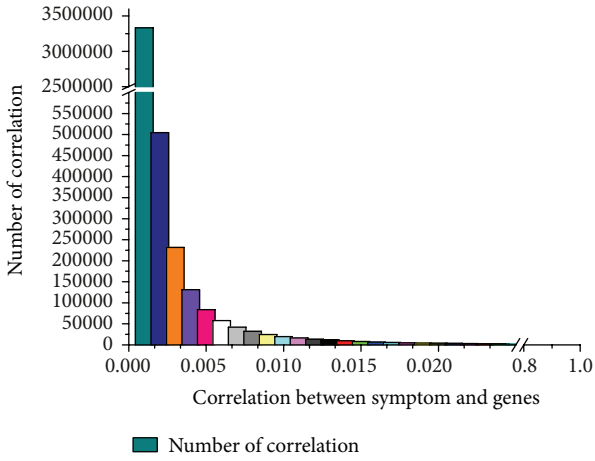


Figure 3: The distribution of correlation between symptom and genes.

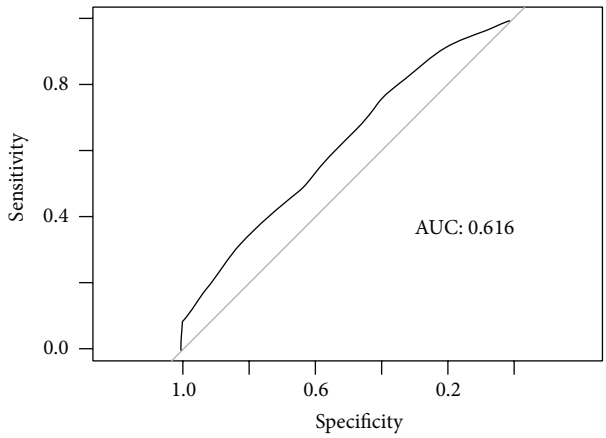


Figure 4: ROC curve to assess prediction performance.

correlation scores bigger than 0.01 have higher possibility than most of the genes (i.e., 83% genes). Therefore, these genes with correlation scores higher than 0.01 are considered to be the potential genes related to symptoms in this study.

Using the HPO benchmark data, we quantify the accuracy of the prediction by comparing the predicted gene list of symptoms with that of the benchmark data. The area under the ROC curves (AUC) of the proposed method is 0.616 (Figure 4).

In order to evaluate the effectiveness of the gene ranking, we also compared the result with random prediction case. We calculate the quantity of genes contained in HPO on the top of our gene list ($P < 0.05$) by comparing with the average quantity of randomly selected the same number of genes. It is noted that the number of true positive candidate genes is 10-fold of the random prediction, with the best case being 249-fold of the random prediction. We take symptom *Muscle Cramp* as an example to compare our result with random case. Given 27 genes in HPO, there are 10 genes included in the top 251 genes ($P < 0.05$) of our candidate genes list. Randomly choosing 251 genes among all the genes (14,221 genes), the possibility of each gene being causing gene is 0.0018986 (27/14,221, we have the hypothesis that the genes in HPO are all causing genes). The expected number of genes in HPO is 0.477 (0.0018986∗251); that is, there is on average 0.477 true causing genes in HPO gene list if 251 genes are randomly selected. So the number of true positive candidate genes is approximately 20-fold (10/0.477) over the random prediction.

To demonstrate the effectiveness of this method, we listed the suggested genes of headache and hemiplegia for instance. Through the analysis of the distribution of all the scores of symptom related genes, we found that most scores (95% in average) are in very low values (i.e., 0.01) with some exceptions of having much larger scores than these row values. Table 2 shows the top 46 ranked genes of the 13,966 genes whose correlation scores are greater than 0.01 with respect to the symptom of headache. We found that TNF and EDNRA are the causing genes for headache as listed in HPO. (the Italic font in Table 2, recall is 6.25% of the 32 genes). Several other genes related to headache in HPO including ENG (rank 52th), ACVRL1 (rank 65th), TGFB1 (rank 74th), VHL (rank 269th), COL4A1 (rank 563th), NF2 (rank 1520th), TTR (rank 2270th), MSX2 (rank 2622th), FGFR2 (rank 2636th), PGK1 (rank 2773th), FAM123B (rank 3002th), SH2B3 (rank 3994th), LRP5 (rank 4286th), NOTCH3 (rank 4386th), SDHB (rank 5618th), and CACNA1A (rank 5855th) are ranked in the top 50%.

We were aware that the HPO is an incomplete database. To have a more comprehensive evaluation on the prediction result, we manually searched the literature in PubMed for the symptom-gene associations. Among the top 10 genes of our list, we found that five additional genes CALCA, TGFBR2, ESR1, KCNK18, and MTHFR (bold font in Table 2) are all considered to be related to headache in recent published literatures [31–34], although they are not recognized in the HPO database. As a result, we recognized totally 7 possible causing genes (CALCA, TGFBR2, TNF, ESR1, EDNRA, KCNK18, and MTHFR) of headache in the top 10 genes.

The relationship between symptoms and diseases is complicated. Some symptoms would be more particularly manifested in several diseases than others. This kind of clinical association would have its underlying molecular mechanisms. To explore the interactions of the related genes of symptoms and diseases in the context of PPI network, we show a subset of protein-protein interactions with respect to headache in Figure 5, which is constructed by the genes connected with 6 diseases related to headache directly. In Figure 5, genes connected with the same diseases are marked in the same colors. We found that 15 genes of 32 genes in HPO (marked in box) in our subnet are the causal genes of diseases or locate on their shortest path. It is possible that the causal genes of a disease, which holds the symptom as particular phenotype, would be the related genes for symptom (marked in pink box), or the candidate genes for symptom would possibly locate on the shortest paths of these genes of the diseases, which have the related symptoms as general phenotypes (marked in red box). To have more clear view of the relationships between the candidate genes of symptoms and the casual genes of the diseases holding

Table 2: The top 46 rank list of genes predicted with respect to headache.

Number	Symbol	Correlation	Number	Symbol	Correlation
1	**CALCA**	0.12828040612489	26	PROC	0.01319213392192
2	F2	0.12201692087326	27	TGFBR1	0.01314425880888
3	**TGFBR2**	0.12033672267710	28	F2RL3	0.01305473906526
4	*TNF*	0.11489921901213	29	MRPL28	0.01303120114399
5	**ESR1**	0.11394663869964	30	RP1-90J20.6	0.01292643824309
6	**KCNK18**	0.11171558541349	31	NDUFB10	0.01268903908242
7	HTR2A	0.10968115391880	32	GRK6	0.01178478736965
8	*EDNRA*	0.10800919199361	33	BBS4	0.01167226005919
9	TREX1	0.10295846965404	34	THBD	0.01165325302184
10	**MTHFR**	0.10278770151638	35	RAMP3	0.01155667772559
11	CALCRL	0.02849332330671	36	HSP90B3P	0.01154759481953
12	RAMP2	0.02819497378019	37	ATP2B4	0.01147183196121
13	RAMP1	0.02816780187933	38	GGCX	0.01145977743914
14	EDN3	0.02015371056218	39	BBS1	0.01135926399377
15	NAA38	0.01897112176120	40	NME3	0.01105404047118
16	GNA11	0.01715273254725	41	F9	0.01092877796805
17	PROZ	0.01677771421601	42	AKR7A2	0.01081719314504
18	NRD1	0.01533854606373	43	TCTEX1D4	0.01078978137476
19	EDN1	0.01522967177446	44	GP5	0.01063032851227
20	NT5DC3	0.01477167543805	45	SERPINA5	0.01041954082501
21	PDIK1L	0.01477167543805	46	PROS1	0.01034320224653
22	PON2	0.01413140552619			
23	SLC31A2	0.01392606686194			
24	RP11-9H12.2	0.01392606686194			
25	ATP2B2	0.01342703847737			

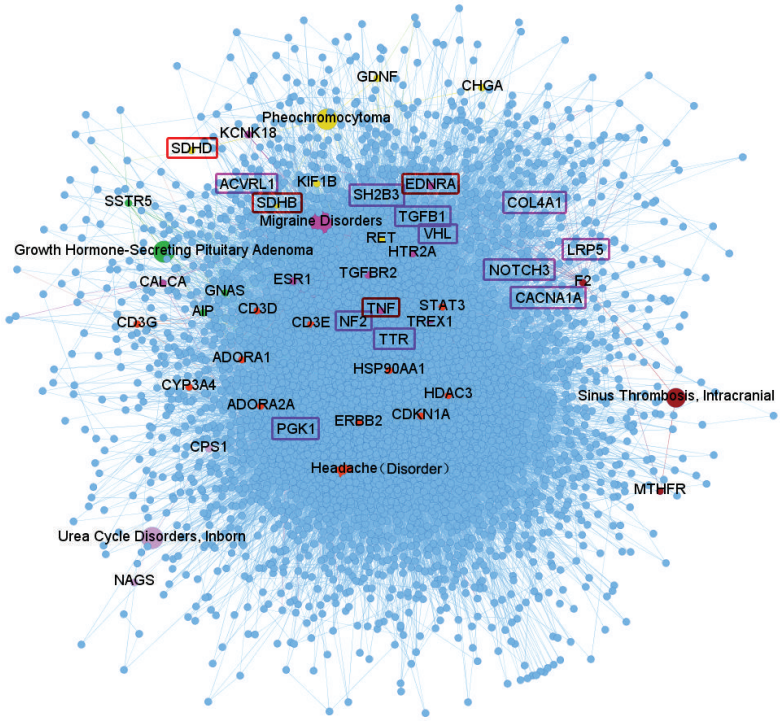


Figure 5: The subset of protein-protein interaction with respect to headache.

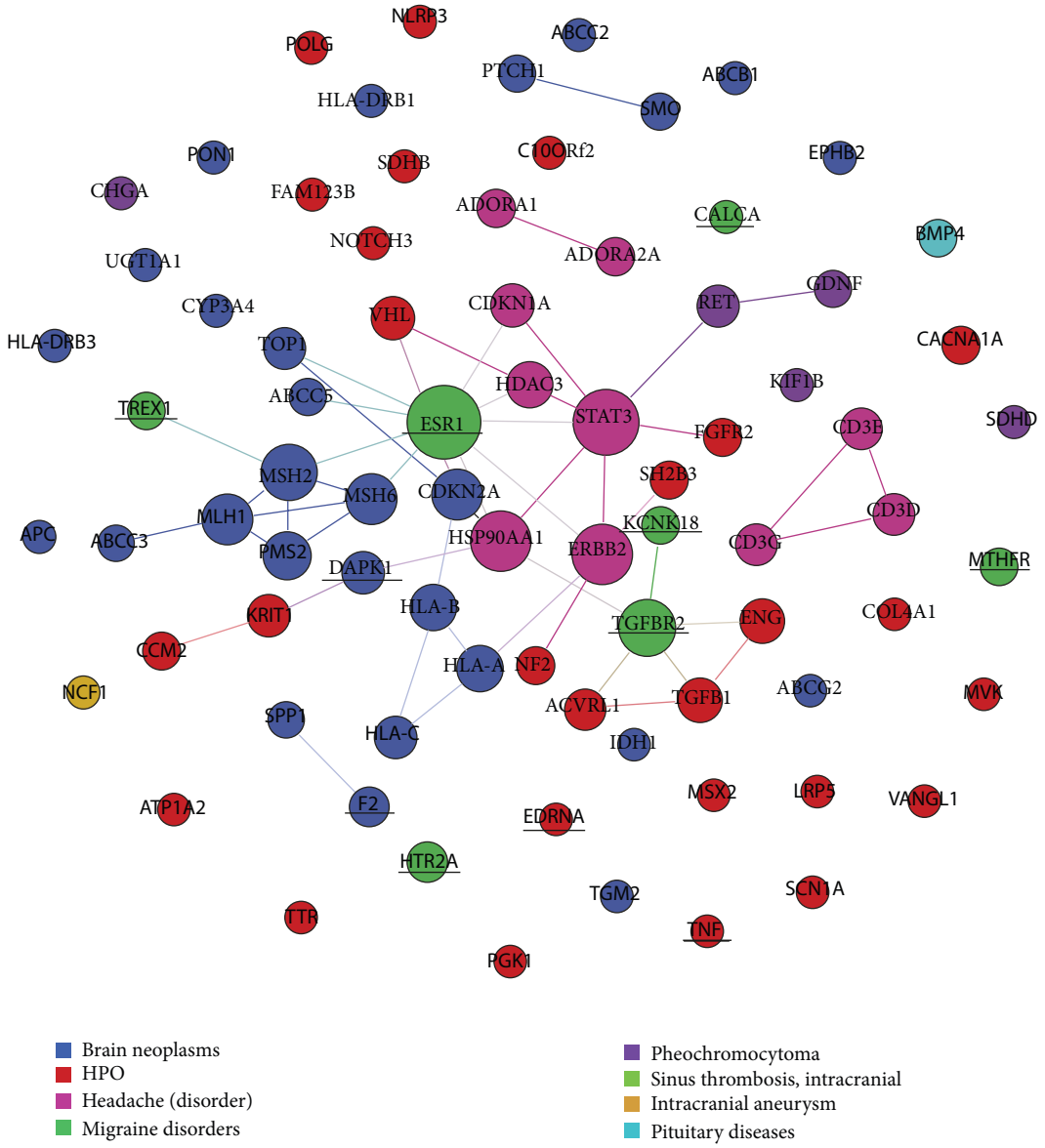


Figure 6: The direct relationship among genes connected with diseases relevant to headache symptom and genes in HPO.

the corresponding symptoms as particular manifestations, we also constructed a network to show the direct relationships among the causing genes of diseases related to headache and the genes in HPO (Figure 6, genes in HPO are marked in red and genes connected with different diseases are marked with different colors). The genes, CALCA, TGFBR2, TNF, ESR1, EDNRA, MTHFR, and so forth, of our top 10 rank list (mentioned above) are marked with underline. We found that the candidate genes with high scores of headache symptom are the causal genes of the diseases, which regard headache as distinct symptom, such as migraine. It is possible that the causing genes of diseases with respect to the distinct symptoms would also be related to their corresponding symptoms.

Table 3 lists the 83 top ranked genes with respect to hemiplegia with correlation greater than 0.01. In the causing genes of hemiplegia in HPO, four genes, namely, COL4A1, CACNA1A, ATP1A2, and SCN1A, are all found in the top 83 candidate genes (recall is 66.7%) except for the gene DOCK8 which is ranked 6667th in whole list of 14,221 genes. However, we found no related publications on indicating the relationships between the 8 genes (except for the 2 genes included in HPO) of the top 10 genes and hemiplegia after manually searching the PubMed literatures.

5. Discussion

As a kind of established clinical manifestations in TCM clinical, symptoms provide key information for the classification of the state of human disease and personalized herb treatment. Symptoms are essentially objective although

Table 3: The rank of candidate genes with respect to hemiplegia.

Number	Gene symbol	Correlation	Number	Gene symbol	Correlation
1	HMOX1	0.1332389455434	46	CACNB4	0.0132782710879
2	MMP9	0.1287040668111	47	KLK6	0.0131043482021
3	*COL4A1*	0.1243046971781	48	CXCL6	0.0130491204825
4	SERPIND1	0.1195621966536	49	CXCL1	0.0130308002686
5	PLAT	0.1133487392385	50	TGFBI	0.0129966667251
6	OFD1	0.1121152624575	51	MMP26	0.0128881745922
7	CDKN1A	0.1107796508251	52	BMP3	0.0128183446929
8	STAT3	0.1106360535810	53	UFD1L	0.0126019281191
9	HDAC3	0.1104973436194	54	KISS1	0.0124874609423
10	*CACNA1A*	0.1072041148709	55	LCN2	0.0123446041858
11	PGK1	0.1041298451985	56	CXCL5	0.0122282204837
12	TREX1	0.1033025393792	57	HAPLN1	0.0121753323548
13	MTHFR	0.1028354996936	58	CTSG	0.0121249443517
14	*ATP1A2*	0.1023789578124	59	SERPINI1	0.0120388399124
15	*SCN1A*	0.1019561553063	60	CABP1	0.0120280578478
16	INPP5E	0.1005153381688	61	CD93	0.0120005018879
17	BLVRB	0.0399295586077	62	COL16A1	0.0119332179348
18	CTA-286B10.6	0.0241333799518	63	PRSS2	0.0119006841698
19	POR	0.0229706000329	64	COL1A1	0.0118755109423
20	COL4A2	0.0203674650216	65	IGHG1	0.0117128695292
21	NAA38	0.0192928165139	66	THBS3	0.0115158569766
22	THBS2	0.0186100092167	67	TFPI	0.0115052630961
23	SAA4	0.0176962732383	68	DCN	0.0112824650786
24	F2	0.0172012123620	69	UBC	0.0110403727935
25	CACNB1	0.0165565034731	70	MMP2	0.0109667286026
26	COL4A4	0.0164204509392	71	COL7A1	0.0109021629899
27	FN1	0.0161953766962	72	LAMA1	0.0106705480427
28	TPT1	0.0159187987328	73	YWHAG	0.0106543097610
29	COL4A3	0.0159043555825	74	IGHA1	0.0104865314738
30	SERPINE2	0.0158807849607	75	RP11-157P1.6	0.0103829026309
31	HABP2	0.0155572576434	76	FAM190B	0.0103401713298
32	COL4A6	0.0154901122445	77	PZP	0.0103015667226
33	COL4A5	0.0153811800445	78	BTC	0.0102664672842
34	SAA2	0.0151416061199	79	NID2	0.0102327719152
35	XXyac-YX65C7_A.1	0.0148026704536	80	TF	0.0101538081025
36	PLG	0.0144321310724	81	RP11-417O11.1	0.0101229813573
37	MATN2	0.0144194792723	82	SERPINA5	0.0101166365026
38	OSM	0.0142128473336	83	NID1	0.0100899609704
39	SNTA1	0.0142116180479			
40	RECK	0.0140016715074			
41	FBLN2	0.0137418115855			
42	COCH	0.0137176242361			
43	MMP10	0.0135697895912			
44	ELANE	0.0134913849548			
45	THBS1	0.0133469773733			

the observation and description of symptoms incorporate subjective factors like human sense and language. Therefore, investigation of the underlying molecular mechanisms of symptoms is more feasible than TCM syndrome. Through integrating disease-symptom associations and multiple phenotype-genotype data sources, this paper proposes a network inference method to predict the candidate gene list for symptoms. Like similar work for disease gene predictions [35, 36], the rank list of symptom-related candidate genes can promote the discovery of molecular mechanisms

of symptoms and thereafter draw the picture of connection between symptoms and genes with respect to diseases. Evaluation shows the effectiveness of the method in identifying genes related to symptoms. Like the predicted genes of headache, more predicted genes could be further investigated to understand the medical insights, which would ultimately support the researchers to confirm the causal genes of symptoms in laboratory study.

It is necessary to mention that this paper is intended to introduce the proposed integrated network framework for predicting the symptom candidate genes. Several aspects related to the method could be improved in future work. Firstly, a carefully curated and evaluated database needs to be established for benchmark data set. Currently, although HPO provides a start point, more effects are needed to obtain high quality symptom-gene databases. While this database is curated, it would offer reliable benchmark platform to evaluations and possible supervision for machine learning methods. On the other hand, due to the complicated confounders involved in symptom-disease relation detection from biomedical literatures, a comprehensive database on disease-symptom relationships would be also very helpful. Secondly, because the similarities between diseases and symptoms indicate different degree of correlations, the similarities between symptoms and diseases could be systematically utilized to improve the iterative computing procedures of random walk related network inference methods. Thirdly, it is highly valuable to investigate the molecular correlations between symptoms and diseases to detect the molecular patterns connecting these two phenotype entities. When some network characteristics underlying the connection are discovered, it would give guideline framework for the development of symptom-gene prediction methods.

Conflict of Interests

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

Acknowledgments

This work was partially supported by NSFC Project (61105055, 81230086), China 973 Program (2014CB542903), The National Key Technology R&D Program (2013BAI02B01, 2013BAI13B04), the National S&T Major Special Project on Major New Drug Innovation (2012ZX09503-001-003), and the Fundamental Research Funds for the Central Universities.

References

[1] M. Jiang, C. Lu, C. Zhang et al., "Syndrome differentiation in modern research of traditional Chinese medicine," *Journal of Ethnopharmacology*, vol. 140, no. 3, pp. 634–642, 2012.

[2] W. Osler, "The principles and practice of medicine: designed for the use of practitioners and students of medicine," *Journal of the American Medical Association*, vol. 82, pp. 1901–1182, 2005.

[3] J. R. Lupski and P. Stankiewicz, "Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes," *PLoS Genetics*, vol. 1, no. 6, pp. 627–633, 2005.

[4] Z. Guo, S. Yu, Y. Guan et al., "Molecular mechanisms of same TCM syndrome for different diseases and different TCM syndrome for same disease in chronic hepatitis B and liver cirrhosis," *Evidence-Based Complementary and Alternative Medicine*, vol. 2012, Article ID 120350, 9 pages, 2012.

[5] J. C. Reubi, "Peptide receptors as molecular targets for cancer diagnosis and therapy," *Endocrine Reviews*, vol. 24, no. 4, pp. 389–427, 2003.

[6] M. Jiang, T. Chen, H. Feng et al., "Serum metabolic signatures of four types of human arthritis," *Journal of Proteomic Research*, vol. 12, no. 8, pp. 3769–3779, 2013.

[7] Z. Wu, X. Zhou, B. Liu, and J. Chen, "Text mining for finding functional community of related genes using TCM knowledge," in *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 459–470, Springer, 2004.

[8] D. L. Wheeler, T. Barrett, D. A. Benson et al., "Database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 35, no. 1, pp. D5–D12, 2007.

[9] 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.

[10] M. S. Cline, M. Smoot, E. Cerami et al., "Integration of biological networks and gene expression data using Cytoscape," *Nature Protocols*, vol. 2, no. 10, pp. 2366–2382, 2007.

[11] A. Rzhetsky, I. Iossifov, T. Koike et al., "GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data," *Journal of Biomedical Informatics*, vol. 37, no. 1, pp. 43–53, 2004.

[12] 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.

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

[14] G. Östlund, M. Lindskog, and E. L. L. Sonnhammer, "Network-based identification of novel cancer genes," *Molecular and Cellular Proteomics*, vol. 9, no. 4, pp. 648–655, 2010.

[15] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinformatics*, vol. 26, no. 8, pp. 1057–1063, 2010.

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

[17] Y. Moreau and L. C. Tranchevent, "Computational tools for prioritizing candidate genes: boosting disease gene discovery," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 523–536, 2012.

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

[19] C. J. Mattingly, M. C. Rosenstein, G. T. Colby, J. N. Forrest Jr., and J. L. Boyer, "The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies," *Journal of Experimental Zoology A: Comparative Experimental Biology*, vol. 305, no. 9, pp. 689–692, 2006.

[20] M. Hewett, D. E. Oliver, D. L. Rubin et al., "PharmGKB: the pharmacogenetics knowledge base," *Nucleic Acids Research*, vol. 30, no. 1, pp. 163–165, 2002.

[21] M. Aguilera, B. Arias, M. Wichers et al., "Early adversity and 5-HTT/BDNF genes: new evidence of gene-environment interactions on depressive symptoms in a general population," *Psychological Medicine*, vol. 39, no. 9, pp. 1425–1432, 2009.

[22] M. Costigan and C. J. Woolf, "Pain: molecular mechanisms," *Journal of Pain*, vol. 1, no. 3, pp. 35–44, 2000.

[23] X. Jeunemaitre, F. Soubrier, Y. V. Kotelevtsev et al., "Molecular basis of human hypertension: role of angiotensinogen," *Cell*, vol. 71, no. 1, pp. 169–180, 1992.

[24] X. Zhou, J. Menche, A. L. Barabasi, and A. Sharma, "Human symptom disease network," *Nature Communications*, 2014.

[25] Z. Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature," *Database*, vol. 2011, p. baq036, 2011.

[26] O. Vanunu and R. Sharan, "A propagation based algorithm for inferring gene-disease associations," in *Proceedings of the German Conference on Bioinformatics on Bioinformatics (GCB '08)*, pp. 54–63, Gesellschaft für Informatik, 2008.

[27] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The human phenotype ontology: a tool for annotating and analyzing human hereditary disease," *American Journal of Human Genetics*, vol. 83, no. 5, pp. 610–615, 2008.

[28] P. N. Robinson and S. Mundlos, "The human phenotype ontology," *Clinical Genetics*, vol. 77, no. 6, pp. 525–534, 2010.

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

[30] J. M. Lobo, A. Jiménez-valverde, and R. Real, "AUC: a misleading measure of the performance of predictive distribution models," *Global Ecology and Biogeography*, vol. 17, no. 2, pp. 145–151, 2008.

[31] M. A. Kaunisto, M. Kallela, E. Hämäläinen et al., "Testing of variants of the MTHFR and ESR1 genes in 1798 Finnish individuals fails to confirm the association with migraine with aura," *Cephalalgia*, vol. 26, no. 12, pp. 1462–1472, 2006.

[32] B. Guldiken, T. Sipahi, T. Remziye et al., "Calcitonin gene related Peptide gene polymorphism in migraine patients," *The Canadian Journal of Neurological Sciences*, vol. 40, no. 5, pp. 722–725, 2013.

[33] T. Freilinger, V. Anttila, B. de Vries et al., "Genome-wide association analysis identifies susceptibility loci for migraine without aura," *Nature Genetics*, vol. 44, no. 7, pp. 777–782, 2012.

[34] I. Rainero, E. Rubino, K. Paemeleire et al., "Genes and primary headaches: discovering new potential therapeutic targets," *The Journal of Headache and Pain*, vol. 14, no. 1, pp. 1–8, 2013.

[35] J. Freudenberg and P. Propping, "A similarity-based method for genome-wide prediction of disease-relevant human genes," *Bioinformatics*, vol. 18, no. 2, pp. S110–S115, 2002.

[36] R. A. George, J. Y. Liu, L. L. Feng, R. J. Bryson-Richardson, D. Fatkin, and M. A. Wouters, "Analysis of protein sequence and interaction data for candidate disease gene prediction," *Nucleic Acids Research*, vol. 34, no. 19, p. e130, 2006.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 207589, 17 pages
http://dx.doi.org/10.1155/2014/207589

Research Article

Qualitative and Quantitative Analysis for Facial Complexion in Traditional Chinese Medicine

Changbo Zhao,[1] Guo-zheng Li,[1] Fufeng Li,[2] Zhi Wang,[2] and Chang Liu[3]

[1] *Department of Control Science and Engineering, Tongji University, Shanghai 201804, China*
[2] *Laboratory of Information Access and Synthesis of TCM Four Diagnosis, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China*
[3] *School of Film & TV Arts and Technology, Shanghai University, Shanghai 200444, China*

Correspondence should be addressed to Guo-zheng Li; gzli@tongji.edu.cn and Fufeng Li; fufeng_lee@hotmail.com

Received 25 January 2014; Accepted 24 February 2014; Published 22 May 2014

Academic Editor: Xing-Ming Zhao

Facial diagnosis is an important and very intuitive diagnostic method in Traditional Chinese Medicine (TCM). However, due to its qualitative and experience-based subjective property, traditional facial diagnosis has a certain limitation in clinical medicine. The computerized inspection method provides classification models to recognize facial complexion (including color and gloss). However, the previous works only study the classification problems of facial complexion, which is considered as qualitative analysis in our perspective. For quantitative analysis expectation, the severity or degree of facial complexion has not been reported yet. This paper aims to make both qualitative and quantitative analysis for facial complexion. We propose a novel feature representation of facial complexion from the whole face of patients. The features are established with four chromaticity bases splitting up by luminance distribution on CIELAB color space. Chromaticity bases are constructed from facial dominant color using two-level clustering; the optimal luminance distribution is simply implemented with experimental comparisons. The features are proved to be more distinctive than the previous facial complexion feature representation. Complexion recognition proceeds by training an SVM classifier with the optimal model parameters. In addition, further improved features are more developed by the weighted fusion of five local regions. Extensive experimental results show that the proposed features achieve highest facial color recognition performance with a total accuracy of 86.89%. And, furthermore, the proposed recognition framework could analyze both color and gloss degrees of facial complexion by learning a ranking function.

1. Introduction

Nowadays, the Traditional Chinese Medicine (TCM) has become a global and important diagnostic approach in the medical field. In TCM, Inspection, auscultation and olfaction, interrogation, and palpation are four diagnostic methods to recognize the human pathological conditions. The traditional facial complexion diagnosis is an important part of inspection examination. The doctors can understand the physiological functions and pathological effects of human body through the inspection of complexion [1]. In retrospect, the inspection of complexion has been studied in TCM for a long time; however, it is still mainly based on the observations with practitioners' nude eyes. The disadvantage of nude observation is that it is likely to get inconsistent diagnostic results due to the large dependence on practitioners' subjective experiences and personal knowledge. Thus, it is worthwhile for TCM doctors and scholars to design an objective and reliable computer-assisted system for facial complexion diagnosis.

Recently, TCM experts stated that the notion facial complexion diagnosis includes two aspects: the facial color diagnosis and the facial gloss diagnosis. Therefore, by avoiding phrasing confusion, we declare here that we will employ the notion facial color instead of facial complexion, if it is basically studying one's facial color recognition. And we may still use the notion facial complexion, if it represents both color and gloss in this paper.

According to the TCM facial diagnosis theory, diagnostic significance of five colors implicates the correlation between five colors of facial skin and diseases. That is, the changes

of facial colors can reveal pathological changes of different viscera and bowels with different natures. In addition, the lustre or gloss of skin refers to the reflective, shiny, and smooth characteristics of facial skin [2] which can help TCM experts to understand the states of visceral and bowels and the severity degrees of diseases.

To make it more detailed, the human facial color always has two cases: normal color and morbid color. The normal or healthy color is always further divided into two parts: normal individual color and varied normal color. The normal individual color refers to the normal natural color of skin that never changes in one's whole life due to ethnic and genetic factors. For the varied normal color, it is influenced by environmental factors and will change slightly with respect to the variations of the climates and seasons.

With these varied colors, which essentially are not morbid colors, the analysis of facial color will be difficult to put into practice; even the TCM practitioners find it hard to discriminate it from morbid colors. That will result in inconsistent diagnosis conclusion under different practitioners' observations. Just like the assessment about consistency of TCM experts' diagnosis results [1].

Besides, different from normal color, morbid color could be reflected on one's face on the condition that the pathological changes of five internal viscera (these are heart, liver, spleen, lungs, and kidney). The morbid color is classified to five colors: reddened facial color, bluish facial color, yellow facial color, pale facial color, and darkish facial color. Their correlation with syndrome in TCM can be simply stated as follows: reddened color implies heat syndrome; bluish color represents blood stasis syndrome; yellow color suggests dampness syndrome and deficiency syndrome, convulsive syndrome, pain syndrome, and cold syndrome; pale color indicates deficiency syndrome and cold syndrome; darkish color hints cold syndrome, blood stasis syndrome, fluid-retention syndrome, and the deficiency of kidney syndrome [3].

Furthermore, various lustre or gloss of skin also manifests different states of diseases. For example, if a patient is considered normal and with healthy facial color, the gloss of skin would be always bright and moist; that is, glossy skin indicates that patient is healthy [2]. However, for the morbid color patient, the gloss of skin reflects the severity of diseases. Patient whose facial gloss is bright and moist would indicate mild illness and easiness to cure no matter what color it is, while dull and dry gloss represent serious illness and difficulty to cure.

Generally speaking, it is significant to study the computerized facial color diagnosis more in-depth with multiple aspects. In the following, we will firstly review related facial color diagnosis techniques and systems in TCM. Then, the shortcomings of existing frameworks will be discussed, eventually leading to our motivation for building a facial color diagnosis with both qualitative and quantitative methods.

A large amount of pattern recognition and data mining technologies have been developed and designed for medical analysis and diagnostic standardization. The literatures [5, 6] propose novel algorithms for medical diagnosis and analysis. Recent machine learning algorithms, such as multi-label algorithm and multi-instance algorithm, have been introduced to deal with special medical data analysis [7–10]. Especially, some works about TCM four diagnostic methods [4, 11] have also been studied recently. There is an extensive literature on TCM four diagnostic methods, but we prefer to review just a few papers relevant with facial complexion analysis.

Early in China, some preliminary results have been reported in facial color analysis using colorimeter or infrared thermograph instrument. Unlike the previous foundation research, Liu and Guo [12] investigated the hepatitis diagnosis with face images by digital camera acquisition device. In their system, five facial regions are firstly segmented using skin detection, facial normalization, and horizontal position of the mouth, nostril, and eyebrow location. Then, the mean value of RGB color is taken as dominant color and k-nearest neighbour (KNN) for color classification. But the segmentation results are not always correct which needs to be adjusted manually. So Wang et al. [13] give an optimized version to improve the segmentation results and health versus hepatitis recognition accuracy. In addition, the fuzzy c-means (FCM) clustering method is introduced to extract the dominant color; finally six color spaces are considered and intensively analyzed for the best classification.

For the facial color analysis, the prior result [1] has established a five-color scale to measure facial color in RGB color space. Another research [14] firstly extracts 15 diagnostic feature points using AdaBoost and Active Shape Model (ASM), and, next, each rectangular region around each point will be used to calculate the facial color similarly as the above studies. More recently, Zhang et al. [2] analyze the facial color gamut and construct six centroids to calculate facial color distribution for health and several diseases classification. The texture features are firstly applied to their facial diagnosis system. For the facial gloss analysis, Zhang et al. [2] also present healthy classification using facial gloss information. Another main work for gloss classification is carried out using feature dimensionality reduction techniques [15]. However, some of the above works share several problems as follows.

(i) Adoptive dominant color as in [1, 13] is strongly depending on greater numbers of clusters which may be encountered failure to represent the morbid color in two cases: one case is the incorrect location of facial region that can derive biased dominant color; another is patients whose disease status is considered as mild, which will give rise to the facial morbid color slightly and inapparently. And, finally, it will extract inaccurate dominant color. One proposed way may overcome these problems as given in [14], but it is impractical to extract the basic skin color of upper arm for all the subjects.

(ii) Only local and a few small facial regions used for the whole facial complexion diagnosis (including both color and gloss) may be difficult to reflect the general facial complexion; it also goes against the TCM overall facial observation theory. As mentioned in [1], the color of different facial regions on one's skin may be judged inconsistently even by the same TCM expert.

But, using more or large facial regions might tend to approach general facial complexion and then would achieve better representation of one's overall facial complexion. In [14], 15 facial regions corresponding to TCM complexion-viscera are used as recognition samples, resulting in good facial color recognition accuracy. Although the facial region numbers may be large enough, the feature points' location algorithm still falls into complication.

(iii) The most important issue is that all the above literatures are essentially qualitative analysis, which only aims to classify the facial complexions or diseases into their respective categories. But yet, in another sense, the severity of disease or the degree of complexion has not been studied as ever (e.g., one case is Yang jaundice with glossy skin and Yin jaundice with lustreless skin with respect to yellow color; for another case, flushed face and flushed cheek, which can be treated as different proportion of red color in the facial face, will indicate excess heat syndrome and deficiency heat syndrome, resp.).

To alleviate those discussed issues, we develop a novel framework for six facial colors diagnosis, which is built with four chromaticity bases and luminance distribution based on patients' whole face. Although it is mainly designed for facial color diagnosis, it also could be applied to the quantitative analysis for facial complexion involving color and gloss degrees.

On one hand, these developed four chromaticity bases are related to four facial colors (normal, redden, bluish, and yellow colors) and would be generated from the respective facial color gamut clustering. Those bases are still considered to be the dominant colors of facial skin but are achieved by means of two steps of fuzzy clustering. With this process, we can obtain more reliable dominant color compared with the works [1, 13].

Moreover, the luminance distribution would split up all chromaticity bases down into certain subbases for better representing the degree of luminosity gradient of our chromaticity bases. With the separated chromaticity subbases, the derived feature representation for the remaining two facial colors (pale and darkish colors) could be more distinctive to discriminate between the other four colors than the previous approach [2], which is basically constructed without considering the effect of luminance distribution on the facial color diagnosis.

On the other hand, not as some previous studies used to do, we would not expect to segment the patient' face image into specified facial regions but to extract one's whole skin color as holistic representation. That is to say, the region segmentation procedure as done by the previous studies [1, 2, 12–14] could be bypassed in our framework. Nevertheless, for the improvement of classification performance, we also find that the combination of locally weighted region and global representation could achieve more significant improvement than only the local or global approaches.

In this regard, facial color distribution would be reliably estimated using our well-established holistic facial complexion representation. This might be more in accordance with TCM overall concept and diagnosis in practice. Furthermore, it would be possible to analyze both qualitative and quantitative issues through our proposed framework.

The remainder of the paper is organized as follows. Section 2 describes the novel facial complexion feature representation and how it is applied to quantitative analysis for both color and gloss degrees. Extensive experimental results and several improvements of the color classification are presented in Section 3. Finally, Section 4 draws some conclusions and future directions.

2. Methods

This section gives detailed descriptions of our developed facial complexion diagnosis chain, which is briefly summarized in Figure 1. As the pipeline shows, our framework is composed basically of four main stages as follows.

(1) At the beginning, four chromaticity bases and their luminance distribution will be constructed on their corresponding facial color images, that is, feature detection and construction stage illustrated in the pipeline.

(2) Then, the designed feature will quantize facial skin color to form complexion distribution on all collected facial images, called feature representation stage.

(3) At the learning stage, a Gaussian Kernel (or RBF kernel) Support Vector Machine (SVM) with the optimal parameters is employed to build the facial color model.

(4) In the final recognition stage, for any testing facial image, the learnt model will determine its facial color category.

In addition, beyond facial color recognition, we further make quantitative analysis of color and gloss degrees for each patient using our established feature representation. Subsequently, we present all procedures step by step.

2.1. Acquisition System and Data Description. The facial image acquisition system is the same as our previous one reported in [11]. This acquisition device mainly consists of annular LEDs and digital camera which are chosen with lots of trials and errors. More specially, because a digital image is produced highly based on the light source and camera, the achieved image might be color distorted, if no color correction was taken under nonstandard light source. In order to reduce the effects of light source, both illumination characteristics of light source and imaging characteristics of camera have been designed experimentally to obtain acceptable facial image. The illumination characteristics have been assessed and analyzed by studying various light sources, and, finally, the best light source with appropriate illumination characteristic (color temperature value is about 5600 K, Ra = 90) is determined as our light source. As for the camera,

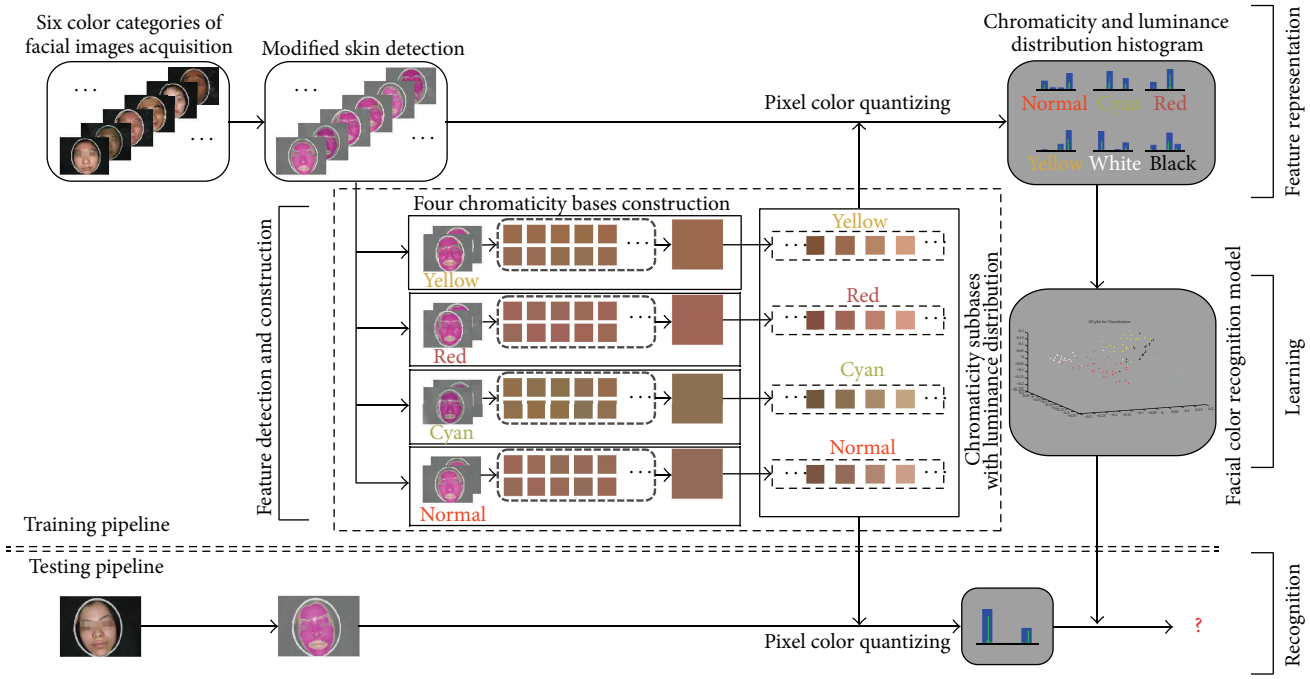


Figure 1: Pipeline of the proposed facial color classification framework.

the imaging characteristics of several available cameras at different modes were also compared and evaluated. Then the best camera (Canon PowerShot series S3 IS) on the best mode with white balance is set up. More details about our acquisition system could be found in [11].

After the development of facial acquisition system, different patients' facial images are obtained and diagnosed on the color corrected monitor by three senior TCM physicians, totally containing six color cases. Such facial images would be recorded with its corresponding color category if more than 2 physicians would make the same diagnosis result.

Consequently, our facial color dataset is built up with a total of 122 cases, containing 24 normal color images, 15 for bluish, 18 for reddened, 24 for yellow, 21 for pale, and 20 for darkish color images. This dataset seems a little small compared with the previous works. This is mainly because our collected facial images need to be filtered by TCM experts firstly if the images are unavailable and then diagnosed by senior TCM physicians. Thus only a small number of facial images are obtained, if the preliminary images are not enough. And, for further research, it is better to collect more data in the future. For color recognition, normal color would be tagged with normal, and all other morbid facial colors (bluish, reddened, yellow, pale, and darkish) would be tagged with its similar color spec named cyan, red, yellow, white, and black. These tags or labels are also illustrated in the framework of Figure 1. In the subsequent parts, for convenience, we would prefer to describe those six facial colors in TCM using their label names.

2.2. Skin Detection and Fine-Tuning. Different from the previous studies, our basic idea is to study the overall complexion condition of patients' face, which may avoid the influence of inaccurate or insufficient local region location and, instead, strengthen the expression of facial complexion in global situation. To meet this goal, an automatic and efficient skin detection approach which is published recently in [16] will be introduced. To put it simply, this approach is a fusion strategy that firstly utilizes a smoothed 2D histogram with I and By channel in log opponent chromaticity (LO) space, and then combines Gaussian Mixture Model (GMM) for modeling the threshold of skin-color distribution. Although there exist a large amount of skin-color detection solutions, we still adopt this fusion version due to two benefits: (1) it is able to cope with the variety of human skin colors even across different ethnic origin, which is especially fit to detect facial skin with different color variations in TCM; (2) it can be managed with low computational cost as no training stage is required.

Nonetheless, one case has been long-term existed in skin detection, which is known as the prone false skin detection of lip. This is presumably because the facial color gamut is highly overlapped with lip color gamut, leading to unexpected skin detection result, especially when skin detection approach is performed only with color feature. Such case is not expected to occur in our facial color recognition system because of its interference on original facial color distribution. Thus, we strongly hope to do a fine-tuning processing so as to crop the corrected skin and discard the mouth region. But accurate mouth segmentation algorithms may be bound up with high computational cost; it is desirable to consider the algorithm characterized by low computational cost and coarse segmentation, such as returning a simple rectangular or circular window on lip. In other words, we prefer to detect mouth region rather than segment the lip along the boundary.

Fortunately, the appearance of facial profile and components (eyes, mouth, and nose) is distinctive from

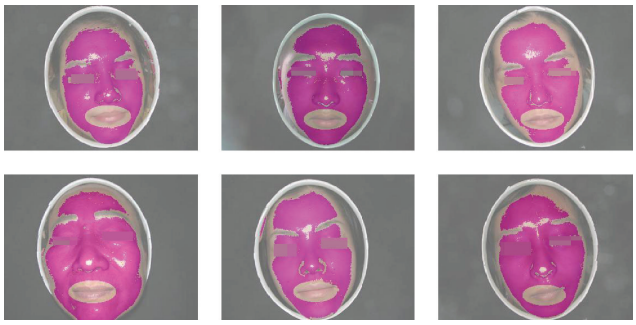


Figure 2: Six typical facial skin regions after skin detection and fine-tuning.

each other, arousing vast works on facial detection. Of these, one face detection approach, known as Viola and Jones' object detection framework [17], shows its excellent performance with low false negative rate and rapid face detection. The detector is firstly done by fast extracting a number of overcomplete haar-like features using integral image, across different scales and different spatial positions on image. Then AdaBoost learning algorithm will be used to select a low number of critical visual features from above large set of features. Finally, a cascade scheme for combining classifiers is built to quickly discard background regions of the image. Another extended work [18] introduces a novel set of rotated haar-like features, which significantly enriches original features and further improves the overall performance. Hence, in this paper, we expect to introduce this well-known framework for localizing facial mouth.

Now, the main issue is how to build a facial mouth detector using Viola and Jones' object detection framework. To be honest, it is time-consuming to prepare the training data and then train a cascade classifier for specific application. Thankfully, the OpenCV community shares a collection of public domain classifiers for facial processing which contains the facial mouth detector. All those trained classifiers are available in the haar cascades repository [19].

Since our facial images are all in frontal view, it is compatible to carry out the detection directly based on this public mouth detector. Then, from the mouth detection result, it should be noted that the discarded mouth region might also remove some skin colors around the lip. Nevertheless, it would not impact the estimation of overall facial color distribution due to such minority skin pixels. This is one reason why we prefer to perform mouth detection instead of accurate but time-consuming lip segmentation. Another reason to employ the Viola and Jones' detector is that it is extremely rapid and reliable, which has been extensively used in computer vision research. Some examples of skin detection and its fine-tuning results are illustrated in Figure 2.

After the skin fine-tuning stage, skin color denoising process would be done to remove the scattered spots (e.g., black moles) on the skin and a small minority of hairs covering the forehead, which is characterized by higher or lower luminance value. We simply employ the cumulative histogram to calculate the gray distribution and remove the top and bottom gray level satisfying certain proportions.

2.3. Chromaticity Bases with Luminance Distribution Construction. Once we got the skin region, we can turn to construct the chromaticity bases and luminance distribution. This is called the feature detection and construction stage. In this stage, three steps of the feature construction are performed: color space transformation, chromaticity bases construction, and final feature representation with luminance distribution.

Color Space Transformation. In order to realize the representation of chromaticity and luminance separately, the transformation from RGB color space to CIELAB color space is realized. CIELAB is a color-opponent space with L component for luminance and remaining components (a and b) representing the chromaticity. Since the CIELAB color space is made on the basis of CIEXYZ color space, we need firstly to transform the RGB color space to CIEXYZ color space [20]. The transformation formula is given as below follows:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4124 & 0.3576 & 0.1805 \\ 0.2126 & 0.7152 & 0.0722 \\ 0.0193 & 0.1192 & 0.9505 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \tag{1}$$

Then, the transformation from CIEXYZ color space to CIELAB color space is done by the following formulas:

$$\begin{aligned} L^{\star} &= 116 f\left(\frac{Y}{Y_n}\right) - 16 \\ a^{\star} &= 500\left[f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right] \\ b^{\star} &= 200\left[f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right], \end{aligned} \tag{2}$$

where

$$f(t) = \begin{cases} t^{1/3} & \text{if } t > \left(\frac{6}{29}\right)^3 \\ \frac{1}{3}\left(\frac{29}{6}\right)^2 t + \frac{4}{29} & \text{otherwise.} \end{cases} \tag{3}$$

Here, X_n, Y_n, and Z_n are the CIEXYZ tristimulus values of the reference white point. Often, its values are assumed as $X = 95.047$, $Y = 100$, and $Z = 108,883$ relative to CIE standard illuminant D65. Based on the color space transformation, all of skin pixels would be converted and applied to construct our feature representation.

Chromaticity Bases Construction. In this step, four chromaticity bases are constructed using two-level clustering on the CIELAB color space. The objective of establishing these four bases is to explore the basic chromaticity with respect to each facial color. Based upon basic chromaticity, it would be possible for us to assign the skin pixels to its closely measured distance. And, thus, final established feature representation could be regarded as the facial chromaticity ratio, which would be easy to distinguish among different facial colors. (e.g., cyan facial color would have a large number of skin pixels approaching cyan chromaticity but less number of pixels getting close to other chromaticity bases.)

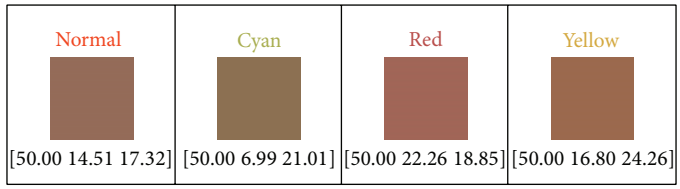


Figure 3: Four chromaticity bases constructed by the proposed approach.

This idea is mainly inspired from the bag-of-words method in computer vision [21]. And similar approach in facial diagnosis has been introduced by Zhang et al. [2]. Our constructed bases are similar but exactly not identical. In their approach, six centroids (similar as bases) are manually selected from the facial color gamut. But in our framework, we tend to automatically select four bases by clustering algorithm and expect to make full use of label information of given facial color category.

To be more specific, from the previous studies [1, 13], the dominant color of facial skin is treated as the greater number of cluster center by FCM. In other words, they assumed that the potential facial color in TCM appeared in the dominant color, but, exactly in practice, dominant color may fall into the minor number of cluster center as we stated before. Besides, TCM theory considers the facial color consisting of two kinds of colors: facial color and basic skin color [14]. In general, we expect to assume that facial chromaticity basis basically appeared in dominant color or subdominant color.

According to this assumption, the purpose of our chromaticity basis extraction framework is to find out facial color chromaticity regardless of it may lie in facial dominant or subdominant chromaticity. So, the major construction procedures of chromaticity bases can be summarized as two levels of clustering:

(i) *pixel-level clustering*: for each patient's facial image, the dominant and subdominant chromaticities of facial skin are extracted by FCM clustering; that is, cluster center number is set to 2; this step is the same as the one in the previous studies, but we further drive it to generate the color chromaticity base with second level clustering;

(ii) *chromaticity-level clustering*: based on dominant and subdominant chromaticity, the second clustering is performed; then, in this step, chromaticity base chooses the cluster center with a greater number of members by FCM; it should be noted that all bases are built under respective facial color categories. Thus, we will obtain four bases for four facial colors (normal, red, cyan, and yellow colors).

As seen in Figure 1, inside the feature detection and construction stage titled "four chromaticity bases construction," we demonstrate the procedures to generate four bases. It can be illustrated that all bases are constructed separately, so each basis will not be affected by other colors. In the first level clustering, it would be possible to collect multiple potential facial color chromaticities within dominant and subdominant chromaticities. This process makes sense because images with the same color class can extract similar color chromaticity in dominant or subdominant chromaticity, no matter which one is our expected chromaticity. Then, it would be reliable to choose the final cluster center with a greater number of members as our final color chromaticity basis in the second level clustering.

We also give those produced bases in Figure 3 their respective chromaticity values at the bottom. Values inside the square brackets represent *[L, a, and b]* (the component values in CIELAB color space). But the L component is meaningless here, just for color display only. Because only two chromaticity values in CIELAB color space are unable to produce color, so we simply set the luminance value as 50 on all chromaticity bases for pure illustration.

Luminance Distribution Construction. Using the above four chromaticity bases, four facial colors can be discriminated from each other. However, the remaining two facial colors, white and black, would fail to be recognized from others. This is reasonable mainly because their distribution of chromaticity gamut is highly overlapped with other four facial colors, leading to impossibly distinguished characteristic through the chromaticity only. Fortunately, the luminance characteristics of them are diametrically opposite, which is easy to represent by the L component of CIELAB color space. This is our basic motivation of building white and black colors feature representation.

In support of this goal, the luminance distribution should be constructed to integrate the bases we designed before. Recall that four chromaticity bases are constructed only by the a and b components of CIELAB color space, one conceivable way for constructing luminance distribution is by splitting up the basis into several subbases by spanning a range of luminance. By this method, final feature representation could be characterized by two advantages:

(i) for each chromaticity base, it can represent the chromaticity distribution of facial skin color, making four facial colors separable. And, moreover, we could make quantitative analysis for facial color degree;

(ii) for each subbase, the luminance distribution would be calculated for each chromaticity base, providing a solution to easily recognize white and black colors. It would also support us to make quantitative analysis for facial gloss degree.

The designed chromaticity subbases with luminance distribution have been shown in Figure 1. We simply set the range of luminance by default (L component value ranges from 0 to 100) and set the interval value of adjacent luminance as ten.

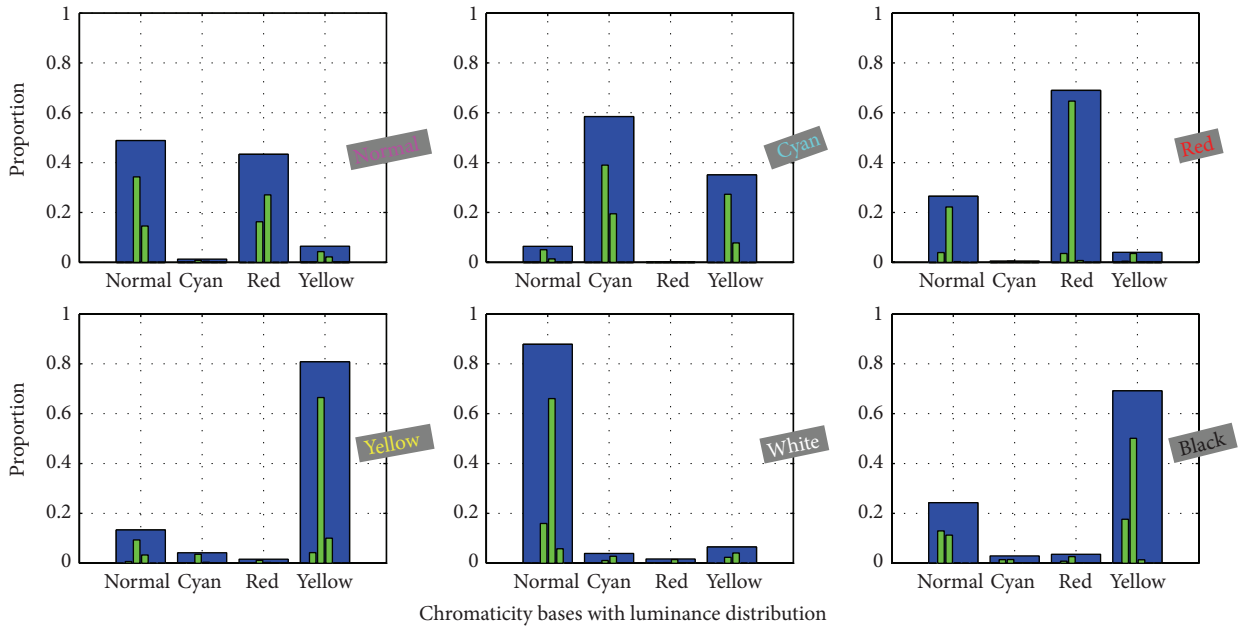


Figure 4: Six typical facial complexion histograms.

Therefore, totally a vector length of 40 features is obtained for each facial color image. Although this process is coarse at first, we will refine it to explore the best luminance distribution by experimental comparison.

2.4. Facial Complexion Representation. By the well-established subbases with both chromaticity distribution and luminance distribution, the facial skin pixel color assignment is performed on each image, thereby quantifying the facial complexion and building our feature representation. We still call the feature representation as facial complexion feature instead of facial color feature due to its expression ability for both color and gloss, so enabling us to make quantitative analysis for the facial complexion degree.

For every patient's image, its facial skin pixels are assigned to nearest subbases using color distance measurement in CIELAB color space. After the assignment, a complexion histogram is formed and would be used to measure for both chromaticity and luminance distribution. In order to represent the histogram as the facial complexion ratio distribution, and especially for quantitative complexion analysis, histogram normalization is carried out to achieve this purpose.

Figure 4 shows six facial color categories of established complexion histograms (which has been adjusted with the optimal luminance distribution). Four chromaticity bases are rendered as blue color, and subbases for luminance distribution are rendered as green color. In the complexion histogram, the highest blue peaks of chromaticity bases correspond to dominant color of facial skin, and subdominant color would be identified by the second highest blue peaks. For these subbases rendered with green color, it denotes the luminance distribution with respect to each chromaticity base. Such process would be more efficient to represent the facial complexion than without consideration of luminance distribution. Furthermore, seen from the histogram, white facial color would hold higher values in luminance distribution at each chromaticity base, while black facial color are always with lower luminance values. Essentially, the proposed feature is only involving color feature. We note that the texture and the shape features are widely used in image representation. But considering the facial color recognition, the texture and the shape features seem not as important as the color feature in the description of facial colors. Thus, in this framework, we put the texture and the shape features aside.

Benefiting from the proposed feature representation, colors of normal, cyan, red, and yellow which are separable on chromaticity could be easier discriminated from each other. Besides, white and black colors which are easily classified by luminance values could be correctly recognized. More important advantage from the combined chromaticity and luminance distribution is probably to analyze color and gloss degrees in quantitative way, given in later stages.

2.5. Facial Color Recognition. Before we discuss the facial color recognition model, it is noted that the proposed feature is unnecessary to be normalized along each feature dimension before modeling. This is not as we usually do when building a machine learning model. Because the constructed features are normalized histograms, denoting the color proportions on one's facial skin. All feature values have normalized from 0 to 1. Especially, we would not perform normalization process again, avoiding destructing the color proportions information.

In fact, facial color recognition is basically considered as a multiclass classification problem. There are various classifiers available for tackling this problem. So we have compared

SVM with several type classifiers (e.g., K-nearest Neighbor and Nave Bayes); the classification result shows that the SVM could achieve better performance than compared classifiers. This is the main reason why we prefer to adopt SVM for facial color recognition here.

It is easier to implement the SVM model with LIBSVM toolbox [22]. For multiclass problem, LIBSVM supports one-against-one strategy for multiclass SVM with building several binary classification models. In our paper, 6 facial color labels would produce 15 pairwise SVM models. And, finally, any test facial color will be assigned the maximum voting label, where each SVM model votes for one label. In addition, we choose a Gaussian kernel SVM with the optimal penalty factor C and width parameter γ in our experiment.

2.6. Facial Complexion Quantitative Analysis. The facial color recognition model is constructed for qualitative analysis of facial complexion, especially for facial color. But for patients belonging to particular facial color category, the severity or degree of disease is unknown for the previous studies. In addition, given a set of patients with specific facial color category, the relative degrees of facial color for patient have not been studied ever. These relative degrees mean that the patient's facial color or gloss degree is adjacent to someone. So in this paper, after facial color recognition stage, we also explore some experiments to make quantitative analysis based on our feature representation. Besides, this quantitative diagnosis has not been developed yet, we attempt to implement some unsupervised methods to do some preliminary trials.

To make quantitative analysis, we assume that facial color pixel numbers of specific color could represent the degree of color, and glossy skin is characterized by reflective, shiny pixels. Thus, after performing facial color recognition, two situations could be studied for quantitative analysis: color degree without luminance distribution and gloss degree without chromaticity distribution.

However, we should notice that our feature representation for facial complexion can be regarded as chromaticity distribution for four facial colors (normal, cyan, red, and yellow). For each chromaticity, it is splitting up by luminance distribution, making it possible to recognize black and white facial colors. Due to the characteristic of our feature, both white and black facial colors are neglected to conduct quantitative analysis for color degree. On the contrary, gloss degree will take all facial colors into consideration.

2.6.1. Color Degree of Quantitative Analysis. Color degree presented here is defined as the quantity of color pixels in one's facial skin. This definition makes sense because large quantities of color indicate obvious color reflection, and corresponding to the degree of pathological changes of five internal viscera.

Therefore, our method for color degree quantitative analysis is to build the scope of pixels numbers on the specific color. In other words, the ranking function of color degree is learnt. So for any patient, its color degree would be estimated by this ranking function.

To achieve each color ranking function, we firstly quantify facial skin pixels into a chromaticity distribution histogram without splitting up by luminance distribution, deriving a vector length of four features. The histogram feature indicates the proportions of each color on facial skin. Then given a set of the same facial color category of patients, their corresponding facial color proportion values in histograms are used to determine facial color degree ranking function. In this paper, we simply utilize normalization to quantify color degree into range from 0 to 1. Finally, for any patient whose predicted facial color belongs to given color, their facial color degree would be estimated by built ranking function. Besides, its relative color degree patients are also obtained with this ranking strategy.

2.6.2. Gloss Degree of Quantitative Analysis. Gloss degree is defined as the pixel quantity of higher luminance value; higher degree suggests glossy skin. This definition is raised by the characteristics of glossy skin mentioned before. With gloss degree, TCM experts could analyze the physical condition of patient. The facial skin of normal color patients would always be glossy; so it also provides complementary information to recognize normal facial color. For morbid facial color, higher gloss degree indicates mild illness and is easier to cure.

Similar to facial color degree, gloss degree is also quantified by a ranking function with normalization. But the gloss representation is slightly different due to its definition. Given all facial complexion histograms we used in recognition stage, the four bases describing chromaticity distribution will be merged. We sum over these bases and obtain a novel histogram only with luminance distribution. This is opposite to color degree which only describes chromaticity distribution.

To achieve gloss degree estimation, we should firstly compute gloss score for derived histogram feature. This is done by summing over all luminance values with weight values in luminance distribution. Those weights are selected by principle that higher luminance values are assigned to large weights. After computing gloss scores for all given patients, the gloss degree ranking function is built by normalized all gloss scores. Then, the gloss degree or relative gloss degrees could be estimated by ranking function for any patient.

3. Experiments and Discussions

The experimental images are collected by Shanghai University of Traditional Chinese Medicine. All images are diagnosed by three practitioners and labeled only if at least two of them made consistent conclusion. Eventually, we have established a small scale facial color data set, including 122 subjects with two facial color cases. One is normal or healthy facial color (24), another is morbid facial color with 5 categories: cyan (15), red (18), yellow (24), white (21), and black (20), respectively.

3.1. Experimental Setup. In our experiment, leave-one-out cross validation (CV) is used for evaluating the model

performance. To build the facial color recognition model, all the training images will be employed to construct the facial chromaticity bases and calculate complexion histograms for all images in the data set. Our images are collected by high definition camera (5-megapixel), which requires high computational cost to manage these facial images. So the height and width size of each facial image are resized by one-eighth of origin, deriving one-sixtieth reduced size of every facial image.

For the SVM with RBF kernel classifier, we use grid search to obtain the optimized parameters and select C and γ values from the range 2^{-8} to 2^{15} and 2^{-8} to 2^{8}, respectively. Then, in each fold of leave-one-out CV, we yield the best classification performance with these combined optimal parameters. Finally, the evaluation criteria are calculated based on the test results.

3.2. Evaluation Metrics. There are many criteria used for evaluating model performance, we consider several generalization measurements which is more reliable to assess multiclass problems [23].

Given the confusion matrix A and classes C, we can derive one-versus-all confusion matrix for each class; then the generalized precision and recall criteria formulas for each class are defined as follows:

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}, \qquad \text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}. \tag{4}$$

Here, TP_i are the number of true positives for C_i at confusion matrix element A_{ii}, TN_i are true negative numbers, FP_i indicate false positive counts, and FN_i indicate false negative counts, respectively.

Then, the overall F-measure score can be calculated with precision and recall, which could measure the overall classification quality of multiclass problems. Especially, it contains two types of F-measure, called microaverage and macroaverage. In addition, we also present confusion matrices visualization both for recall and precision.

Microaverage and Macroaverage F-Measure Scores. The F-measure score is essentially a weighted combination of precision and recall, which is defined as

$$F\ \text{score}_M = \frac{\left(\beta^2 + 1\right)\text{Precision}_M \text{Recall}_M}{\beta^2 \text{Precision}_M + \text{Recall}_M}, \tag{5}$$

where β is generally considered as 1. If the indice M indicates microaverage, the formula would derive the microaverage F-measure score; then its precision and recall are then computed as

$$\begin{aligned} \text{Precision}_{\text{micro}} &= \frac{\sum_{i=1}^{L} \text{TP}_i}{\sum_{i=1}^{L} \left(\text{TP}_i + \text{FP}_i\right)} \\ \text{Recall}_{\text{micro}} &= \frac{\sum_{i=1}^{L} \text{TP}_i}{\sum_{i=1}^{L} \left(\text{TP}_i + \text{FN}_i\right)}. \end{aligned} \tag{6}$$

In microaverage, it can be seen that precision and recall are calculated over all samples. So it gives equal weight to every sample, which is useful to measure the performance on the common numbers of classes.

If the indice M indicates macroaverage, we obtain the macroaverage F-measure score whose precision and recall are computed in another way as below:

$$\begin{aligned} \text{Precision}_{\text{macro}} &= \frac{1}{L}\sum_{i=1}^{L} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \\ \text{Recall}_{\text{macro}} &= \frac{1}{L}\sum_{i=1}^{L} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}. \end{aligned} \tag{7}$$

From the above formulas, we can figure out that the macroaverage is the harmonic average across each class, which would give equal weight to every class regardless of the class numbers. In this way, the macroaverage F-measure score could evaluate the performance on rare numbers of classes. So the microaverage and macroaverage are complementary to each other, and both of them are informative for performance evaluation.

Confusion Matrix. Besides, we introduce two confusion matrices for recall and precision. The reasons employ two metrics of confusion matrix for facial color recognition model are as follows:

(i) the confusion matrix for recall could clearly show the proportion distribution of predicted facial colors with respect to each actual facial color, which illustrates the model performance for each actual facial color;

(ii) while the confusion matrix for precision could demonstrate the proportion distribution of actual facial colors with respect to each predicted facial color, which measures the quality of each predicted facial color.

In the following, we perform extensive experiments to study the performance of our proposed facial color framework. Besides, not only will we present both qualitative and quantitative experimental results, but also some more details about the improvement of our framework are explored, which yield excellent performance compared with the previous methods.

3.3. Preliminary Results on Qualitative Analysis. In this section, some preliminary results in qualitative aspect are produced and compared using above metrics. It should be noted that qualitative analysis mentioned here is referred to as facial color classification problems. The quantitative analysis is considered as the severity or degree of one's facial complexion, which will be presented later.

Each Facial Color Category Classification Performance. After performing leave-one-out CV, confusion matrices for recall and precision are firstly calculated, as shown in Figure 5. The row of confusion matrix means the actual class in database, and the column means the predicted class by our model. Each table cell indicates the ratio value according to specific metrics, and, the darker the table cell, the higher the ratio. Left

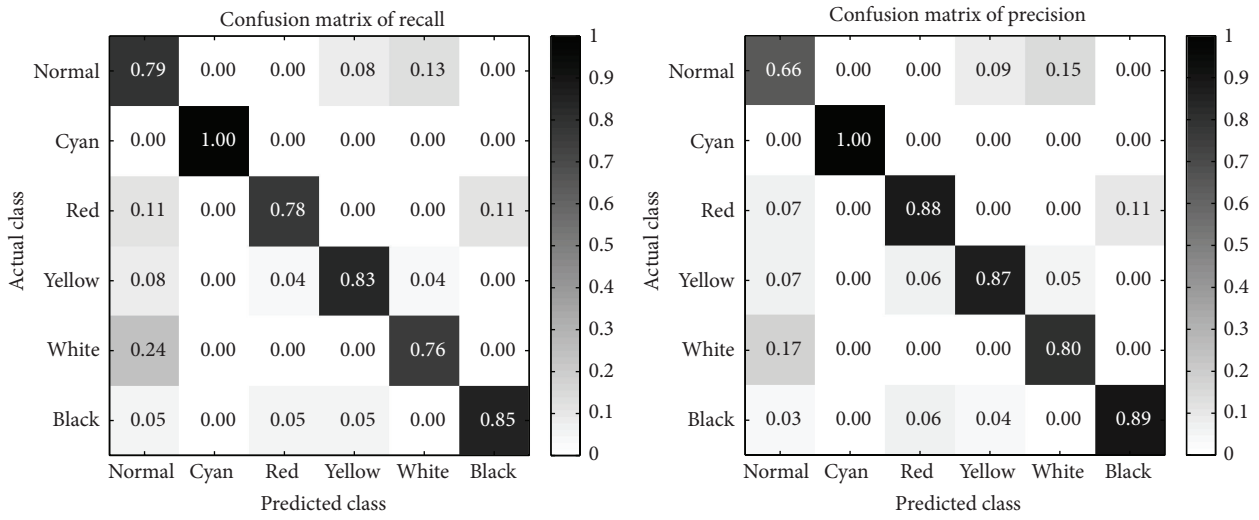


Figure 5: The confusion matrices produced by our model.

confusion matrix is computed for recall metrics and right one for precision metrics. It should be noted that the confusion matrix for recall will only make sense along the column, while precision makes sense along the confusion matrix table row.

According to the normal and white rows on confusion matrix for recall, we observe that normal facial color is slightly prone to be predicted as white and vice versa. Red facial color images have some cases misclassified as normal or black facial color. And yellow and black facial colors achieve better recall values than normal and red.

However, all cyan facial color images have been recognized perfectly. This result for cyan is presumably because its chromaticity base is highly distinctive to other colors which might be easier to be discriminated. But another potential reason may give rise to this outstanding result only for cyan facial color is as follows: cyan color image numbers of our database are not enough (only fifteen images). So the classification performance of cyan facial color on a larger amount of images is still unknown for us.

Observing the right confusion matrix for precision along the row, we note that the worst quality of predicted classes is shown on normal facial color. Predicted normal facial colors are almost misclassified across all other colors, especially those inclining to white facial color. But our model achieves superior precision for other predicted colors.

From Figure 5, we discover that normal facial color is the most difficult class to be recognized. Indeed, in the TCM theory, normal facial color is defined as hybrid colors composed of red and yellow chromaticities with higher luminance. So it is reasonable to be misclassified as white, red, or yellow for normal color, as shown in our confusion matrix results. This also makes us consider how to build the normal color chromaticity base, but we have not studied it in this paper, we hope to put it further in our future study.

From another perspective of each facial color classification performance, it is also considerable to prove the argument we highlighted before that our bases is reliable to extract the dominant color from facial skin, with two-level clustering method. But the chromaticity bases are constructed only with four facial colors. To this end, we ignore white and black facial colors and use only four chromaticity bases histogram to performance classification. For classification, the classifier SVM is not adopted; four facial colors will be predicted by the dominant color or peak bin in histogram.

For comparison, another chromaticity bases are constructed without two-level clustering, just clustered by FCM once as in [1, 13] and then calculates mean value of large cluster as dominant color. We show confusion matrix counts comparison between our proposed and the previous methods in Figure 6, each cell record the predicted facial color counts with respect to actual class. Compared with these two tables, the classes normal, cyan, and yellow are quite competitive between two methods. But the previous approach incorrectly predicts all red facial colors as normal class. This happens presumably because red facial color images in our database are hard to distinguish from normal facial color. And then the previous method may probably derive similar chromaticity base between normal and red facial colors. But our proposed dominant color extraction for constructing chromaticity base could avoid this situation with two levels clustering, due to the full use of label information of each class. Moreover, with such simple classification criteria and lower dimensional features, our total accuracy for four facial colors recognition reaches 87.65%, but the previous method only achieves 67.90%.

Overall Facial Colors Classification Performance. We then present the overall facial color classification performance of our proposed complexion feature representation. For the purpose of validating the distinctive feature representation, we proposed several previous features [1, 12, 13] for facial color recognition that are implemented in this paper. All features

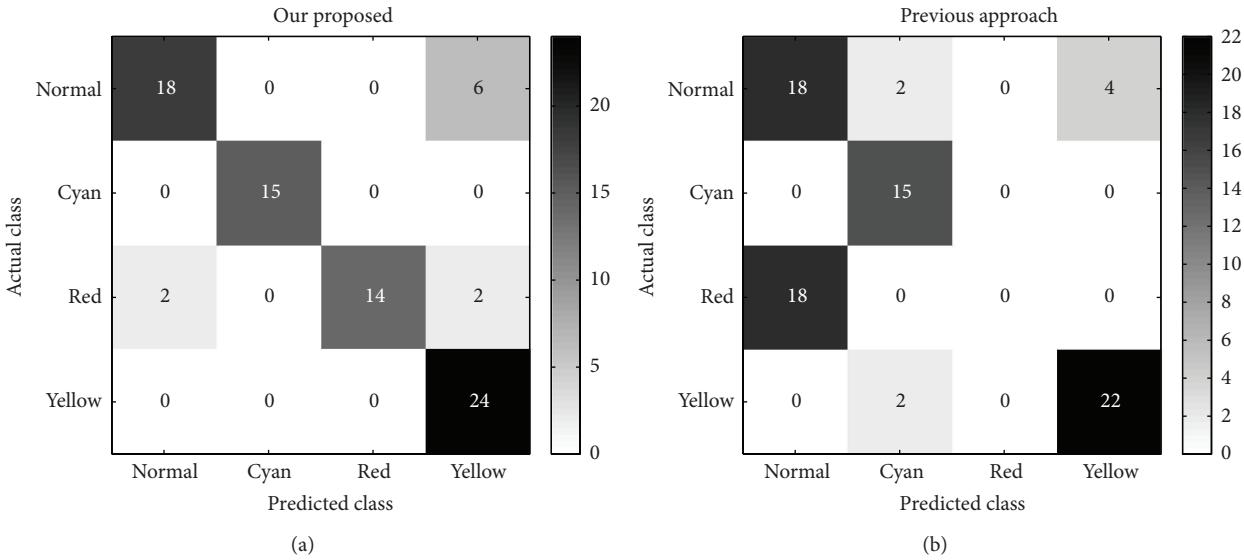


Figure 6: Comparisons of dominant color extraction methods on four facial colors.

Table 1: Comparisons of overall performance by using all methods.

Methods	Whole face			
	$\text{Pre}_{\text{macro}}$	$\text{Rec}_{\text{macro}}$	F_{macro}	F_{micro}
Li et al.'s [1]	0.7306	0.7200	0.7252	0.7131
Liu et al.'s [12]	0.7967	0.8001	0.7984	0.7869
Wang et al.'s [13]	0.7382	0.7306	0.7344	0.7131
Our proposed method	**0.8491**	**0.8358**	**0.8424**	**0.8279**

are extracted on whole facial skin pixels. In this process for compared features, we expect to only explore the facial color representation ability regardless of the influence of other factors, such as local regions versus whole face.

On the implementation of [1], six facial color RGB values are derived and used to measure the similarity of each test facial color, which could be similarly treated as the color distribution feature. But, slightly different from their classification strategy, we implement the distance metrics using nearest neighbor instead of radius scope, due to its higher performance with our database. In [13], dominant color of facial skin pixels is extracted in RGB color space and transformed into other five color spaces; finally, 18 features are obtained for whole face. And, for [12], average RGB values of whole facial pixels are used as color feature. Sincerely, Wu et al. [14] may achieve superior accuracy for facial color recognition problem, but the operation to collect skin color of patients' upper arm is impractical for our current database, so we do not implement it for comparison.

We list our comparison results in Table 1, where "F_{micro}," "F_{macro}," "$\text{Pre}_{\text{macro}}$," and "$\text{Rec}_{\text{macro}}$" represent microaverage F-measure score, macroaverage for F-measure score, precision, and recall. We do not present microaverage precision and recall because they are numerically equal to microaverage F-measure score according to their definitions. From classification evaluation, we find that our complexion representation could achieve highest performance. Even though our existing facial complexion representation could perform better than the previous methods, we still consider some possible improvements for our framework. So in the following section, we discuss some adjustable parameters and regional fusion strategies to further improve our facial color diagnosis framework.

3.4. Improvements on Facial Color Classification. Above results are preliminary since some factors are still without consideration, so we would like to explore more details about our facial complexion representation, which could not only reduce the dimension feature representation but also improve our model recognition rate.

The Effect of Luminance Distribution. Some previous qualitative results are obtained by setting the interval of adjacent luminance as ten by default. Since the luminance value of L component in CIELAB color space ranges from 0 to 100, so a vector length of 40 features is formed for each facial complexion feature representation. But it is considerable to study how many intervals to quantify one's facial luminance would provide higher discrimination; so the finer quantization of the luminance distribution is selected by experiment comparison.

From the preliminary parameters setting of luminance distribution spacing, we have found that extreme luminance value of facial complexion histogram is nonexistent. It happens because we have conducted the preprocessing step to remove spots (e.g., black moles) on facial skin. Thus, the refined luminance range is selected from value 25 to 95 based on the observation of final histograms.

Then the interval of adjacent luminance is determined experimentally. In Figure 7, we show the classification accuracy results over luminance intervals from 1 to 35, deriving feature dimensions from 12 to 284. The experimental intervals

Table 2: Performance comparisons using different region strategies.

Region strategies	$\text{Pre}_{\text{macro}}$	$\text{Rec}_{\text{macro}}$	F_{macro}	F_{micro}
Left cheek	0.8219	0.8274	0.8247	0.8115
Right cheek	0.8286	0.8345	0.8315	0.8197
Forehead	0.6146	0.6254	0.6200	0.5984
Nose	0.6478	0.6401	0.6439	0.6230
Jaw	0.7115	0.7327	0.7220	0.7049
Whole face (no fusion)	0.8491	0.8358	0.8424	0.8279
Classifier level fusion	0.8648	0.8688	0.8668	0.8525
Feature level fusion	**0.8758**	**0.8798**	**0.8778**	**0.8689**

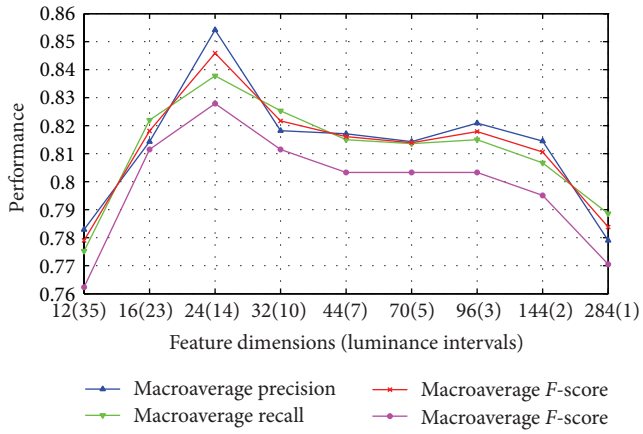


Figure 7: Classification accuracy with different luminance distribution intervals.

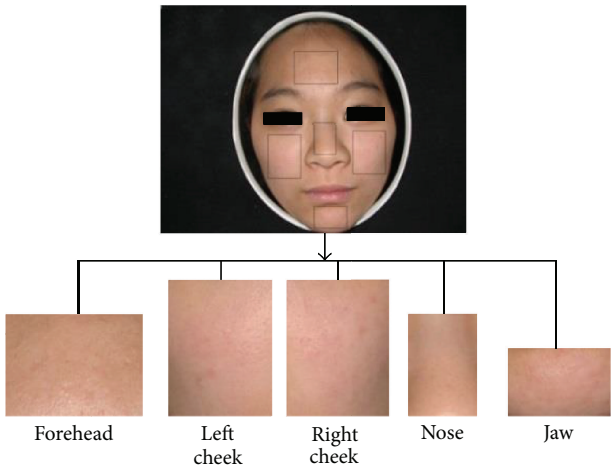


Figure 8: An example of facial image segmentation.

are selected if the luminance range can be divisible. In addition, interval values 23 and 3 are also performed due to some large spacing at luminance intervals or feature dimensions. This selection of feature dimensions is not uniform but rational because the main effect of luminance distribution for facial color classification is the luminance interval number. For easy calculation and even intervals, we take the given intervals as our testing experiment. In summary, the best classification performance is achieved when luminance interval is set as 14. So totally, only 24 feature dimensions and luminance ranging from 25 to 95 are formed to represent facial complexion and build recognition model. We also find that extreme luminance intervals would achieve inferior classification performance in our experiment.

Holistic Representation with Weighted Local Spatiality. Sincerely, studying on the local regions is reasonable for facial color. According to TCM experts, color recognition could behave well just on some local regions (e.g., cheek regions). To explore the facial color expression ability of different local regions, we start to analyze facial color recognition performance on different local regions.

For making experiments on this case, we simply do the local regions segmentation manually, as shown in Figure 8. We partition the whole face into five regions similarly as the previous works [1, 2, 12, 13]. Then each region is used to extract specific features based on aforementioned experiment (total 24 features used) and perform classification separately. Finally, we list the classification results in Table 2. From the classification performance on different local regions, we find that cheek regions could achieve more superior results than other regions. This result is consistent with TCM experts' experiences that cheek regions provide more facial color information than others. In addition, we also list our whole face result here. It achieves highest performance than any local region strategies for classification. This also proves our argument that whole facial color representation would be better than only local regions.

Therefore, different regions contribute different facial color information, it makes us think that spatial distribution information might be useful to represent facial color. Considering our previous feature representation, two levels of information have been described on whole face: chromaticity-level and luminance-level distribution for the histogram. But our features for facial color have not yet contained information about regional level (or spatial level) on different local regions. To this purpose, the regional level color distribution is built to achieve a novel holistic representation. So we need additional stage to segment local regions and rebuild our facial complexion representation.

Our facial images are all frontally scanned with small pitch; so for avoiding complicated local region segmentation process, a simple partition procedure based on the previous

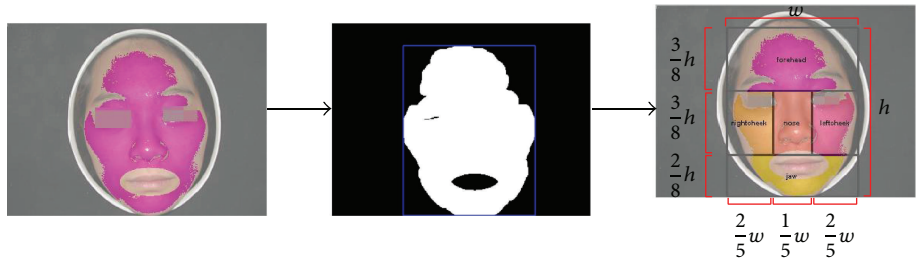


Figure 9: The flowchart of local regions segmentation.

skin detection results is used to produce five regions. We firstly utilize morphological close operation to connect separate regions of previous skin detection results. Then surrounding rectangle is located from skin pixel boundary. After that, each region boundary is determined by self-defined proportions according to each facial skin scope. As seen in Figure 9, we illustrate the flowchart of local regions segmentation. Although the local region segmentation method is simple, it still works to improve recognition rate.

According to classification performance of each region in Table 2, we consider assigning five regions with different weights, higher performance, and larger weights. In our experiment, left and right cheeks are set as 0.3, forehead and nose are 0.1, and jaw is 0.2. Our weights are selected only by simple observation of each region classification performance, without optimization experiment. So those weights are probably not the best choice. Despite that, it still obtains an improvement compared with our preliminary results.

After local region weights assignment, we look for two levels of fusion strategies to perform facial color recognition: feature level fusion and classifier level fusion. To achieve feature level fusion, each region complexion histogram is constructed with its respective weight. Then regional histograms are concatenated to a holistic histogram to represent facial complexion. After histogram normalization, it is possible to perform model training and test with leave-one-out CV. For classifier level fusion, each region builds its complexion histogram and performs model training and test separately. Then the weights are used to weight vote of all predicted colors. The highest vote is selected as final facial colors.

The results of these fusion strategies are listed in Table 2; it clearly shows that fusion of local regions would achieve higher performance than any local regions and whole face without weighted local spatiality. Of these, feature level fusion would achieve highest performance. So this suggests that for facial color recognition problem, more than color-level issues should be explored, but the spatiality-level distribution of skin color should be studied by experts in order to obtain good recognition rate.

After the improvement of classification performance, we are supposed to analyze new performance on each facial color category. In Figure 10, what we observed is that the improved model could achieve higher recognition rate on red, yellow, white, and black colors. Hence, it shows why our overall classification results could reach higher performance. However, normal facial color still suffers inferior performance on recall metric. It informs us that normal color needs to be studied in-depth. In the future, how to improve the normal color recognition performance would be the focus of our researches.

3.5. Quantitative Results. Qualitative results show the facial color classification performance of built model. In this section, we demonstrate some quantitative analysis results for complexion degree including the color degree and the gloss degree.

Based on the features we improved in the previous section, from a total of 120 feature dimensions with five local facial regions information, we weighted each region as the previous task for color classification. Then we sum the feature values over all regions, deriving the overall facial color features which would be utilized for estimating the color and gloss degree later.

For a given set of facial color database, we can learn the ranking function from data in an unsupervised way. That is simply implemented by normalizing each chromaticity base for color degree estimation and weighted summing over all luminance values for gloss degree.

Once we obtain all degrees, we could also achieve the ranking distribution to show relative degrees of each patient for both facial color and gloss degrees. Therefore, in order to illustrate the learnt ranking distribution, we present a two dimensional space visualization in Figure 11. These four visualization figures are a picture subset of our facial database of four facial colors, respectively, which are projected in a two-dimensional space corresponding to the facial color and gloss dimensions.

For each complexion ranking distribution (including facial color and gloss degrees), it demonstrates a few of useful knowledge for quantitative analysis:

(i) the facial pictures are prone to larger facial color degree values indicate one's dramatic color reflection, hinting critical pathological changes of five internal viscera; the additional knowledge of this case would indicate easier facial color recognition, which is also validated in the previous experiments;

(ii) pictures close to lower color degree values denote inapparent color reflection, leading to difficulty of correct recognition; in our experiments, these pictures are always incorrectly predicted by other facial

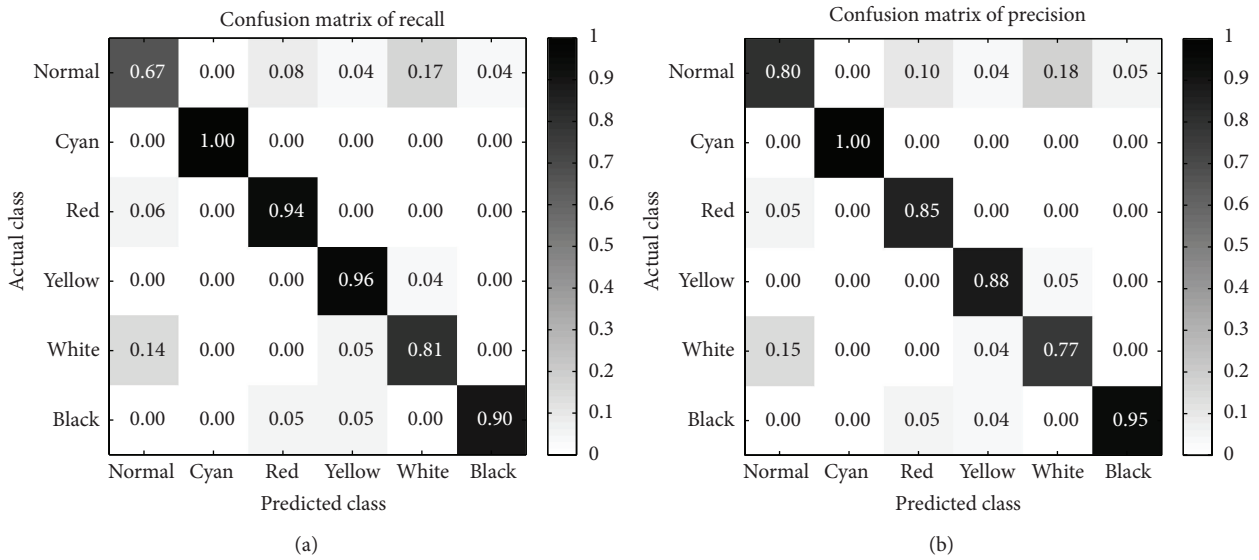


Figure 10: The confusion matrices produced by our improved model.

colors, despite the improvement of our feature representation and model;

(iii) in addition, based on relative degrees, pictures closed in facial color degree should have similar color category distribution in facial skin; similar pictures in color degrees also represent similar difficulty levels for facial color recognition task;

(iv) considering the pictures distribution along gloss degree, larger degree values show the severity of specific facial color; especially for the yellow color, Yang jaundice patients would have glossy skin and lustreless skin would be reflected on the face of Yin jaundice patients. So higher gloss degree can be diagnosed as Yang jaundice, while lower gloss degree can be diagnosed as Yin jaundice; so it will be meaningful if the gloss degree distribution is well established;

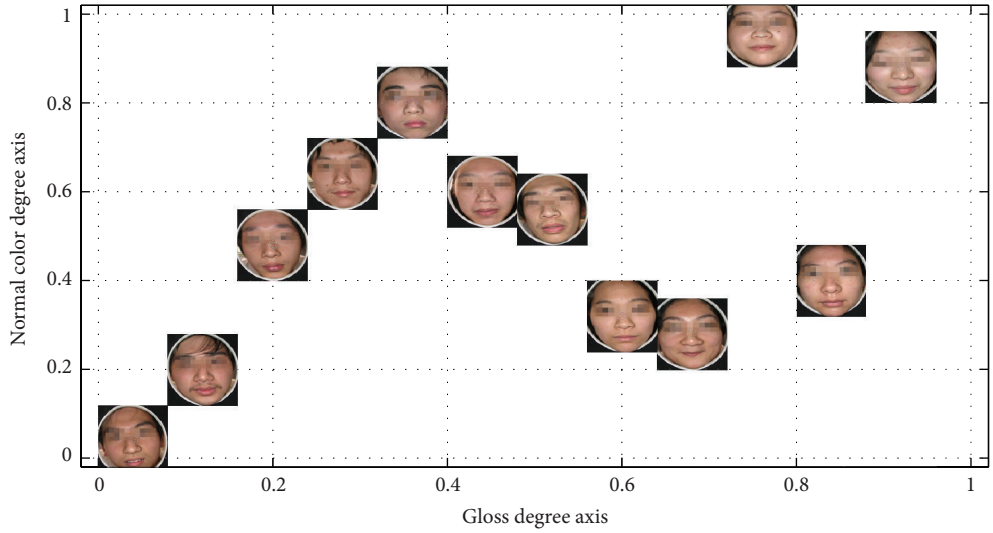
(v) making a general analysis on two-dimensional ranking distribution, we conclude that with both higher facial color and gloss degrees, patient would be healthier for normal facial color and mild illness for morbid facial color; on the contrary, lower degrees denote serious illness and hardness to be cured.

In general, we can expect that with ranking distribution, we could firstly make some preliminary analysis for classification. This may be not only helpful to visualize feature representation but also contribute to analyze the reason of some incorrect classification results for designed feature. This is validated by our classification experiments, which incorrectly predicted that facial color is always with lower color degree values.

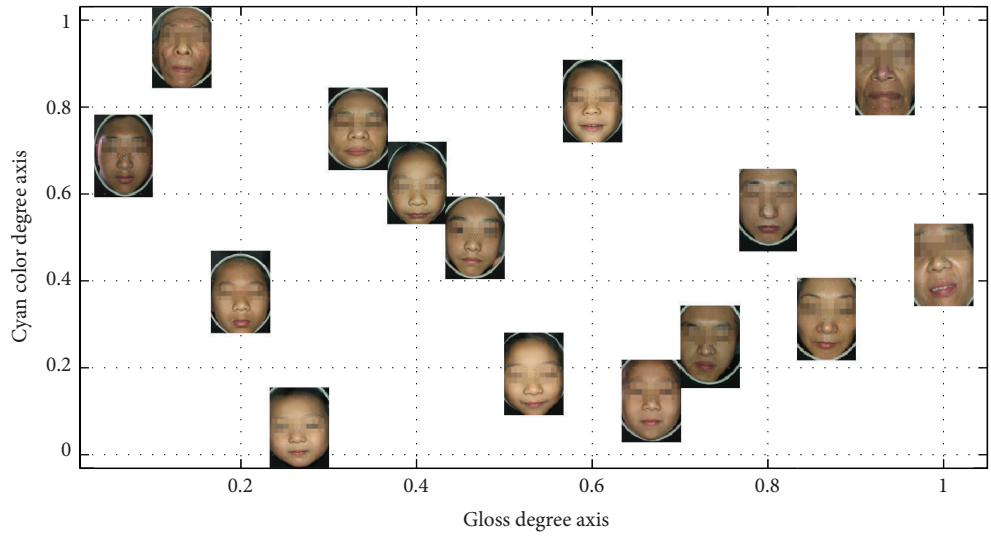
Moreover, for quantitative analysis, it provides both facial color and gloss degrees ranking function. In other words, we could make an in-depth analysis for each patient, examining detailed color degrees and pathological changes of five internal viscera. Sincerely, above quantitative analysis in different aspects has not been fully proved yet by TCM experts, due to unmentioned issue in the previous studies. So it is hard for us to claim that we provide authentic quantitative analysis for facial complexion. But beyond facial color recognition issue, we still make some preliminary researches on quantitative perspective. In the end, some providable quantitative conclusions are derived for deeply studying the degrees of facial color and gloss. Besides, concluded quantitative results are correlative to our previous qualitative results, interpreting why some patients are prone to be incorrectly predicted as other facial color categories.

4. Conclusions and Future Directions

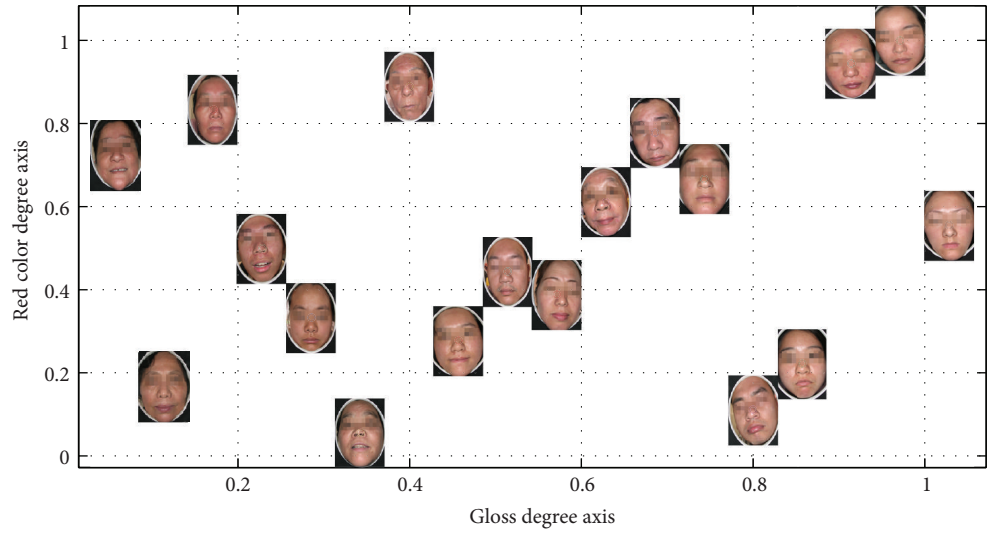
In this paper, a novel feature representation for facial complexion (including facial color and gloss) is proposed and discussed. Based on the proposed feature, we perform extensive experiments to validate and improve its performance for facial color classification, addressed as qualitative analysis in our perspective. Both each and overall facial color classification performance are studied and discussed in our experiments. Considering the effect of parameters, the optimal luminance distribution is obtained. Furthermore, we compare the facial color information on five different local regions, respectively, and then develop improved feature representation of all hybrid regions with different weights. This result tells us that not only the chromaticity-level and luminance-level information are important, but also the spatiality-level information is useful for facial color classification. In addition, we prove that the dominant color of facial color would be extracted more reliably with our two-level clustering method. Eventually we achieve significantly improved classification performances on facial color problems.



(a)

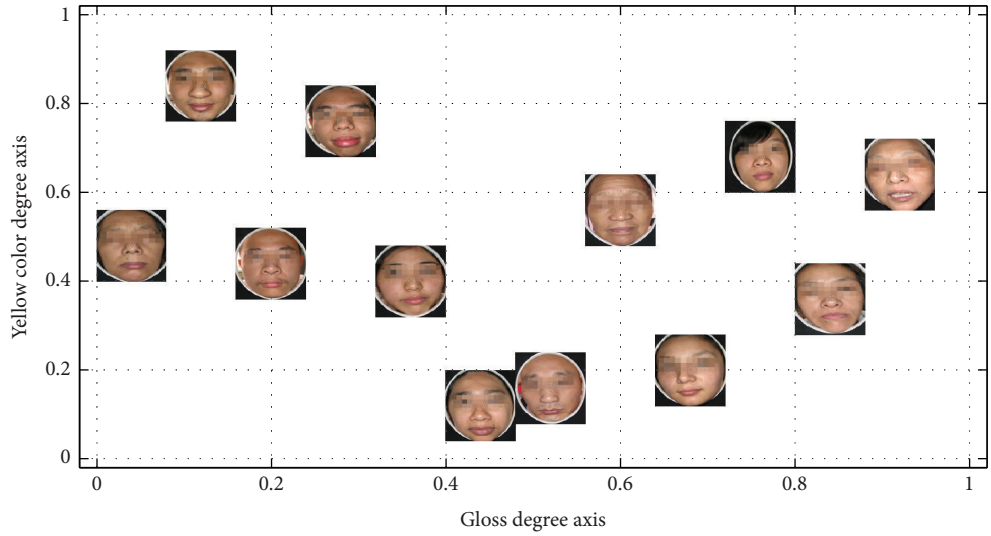


(b)



(c)

Figure 11: Continued.



(d)

Figure 11: The ranking distribution according to facial color degree and gloss degree.

For further researches, we also produce some quantitative analysis results for complexion degree. Although the presented ranking distribution has not been studied further on color and gloss aspects, it still shows highly correlative properties with the proposed feature for facial color classification. This may provide us with an analytical method to consider falsely predicted facial color, so it would support us to redesign a more distinctive feature representation. Besides, some preliminary quantitative conclusions are derived for further research in facial complexion diagnosis problem. However, some issues are still remaining to be addressed through current facial complexion diagnosis framework as follows.

(i) Although our facial image is produced under well-designed light source and camera, digital image still suffers from device-dependent color space rendering. In other words, the color information of an image is also dependent on the imaging characteristics of specific camera; it is almost impossible to be solved only by the adjustment of camera mode. Hence, more accurate color correction is necessary to be done for rendering color information into device-independent color space, as given in [24].

(ii) In our framework, normal facial color always gets into trouble for correct recognition due to its hybrid facial colors property shown previously. Therefore, how to develop an appropriate feature representation for normal facial color would be studied in the future.

(iii) We should also notice that the quantitative analysis presented here is preliminary in an unsupervised way. Thus, in the future work, we expect to require TCM experts to provide complexion degree ranking information of existing database, both on the color and gloss perspectives. With this prior knowledge, we could take some more outstanding supervised techniques to build complexion ranking distribution. Afterwards, more reliable quantitative analyses may be carried out on each patient.

(iv) What has been proved in this paper is that our complexion feature representation would be more distinctive than the previous features for facial color classification problem. However, the facial gloss classification performance has not been presented here. Hence, we will study the proposed complexion feature representation for gloss classification later.

(v) All experiments in this paper are performed in the data set of 122 facial color images. Although our framework is effective and available in this small scale database, the performance in larger scale facial color database or in actual world is still unknown. So further experiments would be expected in future works.

Conflict of Interests

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

Acknowledgments

This work was supported by the Natural Science Foundation of China under Grant nos. 61105053, 61273305, and 81373555, as well as the Fundamental Research Funds for the Central Universities.

References

[1] X. Li, F. Li, Y. Wang, and P. Qian, "Computer-aided disease diagnosis system in TCM based on facial image analysis,"

International Journal of Functional Informatics and Personalised Medicine, vol. 2, no. 3, pp. 303–314.

[2] B. Zhang, X. Wang, F. Karray, Z. Yang, and D. Zhang, "Computerized facial diagnosis using both color and texture features," *Information Sciences*, pp. 49–59, 2013.

[3] "Inspection of complexion in TCM," http://tcmdiscovery.com.

[4] G. Z. Li, S. Sun, M. You, Y. L. Wang, and G. P. Liu, "Inquiry diagnosis of coronary heart disease in Chinese medicine based on symptom-syndrome interactions," *Chinese Medicine*, vol. 7, article 9, 2012.

[5] M. You, R.-W. Zhao, G. Z. Li, and X. Hu, "MAPLSC: a novel multi-class classifier for medical diagnosis," *International Journal of Data Mining and Bioinformatics*, vol. 5, no. 4, pp. 383–401, 2011.

[6] M. Shi, G. Li, and F. Li, "C2G2FSnake: automatic tongue image segmentation utilizing prior knowledge," *Science China Information Sciences*, vol. 56, no. 9, pp. 1–14, 2011.

[7] X. Wang and G. Z. Li, "Multi-label learning via random label selection for protein subcellular multi-locations prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 2, pp. 436–446, 2013.

[8] G. Zhang, J. Yin, Z. Li, X. Su, G. Li, and H. Zhang, "Automated skin biopsy histopathological image annotation using multi-instance representation and learning," *BMC Medical Genomics*, vol. 6, supplement 3, p. S10, 2013.

[9] H. Shao, G. Li, G. Liu, and Y. Wang, "Symptom selection for multi-label data of inquiry diagnosis in traditional chinese medicine," *Science China Information Sciences*, vol. 56, no. 5, pp. 1–13, 2013.

[10] G.-Z. Li, S. Yan, M. You, S. Sun, and A. Ou, "Intelligent zheng classification of hypertension depending on ml-knn and feature fusion," *Evidence-Based Complementary and Alternative Medicine*, vol. 2012, Article ID 837245, 5 pages, 2012.

[11] F. Li, C. Zhao, Z. Xia, Y. Wang, X. Zhou, and G. Z. Li, "Computer-assisted lip diagnosis on traditional chinese medicine using multi-class support vector machines," *BMC Complementary and Alternative Medicine*, vol. 12, no. 1, pp. 1–13, 2012.

[12] M. Liu and Z. Guo, "Hepatitis diagnosis using facial color image," in *Medical Biometrics*, D. Zhang, Ed., vol. 4901 of *Lecture Notes in Computer Science*, pp. 160–167, Springer, Berlin, Germany, 2007.

[13] X. Wang, B. Zhang, Z. Guo, and D. Zhang, "Facial image medical analysis system using quantitative chromatic feature," *Expert Systems With Applications*, vol. 40, no. 9, pp. 3738–3746, 2013.

[14] T. Wu, B. Bai, F. Sun, C. Zhou, and P. Wang, "Study on complexion recognition in TCM," in *Proceedings of the International Conference on Computer and Communication Technologies in Agriculture Engineering (CCTAE'10)*, pp. 487–490, Chengdu, China, June 2010.

[15] R. Zhou, F. F. Li, Y. Q. Wang, X. Y. Zheng, R. W. Zhao, and G. Z. Li, "Application of PCA and LDA methods on gloss recognition research in TCM complexion inspection," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW '10)*, pp. 666–669, Hong Kong, December 2010.

[16] W. R. Tan, C. S. Chan, P. Yogarajah, and J. Condell, "A fusion approach for efficient human skin detection," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 138–147, 2012.

[17] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[18] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *Proceedings of the International Conference on Image Processing (ICIP '02)*, pp. I900–I903, September 2002.

[19] "Haar cascades repository," http://alereimondo.no-ip.org/OpenCV/34.

[20] M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta, "A standard default color space for the internet—srgb," 1996, http://www.w3.org/Graphics/Color/sRGB.html.

[21] F. F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 524–531, June 2005.

[22] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[23] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427–437, 2009.

[24] X. Wang and D. Zhang, "An optimized tongue image color correction scheme," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 6, pp. 1355–1364, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 641469, 9 pages
http://dx.doi.org/10.1155/2014/641469

Research Article

Computational Prediction of Protein Function Based on Weighted Mapping of Domains and GO Terms

Zhixia Teng,[1,2] Maozu Guo,[1] Qiguo Dai,[1] Chunyu Wang,[1] Jin Li,[1,3] and Xiaoyan Liu[1]

[1] *Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001, China*
[2] *Department of Information Management and Information System, Northeast Forestry University, Harbin 150001, China*
[3] *Department of Statistical Genetics, Harbin Medical University, Harbin 150001, China*

Correspondence should be addressed to Zhixia Teng; teng_zhixia@hit.edu.cn and Maozu Guo; maozuguo@hit.edu.cn

Received 21 December 2013; Accepted 12 March 2014; Published 23 April 2014

Academic Editor: Xing-Ming Zhao

In this paper, we propose a novel method, SeekFun, to predict protein function based on weighted mapping of domains and GO terms. Firstly, a weighted mapping of domains and GO terms is constructed according to GO annotations and domain composition of the proteins. The association strength between domain and GO term is weighted by symmetrical conditional probability. Secondly, the mapping is extended along the true paths of the terms based on GO hierarchy. Finally, the terms associated with resident domains are transferred to host protein and real annotations of the host protein are determined by association strengths. Our careful comparisons demonstrate that SeekFun outperforms the concerned methods on most occasions. SeekFun provides a flexible and effective way for protein function prediction. It benefits from the well-constructed mapping of domains and GO terms, as well as the reasonable strategy for inferring annotations of protein from those of its domains.

1. Introduction

More and more sequences of proteins are available due to the advanced sequencing technologies, but the biological roles and functions of the proteins are hardly known. As reported by [1], only less than one percent of proteins have been functionally characterized by experiments. In other words, protein sequencing is faster than annotating protein. To fill this gap, a large number of computational methods have been developed to predict protein functions. These methods exploit biological information including amino acid sequence [2–9], genomic context [10–14], protein interaction networks [15–17], protein structure [18–23], microarray [24], and literate to predict protein functions [25, 26]. However, the newly sequenced proteins are often poor in other biological information except the amino acid sequences. Thus, the development of the sequence-based method is crucial and useful for directing further experimental work.

In the past few years, several sequence-based methods [2–9] have been proposed to infer protein functions. These methods annotated the protein with the representative annotations of its homologues. Intuitively, these methods are also called homology-based methods. Usually, the homology-based methods include two stages: searching homologues through BLAST or PSI-BLAST and selecting representative Gene Ontology (GO) terms from annotations of homologues of the unannotated protein. More specifically, Goblet [2] determined the homologues by a predefined threshold of BLAST *e*-value and annotated the unannotated protein with the GO terms of its homologues. GoFigure [3], OntoBlast [4], and Gotcha [5] weighted the GO terms by the BLAST *e*-values and chose GO terms by their weights. PFP [6, 7] made use of both strongly and weakly similar sequences of the query sequence to increase the coverage of functional annotation. ESG [8] exploited cascading homologues of the unannotated protein iteratively to improve the precision of prediction. ConFunc

[9] split the homologues into subgroups according to their annotations and then inferred annotations of the unannotated protein from these subgroups. These methods have a positive impact on protein function prediction. However, the homology-based methods may not work when the unannotated protein has low sequence similarity to other annotated sequences or all of its homologues are not annotated. Furthermore, it is also reported that transferring annotations among homologues may easily produce erroneous results [27].

As is known, domain is the conserved sequence and structure in the evolution of proteins, which plays as the stable and independent functional block of proteins [28]. Besides the detailed sequence, domain also carries some important structural information, that is, active site, which is tightly relevant to biological function [21]. Thus, a domain may be a suitable clue to discover the function of proteins. Statistics on UniProt database (released in May, 2013) show that more than sixty percent of proteins have domains. Moreover, domain databases and tools for efficient domain recognition have been developed including Pfam [29], SCOP [30], RPS-BLAST [31], and HMMER [32]. These databases and tools accelerate the analysis of domains in protein. In general, it seemed that inferring functions from resident domains of the protein is feasible and reasonable.

2. Related Works

So far, many efforts have been made for discovering functional signals carried by domains. Schug et al. [33] generated rules for function-domain associations based on the intersection of functions assigned to gene products which contain domains at varying levels of sequence similarity. Hayete and Bienkowska [34] designed an automated predictor based on decision tree to assign functions for domains. Mulder et al. [35] mapped GO terms to the domain if all proteins with the given domain do not exist in the set of proteins without the given GO term. Song et al. [36] transferred functions based on alignment of domain content. In analogy with [35], Forslund and Sonnhammer [37] assigned GO term to domain set if and only if all proteins containing the domain set also are annotated with the given GO term. Rentzsch and Orengo [38] transferred annotations in single profile-based sequence cluster. These methods are easily understood and realized, but they are readily misled into making an error-prone prediction by spurious and missing annotations of proteins. Even a single protein missing a valid GO term is enough to mislead the functional inferring about its domains.

In addition, Zhao et al. [39] utilized the protein-domain features, domain-domain interaction, and domain coexisting features to predict domain function. Their work extended the coverage of domain annotation effectively and provided solid foundation for predicting protein function. However, their work mainly paid attention to domain function rather than how the annotation of domain affects protein function. In our work, we focus on how to predict protein function based on domain annotation.

Recently, the probabilistic models have become increasingly popular for their remarkable performance on uncertainty inference. Forslund and Sonnhammer [37] utilized Naïve Bayesian (NB) model for assigning terms to domain set. Nevertheless the Naïve Bayesian model required that domain sets occurrence independently, which does not come with practice. Thus, Forslund et al. had attempted to reduce the dependencies between domain subsets using an averaged contribution from each domain subset. However, the conditional independence assumption may still not hold. Subsequently, Messih et al. [40] designed two models based on NB: one is DRDO that an averaged contribution from each subset which contains the sequential neighboring domains is used to solve the problem of dependency; the other is DRDO-NB which took recurrence and order of domains into consideration. Although computational complexity of DRDO is lower than that of NB, it may still not satisfy the conditional independence assumption. Moreover, all of these methods pruned GO terms of resident domains before they assigned GO terms to the host protein. Thus, some weak functional signals which may be amplified by dependencies between domains are likely to be neglected.

Fang and Gough [41] generalized a dcGO predictor for inferring GO terms associated with individual domains and supradomains based on protein-level GO annotation (GOA) and families of protein. dcGO exploited P value to evaluate the association strength (mentioned as relevance in the following sections to simplify) between domain and GO term. Since P value only represents the probability of error involved in null hypothesis, it may not be reasonable for estimating the relevance between domain and GO term by P value. In other words, P value can be used to determine which GO term is related to the given domain from statistical perspective but it is not enough to measure the degree of their relatedness. Thus, an appropriate metric is needed for weighting the relevance between GO term and domain objectively.

In this paper, we design a method to seek functions for proteins (SeekFun) effectively. Under this method, a mapping of GO terms and domains is constructed based on protein-level GOA and domain compositions of proteins. The relevance between domain and GO term is measured by symmetrical conditional probability. Based on the relevance of resident domains and terms, the relevance between host protein and GO terms is computed. Finally, the GO terms with relevance above a predefined threshold are used to annotate the host protein. The performance of SeekFun is validated by a series of experiments. The results suggest that our method is effective and reliable for protein function prediction.

3. Methods

3.1. Step 1: Construct and Weight Mapping of Domains and GO Terms. It is assumed that the resident domains may be associated with GO terms of the host protein. It is a rough assumption about the relationship between domain and GO term and may result in a large number of false

associations. To differentiate the true associations from the false ones, the relevance between domain and GO term need be measured. Judged with this, the true associations will have higher relevance while the false ones will have lower relevance.

As mentioned earlier, P value can be used to determine whether the domain is related to the GO term or not. When the P value of domain and GO term is larger than the given significance threshold, it is considered that the domain can be annotated with the GO term, and vice versa. However, the larger P value does not mean a more tight relationship between domain and GO term. In simple words, P value may be not suitable for measuring relevance between domain and GO term. Suppose that v_j represents that the protein containing domain d_j and u_i denotes that the protein plays the function described by GO term go_i. The conditional probability $\text{pr}(u_i \mid v_j)$ means the probability of that the protein containing d_j is annotated by go_i. The $\text{pr}(u_i \mid v_j)$ can reflect the dependence of go_i on the d_j. Likewise, the $\text{pr}(v_j \mid u_i)$ represents the probability of that the protein annotated by go_i containing the domain d_j. The $\text{pr}(v_j \mid u_i)$ can reflect the dependence of d_j on the go_i. Thus, it can be inferred that simple conditional probability can reflect relevance between domain and GO term partly but not enough. As (1), symmetrical conditional probability may be appropriate to measure the relevance between GO term go_i and domain d_j, $\text{DR}(\text{go}_i, d_j)$. Consider

$$\text{DR}\left(\text{go}_i, d_j\right) = \sqrt{\text{pr}\left(u_i \mid v_j\right) \cdot \text{pr}\left(v_j u_i\right)}. \tag{1}$$

Equation (1) means that the relevance between go_i and d_j is determined jointly by conditional probabilities between v_j and u_i. The bigger the probabilities are, the stronger the relevance between them is. Range of the relevance is from 0 to 1. The higher relevance means that the domain is more probably annotated with the term.

Supposed that $\#\text{prot}(\text{go}_i)$ is the number of proteins which are annotated with the go_i, $\#\text{prot}(d_j)$ is the number of proteins which contain d_j, and $\#\text{prot}(\text{go}_i, d_j)$ is the number of proteins which have to do with both go_i and d_j. Accordingly, (1) can be transformed into (2). Consider

$$\begin{aligned}\text{DR}\left(\text{go}_i, d_j\right) &= \sqrt{\frac{\#\text{prot}\left(\text{go}_i, d_j\right)}{\#\text{prot}\left(d_j\right)} \cdot \frac{\#\text{prot}\left(\text{go}_i, d_j\right)}{\#\text{prot}\left(\text{go}_i\right)}} \\ &= \sqrt{\frac{\#\text{prot}\left(\text{go}_i, d_j\right)^2}{\#\text{prot}\left(d_j\right) \cdot \#\text{prot}\left(\text{go}_i\right)}}.\end{aligned} \tag{2}$$

3.2. Step 2: Transfer GO Terms of Resident Domains to the Host Protein. As is known, GO terms are organized as a directed acyclic graph and may be related to each other. Thus, predicting functions of proteins should take the relationship between GO terms into consideration. GO has a rule called "true path rule", which defines the terms along the pathway from a given term to the root term that must annotate the protein if the protein is annotated with the given term. And a path upward from the given term to the root term in GO hierarchy is regarded as a true path of the term. Considering the true path rule, the mapping of GO terms and domains is extended along true paths of the GO terms in our method. Traditionally, if a domain is associated with a GO term, it is also associated with all ancestral terms of the GO term with equal relevance. However, it is reported that the semantics of GO terms has differences even if they are parent-child relationship. Thus, the relevance between the domain and each ancestor of the GO term may be different and the semantic differences between GO terms should be considered.

In fact, the organization of GO terms can be regarded as a split-flow semantic system (SFSS). In SFSS, the root term is the source of semantics which can describe the general functions while others represent semantic branches of the root term and illustrate specific functions. So the terms along the true path of the given term have different capabilities to describe the functions. Generally, for a given function, the ancestral term is more likely to describe the given function than its descendants because the semantics of its ancestors is more general and has more power to describe function. It can be explained by semantic coverage of GO term, which can be roughly estimated by the number of its descendants [42].

Based on these analyses, we proposed a novel strategy, namely RSC, to measure the relevance between domain and ancestral term based on semantic coverage. That is, given a term go_i which is related to the domain d_j with relevance $\text{DR}(\text{go}_i, d_j)$, the relevance between the domain d_j and the ancestral term go_k of term go_i, can be calculated by (3). In (3), $D(\cdot)$ represent the descendant set of the given term and $\text{Anc}(\text{go}_i)$ consists of the ancestors of the term go_i. Naturally, along the true path, the term which is nearer to root has bigger relevance value with the given domain than others and it is more probably to annotate the host protein

$$\text{DR}\left(\text{go}_k, d_j\right) = \frac{\left|D\left(\text{go}_k\right)\right|}{\left|D\left(\text{go}_i\right)\right|} \cdot \text{DR}\left(\text{go}_i, d_j\right), \quad \text{go}_k \in \text{Anc}\left(\text{go}_i\right). \tag{3}$$

It is supposed that protein is associated with all GO terms which are related to the resident domains of the protein. The relevance between protein and GO term can be derived from the relevance of the term and resident domains of the protein. For example, if a protein P contains a set of domain $D = \{d_1, d_2, \ldots, d_n\}$ and $\text{DR}(\text{go}_i, d_j)$ denote the relevance between go_i and d_j, then the relevance between P and go_i, $\text{PR}(\text{go}_i, P)$, can be computed by (4). Consider

$$\text{PR}\left(\text{go}_i, P\right) = \max_{d_j \in D, 1 \le j \le n} \text{DR}\left(\text{go}_i, d_j\right). \tag{4}$$

After the extension, each protein is associated with a group of GO terms with strong or weak relevance. To facilitate comparison, the relevance of proteins and terms need be normalized. Each of GO categories should be analyzed, respectively, as they have different biological meanings. For each protein, the relevance between the protein P and the root r of subontology (GO: 00003674 for molecular function, GO: 00008150 for biological process, and GO: 00005575 for

Table 1: The details of experimental datasets.

	Uniref50	SwissProt	TrEMBL
Number of annotated proteins	20693	17176	19526
Number of proteins with domains	11673	15810	13588
Number of involved domains	4998	4430	3642
Number of involved GOs	4812	7572	3992

cellular component), PR(r, P), is used as baseline because the real annotations of proteins must be split from the root in the GO hierarchy. The normalized relevance of go_i and P, NPR(go_i, P), can be measured by (5). The relevance has been standardized to scale from 0 to 1. The higher relevance means that the protein is more probably annotated with the term. Consider

$$\text{NPR}\left(\text{go}_i, P\right) = \frac{\text{PR}\left(\text{go}_i, P\right)}{\text{PR}\left(r, P\right)}. \tag{5}$$

Through the above steps, the relevance of proteins and GO terms has been measured already. To select real annotations from candidate annotations, a threshold t of relevance need be defined. If the relevance between protein and term is above the predefined threshold t and the term is assigned to the protein, and vice versa. In our study, the threshold t is about 0.6~0.7 as the proposed model performs well on the given datasets.

4. Results and Discussion

4.1. Experimental Datasets. Three up-to-date protein subsets of UniProt, Uniref50, SwissProt, and TrEMBL, are selected to evaluate SeekFun. The proteins which are only annotated with GO term inferred from electronic annotations are excluded from the experimental datasets. The SwissPfam database is used to determine the detailed domain composition of proteins. All the datasets are downloaded on May 20, 2013. The details of the experimental datasets are listed in Table 1.

4.2. Evaluation Metrics. Consistent with Critical Assessment of Functional Annotations (CAFA) experiments [42], the precision, recall, and f-measure are utilized to judge the performance of methods in our experiments. Given a target protein x and $K(x)$ which is a set of known (true) annotations of x, the precision of the method at threshold $t \in [0, 1]$, pr(t), can be calculated as

$$\text{pr}\left(t\right) = \frac{1}{m\left(t\right)} \sum_{x \in S} \frac{\left|K\left(x\right) \cap P_t\left(x\right)\right|}{\left|P_t\left(x\right)\right|}. \tag{6}$$

In (6), $P_t(x)$ is the set of predictive annotations whose relevance with x is above t. S is the target set for testing. $m(t)$ is the number of proteins which at least has one predictive GO term under given t. Similarly, the recall of method at threshold t, rc(t), can be computed by

$$\text{rc}\left(t\right) = \frac{1}{|\text{S}|} \sum_{x \in S} \frac{\left|K\left(x\right) \cap P_t\left(x\right)\right|}{\left|K\left(x\right)\right|}. \tag{7}$$

The f-measure (the harmonic mean of precision and recall) gives an intuitive number for comparisons of the concerned methods. For each method, the maximal value of f-measure on the overall threshold of relevance, $F_{\max}$, is calculated as

$$F_{\max} = \max_t \left\{ \frac{2 \cdot \text{pr}\left(t\right) \cdot \text{rc}\left(t\right)}{\text{pr}\left(t\right) + \text{rc}\left(t\right)} \right\}. \tag{8}$$

Considering the relationships between GO terms, the comparisons are guided by the true path rule. That is, the $K(x)$ and $P(x)$ are extended by adding all ancestors of their members to them before comparing.

4.3. Comparisons of Relevance Computed by Different Strategies. To illustrate the rationality of weighting strategies, the relevance weighted by symmetrical conditional probability (R_{SCP}) is compared with those measured by P value (R_{PV}) and traditional conditional probability (R_{dSCP}). In fact, it is hard to evaluate the relevance between domain and GO term for lacking of the gold standard. To determine appropriate strategies for weighting relevance, some properties of relevance are analysed. A little random noise may make a difference between observed and real datasets and the relevance should be robust on these similar datasets. To simulate similar datasets, a series of subsets of Uniref50, SwissProt, and TrEMBL is constructed by taking nine of their ten equal-size partitions randomly at a time. The calculations of relevance by different strategies are performed on these subdatasets. The varied distributions of relevance on the different datasets may be good evidence for which strategy is more proper for weighting relevance.

The distributions of relevance derived from different strategies are displayed in Figure 1. In order to facilitate comparison, without loss of meanings, the logarithmic transformation and Z-score transformation are performed on R_{PV}, which are represented by $\log R_{\text{PV}}$ in Figure 1. Observed the figure, it can be found that R_{dSCP} is the most changeful while the distribution curves of both R_{SCP} and $\log R_{\text{PV}}$ have similar trends. All of those suggest that, as for robustness on tiny different datasets, the R_{SCP} and R_{PV} are more proper than R_{dSCP}. What is more, the curves of R_{SCP} and R_{PV} appear to have obvious monotonicity that is beneficial for assigning GO terms to the domain.

Meanwhile, the curves of R_{PV} are steeper than those of R_{SCP} on each dataset, which imply that the resolution of R_{SCP} is lower than R_{PV}. In this paper, the resolution describes how sensitive the relevance is to distinguish true positive association between domain and GO term from other negative ones. The resolution of relevance is inversely proportional to the average density of relevance in their range, which is just indicated by the steepness of the curves in the figures. In simple words, the larger the average density of relevance in their range is, the harder the true association between domain and GO term is determined.

On the other hand, the relevance derived from two significantly different datasets may vary more dramatically than those from the similar datasets. Statistically, the SwissProt and TrEMBL have no intersection while they have 5031 and

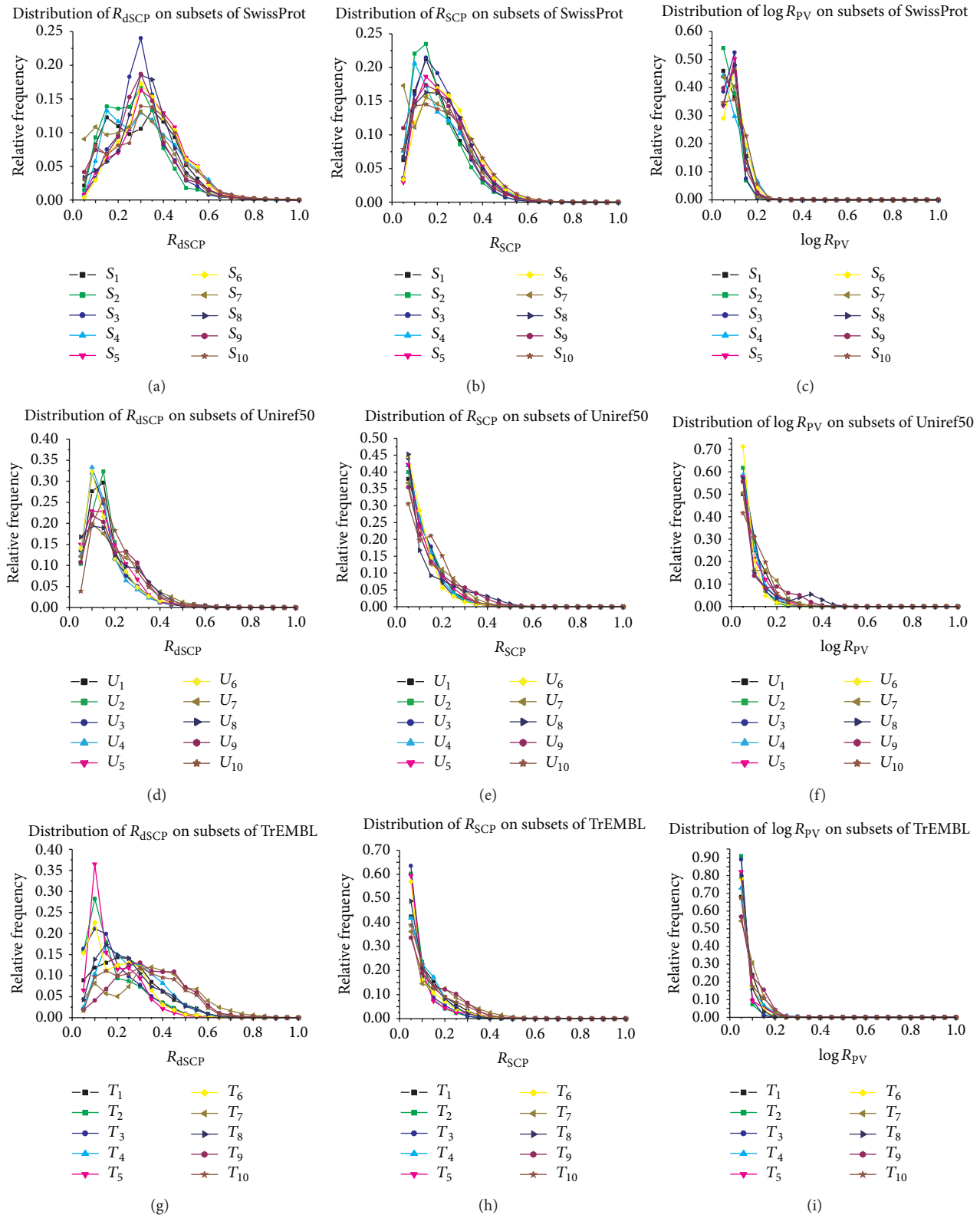


FIGURE 1: Compare distributions of relevance on similar datasets. R_{dSCP}, R_{SCP}, and log R_{PV} represent the relevance computed by conditional probability, symmetrical conditional probability, and P value, respectively. S_i is constructed by taking nine of ten equal-size partitions of SwissProt at a time, $i = 1, 2 \cdots 10$. Likewise, U_j and T_k denote the constructed subsets of Uniref50 and TrEMBL separately, $j, k = 1, 2 \cdots 10$. The curves display the distributions of relevance on similar subsets of the experimental datasets.

Table 2: Compare the impact of R_{SCP} on protein function prediction.

		Uniref50			SwissProt			TrEMBL		
		MF	BP	CC	MF	BP	CC	MF	BP	CC
$Pred_{pfam2go}$	Precision	0.5568	0.6094	0.5978	0.4861	0.532	0.5557	0.3856	0.3482	0.3954
	Recall	0.441	0.2888	0.1747	**0.6951**	0.4496	0.2255	0.6176	0.6027	0.2255
	F_{max}	0.4922	0.3918	0.2704	0.5721	0.4873	0.3208	0.4748	0.4414	0.2872
$Pred_{weighted}$	Precision	0.2979	0.2502	0.1944	0.3514	0.2609	0.2611	0.3472	0.2179	0.2033
	Recall	**0.7805**	**0.6959**	**0.8774**	0.5946	**0.6523**	**0.7603**	**0.7917**	**0.7011**	**0.8213**
	F_{max}	0.4312	0.3681	0.3183	0.4417	0.3727	0.3887	0.4827	0.3325	0.3259
$Pred_{combine}$	Precision	**0.8506**	**0.8622**	**0.7503**	**0.8543**	**0.8577**	**0.7662**	**0.7641**	**0.7939**	**0.835**
	Recall	0.6971	0.5823	0.7655	0.56	0.4093	0.5984	0.7662	0.6371	0.7309
	F_{max}	**0.7662**	**0.6951**	**0.7578**	**0.6765**	**0.5542**	**0.672**	**0.7651**	**0.7069**	**0.7795**

The best results are in bold.

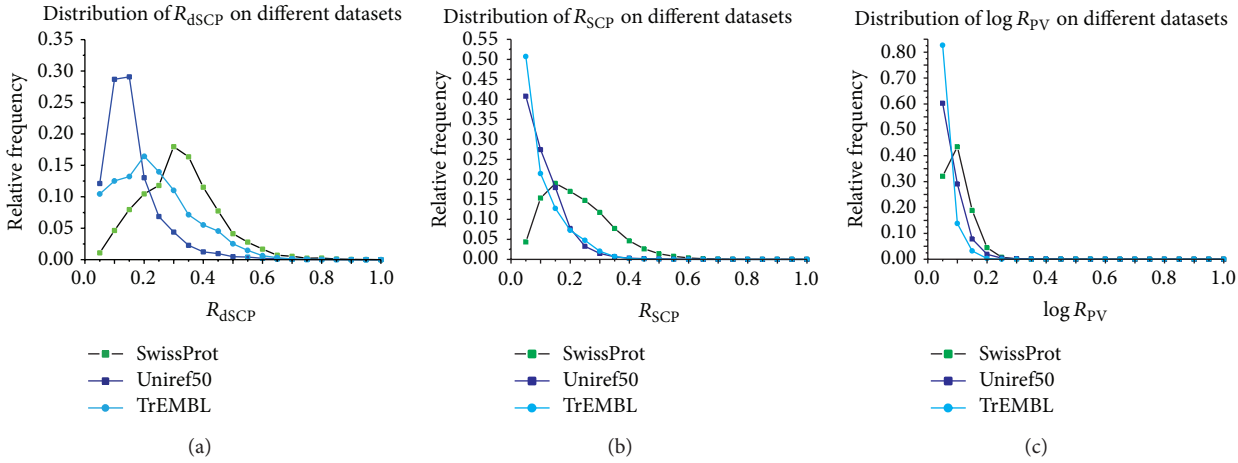


Figure 2: Compare distributions of relevance on significantly different datasets. R_{dSCP}, R_{SCP}, and log R_{PV} represent the relevance computed by conditional probability, symmetrical conditional probability, and P value, respectively. SwissProt, Uniref50, and TrEMBL are the significantly different datasets.The curves display the distributions of relevance on the experimental datasets.

6929 common proteins with Uniref50, about up to their 30% and 36% separately. Consequently, the difference between the curves of relevance on SwissProt and TrEMBL should be larger than those of others. Observing the distributions of relevance on these datasets, as displayed by Figure 2, it can be found that the R_{SCP} and log R_{PV} vary as expected but the log R_{PV} still suffers from low resolution. Generally speaking, it can be concluded that R_{SCP} is a more suitable measure of relevance between domain and GO term.

4.4. The Impact of R_{SCP} on Protein Function Prediction. For validating its impact on protein function prediction, R_{SCP} is tested on experimental datasets: Uniref50, SwissProt, and TrEMBL, respectively. The comparison is performed on the three subontologies of GO: molecular function (MF), biological process (BP), and cellular component (CC) separately. The comparison includes two steps: constructing mapping of domains and GO terms and annotating proteins based on the mapping.

In our experiment, the mapping of Pfam domains and GO terms (pfam2go) is downloaded from the Gene Ontology website in May, 2013. Based on this reliable mapping, all annotations which are associated with the resident domains are assigned to the host protein. This method is named $Pred_{pfam2go}$ in this paper. Meanwhile, the mapping of Pfam domains and GO terms which is weighted by R_{SCP} is also used for prediction, namely, $Pred_{weighted}$. In the comparisons, $Pred_{pfam2go}$ and $Pred_{weighted}$ are validated by performing the same task in the same framework on the basis of different mappings of domains and GO terms. To avoid the influence of domain coverage, the weighted mapping with R_{SCP} just includes the domains in pfam2go when it is applied. Here, to compare the influence of the strategy R_{SCP} and RSC, the method which is the combination of them is also used to perform the same task and marked with $Pred_{combine}$. Their performances are illustrated in Table 2.

As displayed in Table 2, $Pred_{weighted}$ has higher recall than $Pred_{pfam2go}$ while the latter achieves better precision than the former. These results suggest that the $Pred_{weighted}$ could improve the specificity of annotations but it is at the cost of precision.

It also can be found from Table 2 that $Pred_{combine}$ is superior to others in general. Compared to $Pred_{pfam2go}$, $Pred_{combine}$ outperforms on both precision and recall. In

Table 3: Compare the impact of *RSC* on protein function prediction.

		Uniref50			SwissProt			TrEMBL		
		MF	BP	CC	MF	BP	CC	MF	BP	CC
RPE	Precision	0.2709	0.1582	0.184	0.328	0.2334	0.2866	0.2803	0.1664	0.2131
	Recall	**0.8255**	**0.7195**	**0.901**	**0.6801**	**0.5096**	**0.7807**	**0.8625**	**0.7416**	**0.9**
	F_{max}	0.4076	0.2575	0.3044	0.4424	0.3195	0.4184	0.4224	0.2709	0.3443
RSC	Precision	**0.8782**	**0.8804**	**0.768**	**0.8529**	**0.8616**	**0.7751**	**0.8064**	**0.8071**	**0.8163**
	Recall	0.7876	0.6856	0.8163	0.5953	0.4294	0.6083	0.8229	0.6985	0.7716
	F_{max}	**0.8304**	**0.7709**	**0.7914**	**0.7012**	**0.5731**	**0.6816**	**0.8146**	**0.7489**	**0.7933**

The best results are in bold.

Table 4: Compare the performances of the concerned methods.

		Uniref50			SwissProt			TrEMBL			Average
		MF	BP	CC	MF	BP	CC	MF	BP	CC	
NB	Precision	0.7778	0.7339	0.7421	0.8362	0.8121	**0.8408**	**0.8977**	**0.8477**	**0.8927**	0.8201
	Recall	0.0428	0.0319	0.0244	0.5012	0.4212	0.3718	0.5086	0.3721	0.4819	0.3062
	F_{max}	0.0812	0.0612	0.0473	0.6267	0.5547	0.5156	0.6493	0.5172	0.6259	0.4088
DRDO	Precision	0.7716	0.7151	0.7109	0.8232	0.8004	0.8312	0.8644	0.8073	0.8623	0.7985
	Recall	0.1777	0.1385	0.1115	0.5868	0.5023	0.4437	0.5517	0.429	0.5422	0.387
	F_{max}	0.2888	0.2321	0.1928	0.6852	0.6173	0.5786	0.6735	0.5603	0.6657	0.4994
DRDO-NB	Precision	0.8375	0.6906	0.7439	0.7379	0.7186	0.6766	0.8426	0.8471	0.7512	0.7607
	Recall	0.2094	0.232	0.2695	0.2394	0.2272	0.2633	0.157	0.1502	0.1452	0.2104
	F_{max}	0.335	0.3474	0.3956	0.3615	0.3452	0.379	0.2647	0.2551	0.2434	0.3252
dcGO	Precision	0.4342	0.3751	0.3014	0.558	0.5253	0.4375	0.3801	0.3473	0.3494	0.412
	Recall	0.6127	0.503	0.6127	**0.605**	**0.4303**	0.5904	0.6692	0.5137	0.6509	0.5764
	F_{max}	0.5083	0.4297	0.4041	0.5805	0.4731	0.5026	0.4848	0.4144	0.4547	0.4725
SeekFun	Precision	**0.8782**	**0.8804**	**0.7682**	**0.8529**	**0.8616**	0.7751	0.8064	0.8071	0.8163	**0.8274**
	Recall	**0.7876**	**0.6856**	**0.8163**	0.5953	0.4294	**0.6083**	**0.8229**	**0.6985**	**0.7716**	**0.6906**
	F_{max}	**0.8304**	**0.7709**	**0.7914**	**0.7019**	**0.5731**	**0.6816**	**0.8146**	**0.7489**	**0.7933**	**0.7451**

The best results are in bold.

contrast to $\text{Pred}_{\text{weighted}}$, $\text{Pred}_{\text{combine}}$ significantly improved the precision while it does as well as $\text{Pred}_{\text{weighted}}$ on recall. Thus, it can be concluded that R_{SCP} tend to select specific terms for the proteins and RSC balances this bias by propagating in the GO hierarchy. It may be the reason that $\text{Pred}_{\text{combine}}$ shows higher performances.

4.5. The Impact of RSC on Protein Function Prediction. In order to validate the effectiveness of the *RSC*, it is compared with traditional strategy which set the relevance of domain and terms along a true path as equal (RPE). The two strategies are applied to predict protein functions based on the mapping of domains and GO terms weighted by R_{SCP}. Their best performances are listed in Table 3.

As displayed, RPE gives a better recall while RSC has higher precision and F_{max}. In general, RSC may be more beneficial to protein function prediction than RPE. It may be because the resolution of R_{SCP} is effectively promoted by different relevance between protein and each term along a true path. On the contrary, RPE considered that protein has equal relatedness to every term along the true path, which makes it harder to determine the true positive associations between terms and the host protein. Even if the threshold of *RPE* is 1, its precision is still lower than the other one and recall goes down. It confirms that the differences of GO terms have significant influence on their relevance with protein.

4.6. Comparison of the Concerned Methods. To assess the efficiency of SeekFun, it is compared together with NB, DRDO, DRDO-NB, and dcGO on the three benchmark datasets. The performances of concerned methods on different dataset are shown in Table 4. To provide a simple number for comparison between methods, the averages of metrics on each dataset are also listed.

In terms of precision, SeekFun is superior to others while NB, DRDO, and DRDO-NB follow in turn. The dcGO is significantly lower than others. As aforementioned, dcGO measured relevance between domain and GO term by *P* value while other methods calculated it based on conditional probability. These results may indicate again that the relevance estimated by *P* value is not sensitive enough to determine the true positive associations between domain and GO term. In other words, R_{PV} has low resolution for distinguishing real annotations of protein. By contrast, the conditional probability is more suitable for estimating relevance.

As for the recall, SeekFun performs better than others while dcGO follows. It also can be found that the performances of NB, DRDO, and DRDO-NB are not as well as the other methods. Comparing the details of them, NB, DRDO, and DRDO-NB infer functions of protein from annotations of domain combinations, which enhance the precision of function prediction. However, in the process of discovering domain combinations, some slightly weak associations between domain and GO term may be neglected. The resident domains of the host protein may interplay as different combinations to perform different functions. Nevertheless, these methods judge domain combination if the members of the domain combination exist in the protein and the P value of their combination is above predefined threshold. It may miss information covered in the potential domain combinations and domain themselves. We guess this may be the reason that these methods show lower recall of functions.

Overall, SeekFun has better performance than others. It can attribute to the weighted mapping of domains and GO terms and the strategy for transferring annotations of resident domains to the host proteins. The weighted mapping can reflect the relationship between domain and GO term properly. The transferring strategy takes both the differences and connections of terms into consideration, which greatly promote its capability of distinguishing real associations of domains and terms from the false ones.

5. Conclusions

In this paper, SeekFun is developed for protein function prediction. Instead of using amino acid sequence of protein directly, SeekFun takes the resident domains of proteins and protein-level GOA as clues to annotate proteins. We tested the overall performance of SeekFun and the results suggest that SeekFun is superior to the concerned methods: NB, DRDO, DRDO-NB, and dcGO on precision and recall generally.

Meanwhile the effects of relevance computed by symmetrical conditional probability, (R_{SCP}) and the strategy for inferring annotations of protein from the annotations of its resident domains (RSC) are validated, respectively. The results of these experiments confirmed that both of them are effective and can promote the performance of protein function prediction. In the proposed method, R_{SCP} tend to discover specific functions of protein but it cannot ensure the precision and RSC is used to compensate for the lack of R_{SCP}. So the combination of them achieves high performances. The main idea of SeekFun could be used to acquire knowledge from other functional ontologies based on different domain resources easily. SeekFun will facilitate the discovery of protein functions and the insights into the biological roles of proteins.

Conflict of Interests

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

Authors' Contribution

The experiments are conceived and designed by Zhixia Teng and Maozu Guo. The experiments are performed by Zhixia Teng and Chunyu Wang. The data are analyzed by Zhixia Teng, Qiguo Dai, and Jin Li. The paper is prepared by Zhixia Teng, Maozu Guo, Qiguo Dai, and Xiaoyan Liu.

Acknowledgments

Maozu Guo is supported by Natural Science Foundation of China (61271346) and Specialized Research Fund for the Doctoral Program of Higher Education of China (20112302110040). Xiaoyan Liu is supported by Natural Science Foundation of China (61172098 and 91335112).

References

[1] D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler, "The GOA database in 2009—an integrated Gene Ontology Annotation resource," *Nucleic Acids Research*, vol. 37, no. 1, pp. D396–D403, 2009.

[2] S. Hennig, D. Groth, and H. Lehrach, "Automated gene ontology annotation for anonymous sequence data," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3712–3715, 2003.

[3] S. Khan, G. Situ, K. Decker, and C. J. Schmidt, "GoFigure: automated gene ontology annotation," *Bioinformatics*, vol. 19, no. 18, pp. 2484–2485, 2003.

[4] G. Zehetner, "OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3799–3803, 2003.

[5] D. M. A. Martin, M. Berriman, and G. J. Barton, "GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes," *BMC Bioinformatics*, vol. 5, article 178, 2004.

[6] T. Hawkins, S. Luban, and D. Kihara, "Enhanced automated function prediction using distantly related sequences and contextual association by PFP," *Protein Science*, vol. 15, no. 6, pp. 1550–1556, 2006.

[7] T. Hawkins, M. Chitale, S. Luban, and D. Kihara, "PFP: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data," *Proteins: Structure, Function and Bioinformatics*, vol. 74, no. 3, pp. 566–582, 2009.

[8] M. Chitale, T. Hawkins, C. Park, and D. Kihara, "ESG: extended similarity group method for automated protein function prediction," *Bioinformatics*, vol. 25, no. 14, pp. 1739–1745, 2009.

[9] M. N. Wass and M. J. E. Sternberg, "ConFunc—functional annotation in the twilight zone," *Bioinformatics*, vol. 24, no. 6, pp. 798–806, 2008.

[10] W. T. Clark and P. Radivojac, "Analysis of protein function and its prediction from amino acid sequence," *Proteins: Structure, Function and Bioinformatics*, vol. 79, no. 7, pp. 2086–2096, 2011.

[11] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 8, pp. 4285–4288, 1999.

[12] M. Huynen, B. Snel, W. Lathe III, and P. Bork, "Predicting protein function by genomic context: quantitative evaluation

and qualitative inferences," *Genome Research*, vol. 10, no. 8, pp. 1204–1210, 2000.

[13] F. Enault, K. Suhre, and J.-M. Claverie, "Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis," *BMC Bioinformatics*, vol. 6, article 247, 2005.

[14] B. E. Ersgelhardt, M. I. Jordan, K. E. Muratore, and S. E. Brersfser, "Protein molecular function prediction by Bayesian phylogenomics," *PLoS Computational Biology*, vol. 1, no. 5, article e45, 2005.

[15] P. Gaudet, M. S. Livstone, S. E. Lewis, and P. D. Thomas, "Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium," *Briefings in Bioinformatics*, vol. 12, no. 5, pp. 449–462, 2011.

[16] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein-protein interaction data," *Journal of Computational Biology*, vol. 10, no. 6, pp. 947–960, 2003.

[17] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, "Global protein function prediction from protein-protein interaction networks," *Nature Biotechnology*, vol. 21, no. 6, pp. 697–700, 2003.

[18] F. Pazos and M. J. E. Sternberg, "Automated prediction of protein function and detection of functional sites from structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 41, pp. 14754–14759, 2004.

[19] D. Pal and D. Eisenberg, "Inference of protein function from protein structure," *Structure*, vol. 13, no. 1, pp. 121–130, 2005.

[20] D. Bandyopadhyay, J. Huan, J. Liu et al., "Structure-based function inference using protein family-specific fingerprints," *Protein Science*, vol. 15, no. 6, pp. 1537–1543, 2006.

[21] Z.-P. Liu, L.-Y. Wu, Y. Wang, L. Chen, and X.-S. Zhang, "Predicting gene ontology functions from protein's regional surface structures," *BMC Bioinformatics*, vol. 8, article 475, 2007.

[22] J. Skolnick and M. Brylinski, "FINDSITE: a combined evolution/structure-based approach to protein function prediction," *Briefings in Bioinformatics*, vol. 10, no. 4, pp. 378–391, 2009.

[23] L. Sael, M. Chitale, and D. Kihara, "Structure- and sequence-based function prediction for non-homologous proteins," *Journal of Structural and Functional Genomics*, vol. 13, no. 2, pp. 111–123, 2012.

[24] C. Huttenhower, M. Hibbs, C. Myers, and O. G. Troyanskaya, "A scalable method for integration and functional analysis of multiple microarray datasets," *Bioinformatics*, vol. 22, no. 23, pp. 2890–2897, 2006.

[25] S. Brady and H. Shatkay, "EpiLoc: a (working) text-based system for predicting protein subcellular location," *Pacific Symposium on Biocomputing*, pp. 604–615, 2008.

[26] A. Wong and H. Shatkay, "Protein function prediction using text-based features extracted from the biomedical literature: the CAFA challenge," *BMC Bioinformatics*, vol. 14, supplement 3, article S14, 2013.

[27] D. Devos and A. Valencia, "Practical limits of function prediction," *Proteins*, vol. 41, no. 1, pp. 98–107, 2000.

[28] M. Bashton and C. Chothia, "The generation of new protein functions by the combination of domains," *Structure*, vol. 15, no. 1, pp. 85–99, 2007.

[29] 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.

[30] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995.

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

[32] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching," *Nucleic Acids Research*, vol. 39, supplement 2, pp. W29–W37, 2011.

[33] J. Schug, S. Diskin, J. Mazzarelli, B. P. Brunk, and C. J. Stoeckert Jr., "Predicting gene ontology functions from ProDom and CDD protein domains," *Genome Research*, vol. 12, no. 4, pp. 648–655, 2002.

[34] B. Hayete and J. R. Bienkowska, "Gotrees: predicting go associations from protein domain composition using decision trees," *Pacific Symposium on Biocomputing*, pp. 127–138, 2005.

[35] N. J. Mulder, R. Apweiler, T. K. Attwood et al., "New developments in the InterPro database," *Nucleic Acids Research*, vol. 35, supplement 1, pp. D224–D288, 2007.

[36] N. Song, R. D. Sedgewick, and D. Durand, "Domain architecture comparison for multidomain homology identification," *Journal of Computational Biology*, vol. 14, no. 4, pp. 496–516, 2007.

[37] K. Forslund and E. L. L. Sonnhammer, "Predicting protein function from domain content," *Bioinformatics*, vol. 24, no. 15, pp. 1681–1687, 2008.

[38] R. Rentzsch and C. A. Orengo, "Protein function prediction using domain families," *BMC Bioinformatics*, vol. 14, supplement 3, article S5, 2013.

[39] X.-M. Zhao, Y. Wang, L. Chen, and K. Aihara, "Protein domain annotation with integration of heterogeneous information sources," *Proteins: Structure, Function and Genetics*, vol. 72, no. 1, pp. 461–473, 2008.

[40] M. A. Messih, M. Chitale, V. B. Bajic, D. Kihara, and X. Gao, "Protein domain recurrence and order can enhance prediction of protein functions," *Bioinformatics*, vol. 28, no. 18, pp. i444–i450, 2012.

[41] H. Fang and J. Gough, "A domain-centric solution to functional genomics via dcGO *Predictor*," *BMC Bioinformatics*, vol. 14, supplement 3, article S9, 2013.

[42] Z. Teng, M. Guo, X. Liu, Q. Dai, C.-Y. Wang, and P. Xuan, "Measuring gene functional similarity based on group-wise comparison of GO terms," *Bioinformatics*, vol. 29, no. 11, pp. 1424–1432, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 351095, 12 pages
http://dx.doi.org/10.1155/2014/351095

Research Article

Pathway Bridge Based Multiobjective Optimization Approach for Lurking Pathway Prediction

Rengjing Zhang,[1] Chen Zhao,[2] Zixiang Xiong,[1] and Xiaobo Zhou[2]

[1] *Electrical and Computer Engineering Department, Texas A&M University, College Station, TX 77840, USA*
[2] *Radiology Comprehensive Cancer Center Cancer Biology, Wake Forest University, Winston-Salem, NC 27103, USA*

Correspondence should be addressed to Xiaobo Zhou; xizhou@wakehealth.edu

Received 28 January 2014; Accepted 16 March 2014; Published 16 April 2014

Academic Editor: Xing-Ming Zhao

Ovarian carcinoma immunoreactive antigen-like protein 2 (OCIAD2) is a protein with unknown function. Frequently methylated or downregulated, OCIAD2 has been observed in kinds of tumors, and TGFβ signaling has been proved to induce the expression of OCIAD2. However, current pathway analysis tools do not cover the genes without reported interactions like OCIAD2 and also miss some significant genes with relatively lower expression. To investigate potential biological milieu of OCIAD2, especially in cancer microenvironment, a nova approach pbMOO was created to find the potential pathways from TGFβ to OCIAD2 by searching on the pathway bridge, which consisted of cancer enriched looping patterns from the complicated entire protein interactions network. The pbMOO approach was further applied to study the modulator of ligand TGFβ1, receptor TGFβR1, intermediate transfer proteins, transcription factor, and signature OCIAD2. Verified by literature and public database, the pathway TGFβ1- TGFβR1- SMAD2/3-SMAD4/AR-OCIAD2 was detected, which concealed the androgen receptor (AR) which was the possible transcription factor of OCIAD2 in TGFβ signal, and it well explained the mechanism of TGFβ induced OCIAD2 expression in cancer microenvironment, therefore providing an important clue for the future functional analysis of OCIAD2 in tumor pathogenesis.

1. Background

Tumor microenvironment has been largely studied as a dynamic system to define the behaviors of cancer. This system is orchestrated by cytokines, growth factors, inflammatory cells, cancer cells, stroma, as well as the extracellular matrix [1]. Tumor-associated fibroblasts (TAFs) are major elements of tumor stroma and have been shown to play an important role in tumor growth and progression. Epithelial-to-mesenchymal transition (EMT) is a major source of TAFs. In tissue fibrosis, it is well-established that epithelial cells contribute to the accumulation of fibroblasts by undergoing EMT in response to stimuli from the microenvironment [2]. TGFβ remains among the key factors responsible for the recruitment of Tumor Associated Fibroblasts (TAFs) and induction of EMT. TAFs, meanwhile, strongly contribute to the production and activation of TGFβ in the activated stroma and thereby generate the autocrine feed-forward loop that is characteristic for persisting fibroblasts activities [3]. However, the exact regulation between TGFβ signals and TAFs in tumor microenvironment is yet to be completely understood.

OCIAD2 was originally immunoscreened from ascites of a patient with ovarian cancer and was found to be an immunoreactive antigen [4]. However, the function of OCIAD2 protein, involved pathways, and molecular mechanisms has never been reported. Based on our preliminary data analysis, we hypothesized that human OCIAD2 represents a potential tumor suppressor gene in some tumor types and its dysregulation is involved in TGFβ regulated signaling in tumor microenvironment for the following reasons: (1) high-throughput profiling data and public database analyses showed that OCIAD2 is frequently methylated and/or down-regulated in some kinds of cancers [5–7]; (2) GEO database revealed that the expressions of OCIAD2 are induced by TGFβ signal in pancreatic (GSE23952), lung adenocarcinoma (GSE17708), and ovarian cancer cells (GSE6653); (3) moreover, a computational analysis with TCGA database revealed that methylation site of OCIAD2 is top-ranked signature in ovarian Metastasis-Associated Fibroblasts (MAFs) [8]. This

evidence indicated a potential biological milieu of OCIAD2. We hereby speculate that downregulated OCIAD2 expression in tumor microenvironment facilitates deregulated TGFβ signaling. As a consequence of these changes, tumor cells escape immunosurveillance and exaggerate tumor progression and metastatic spread.

To predict molecular network of OCIAD2 in TGFβ regulated tumor microenvironment, a nova pathway analysis approach with bioinformatics methods has been developed. Current signal analysis methods typically have three steps: build literature based preliminary signaling pathways model; generate gene expression experimental data; detect the shortest path as the specific signal and verify biological meaning. Pathways consisting of highest differentially expressed genes and reported interactions would be shown as the results in this kind of pathway study methods. However, not all the targets or receptors of ligands are with top expression changes; that is, TGFβ regulates numerous other growth factors positively and negatively, some of which are not the most obviously changed ones but still give response to the stimulation of TGFβ. Moreover, new genes with seldom previous reports, such as OCIAD2, cannot be included in any pathways because of the lack of known interactions with other proteins. To study the mechanism of OCIAD2 changes induced by TGFβ stimulation in cancer cell lines, a new approach to inferring the signaling paths based on the pathway bridges between the stimulant TGFβ and its target gene OCIAD2 using the multiobjective optimization approach, named pbMOO, was developed. Pathway Bridge was defined as a subset of protein interactions network that consisted of clustering loop motifs with extremely high frequency occurring in cancer related processes than by chance. All four-vertex motifs, among which the triangle and rectangle were shown with significantly higher occurrences than randomized ones, were detected from a network generated from HPRD database with 12794 proteins and 39031 interactions. Rather than traversing the entire protein interaction network with enormous nodes and edges, all the loop motifs were clustered as a "Pathway Bridge" between TGFβ signaling pathway and cancer signaling pathway. Relatively, the time saving approach returned to highly reliable protein paths only by searching connecting nodes on the bridge. Moreover, motifs on the bridge were concentrated on cancer related processes, which guaranteed that the nodes chosen for the path are specified for cancer microenvironment. Then, the cost of a protein path was defined by gathering up the cost of each edge, which is the P value sum of two interacted protein nodes. According to the property of P value, the path cost is the probability of obtaining a path whose cost is no more than the one that was actually observed. Hence, the reliability of a predicted pathway can be represented by its cost.

177 transcription factors of Homo sapiens were analyzed from Transcriptional Regulatory Element Database [9], and androgen receptor (AR) was discovered as the most credible transcription factor of OCIAD2. Applied the approach on GSE42357 and GDS3634 expression data from NCBI, and the paths with the lowest cost were picked out as the responsible possible molecular mechanisms between TGFβ and OCIAD2 in hepatocellular carcinoma (HCC) samples and prostate cancer cell lines. After verifying the biological meaning of the low-cost paths, the signal TGFβ1- TGFβR1- SMAD2/3- SMAD4- AR-OCIAD2 was discovered and explained TGFβ's stimulation on OCIAD2 expression in cancer.

2. Methods

In order to see how the signature gene is enrolled in the ligand stimulation signal, a pathway bridge based multiobjective optimization approach (pbMOO) was designed and summarized, as shown in Figure 1. Three kinds of data were chosen as the initial input data: protein interactions (Figure 1(a)) from HPRD [10] were applied for building the entire protein-protein interactions (PPI) network; signaling pathways (Figure 1(b)) from both KEGG [11] and IPA [12] were selected as the background pathway library based on which the pathway bridges were constructed; groups of microarray data (Figure 1(c)) were used for assigning the path cost in the optimization problem and presenting the correlation between genes. Calculated by FANMOD [13], loop motifs (Figure 1(a1)), which were the higher frequency occurring subnetworks out of the entire PPI network, were shown to be enriched in cancer and related pathways. Searching on the pathway bridge (Figure 1(b2)), which was defined as a set of loop motif clusters (Figure 1(a2)) connecting ligand and signature genes, a multiobjective optimization problem was solved by finding the pathways with the lowest path cost that was assigned by gene expression P value. When multiple experimental gene expression data were used, the cost of each path was then defined as the summation of average P value of connected genes in the optimization problem. Then the modular study (Figure 1(c1)) was applied on the calculated results of the optimization problem. Finally, the integrated signals, which began with ligand and its receptor, passing through transduction proteins and targeting transcription factor and finally the signature, were output as the most reliable predicted pathways (Figure 1(c2)) explaining how the ligand changes affected the signature.

2.1. Cell Lines and Drug Treatment. Hep-3B and Du-145 were obtained from American Type Culture Collection. All cell lines were cultured in DMEM with 10% fetal bovine serum (FBS) and antibiotics. TGFβ1 (R&D Systems, Minneapolis, MN) were applied at concentrations of 5 ng/mL. TGFβR inhibitor LY2109761 were purchased from Selleck Chemicals LLC (Houston, TX), using 2 μM. For the drug treatment, human liver and prostate cancer cell lines, Hep-3B and Du-145 were treated with 5 ng/mL TGFβ1, 2 μM LY2109761, and combination for 24 hours in serum free media, and OCIAD2 mRNA levels were determined by quantitative real-time RT-PCR analysis.

2.2. RNA Extraction and Quantitative Real-Time PCR. Total RNAs were isolated from tumor cells using TRIZOL reagent (Life Technologies, USA) following the manufacturer's recommendations. RNA concentration and purity were determined by measuring absorbance at 260 and 280 nm with a NanoDropTM 1000 Spectrophotometer (Thermo Scientific,

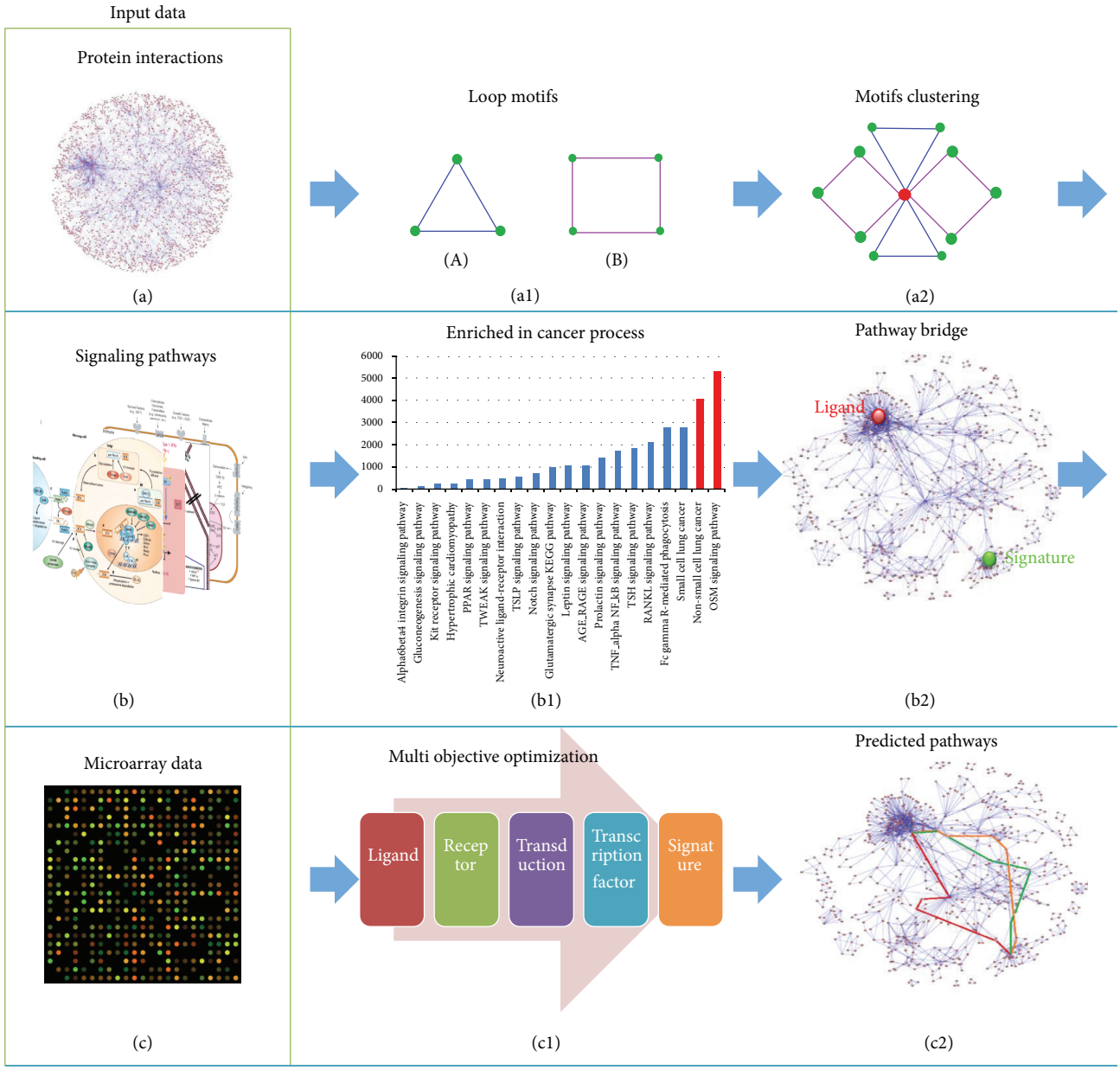


Figure 1

USA). cDNA synthesis was performed with Superscript III reverse transcriptase kit (Life Technologies, USA). Quantitative real-time reverse-transcription polymerase chain reaction (RT-PCR) was performed using an Applied Biosystems 7300 Sequence detection system (Applied Biosystems, Life Technologies, USA). The primer sets of OCIAD2 are described below: 5′-TGCGAGAATGTCAGGAAGAA-3′ and 5′-AAATCCCAAGAGACCAGCAA-3′.

2.3. Motif Detection of Protein-Protein Interaction Network. "Network Motifs" [14] are interconnected patterns (subgraphs) with significantly higher occurring in complicated networks than in randomized ones. In biological networks, almost all of the four vertex motifs were combinations of smaller motifs [15], and loop-structural motifs have been proved to be enriched in a protein-protein interaction (PPI) network generated from PPI database, that is, HPRD. As a literature-collected public database, HPRD has 12794 proteins and 39031 pairs of interactions for 9605 of them. The sufficient data capacity helps a lot on unclear reciprocities prediction.

In this paper, the outstanding tool Fast Network Motif Detection (FANMOD) [13] was applied to census four-vertex subgraphs in undirected PPI network by using the Randomized Enumeration (RAND-ESU) algorithm.

Motifs detection results from PPI network were shown as an example in Figure 2 and detailed in Supplementary 1.1 in Supplementary Material available online at

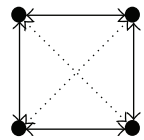
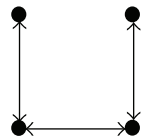
ID	Adj	Frequency (original)	Mean-frequency (random)	Standard-dev. (random)	Z-score	P value
31710		0.017858%	$1.1634e-007\%$	$1.6208e-008$	11018	0
8598		44.115%	45.117%	0.0016806	−5.9588	1

Figure 2

http://dx.doi.org/10.1155/2014/351095. Motifs ID lied in column one and adjacency matrix was presented in the second column; Frequency was the probability of each motif in original PPI network, and Mean-Frequency was of the motif that occurred in random networks; the standard deviation from the mean frequency was listed in the fifth column; Z-score meant the value of the difference of two frequencies divided by the standard deviation; and P value was the difference of motif numbers between random networks and original one then divided by the total number of random networks. ID 8598 in Figure 2 had relatively higher occurring frequency in random networks than the original one, and thus its Z-Score was negative and P value was large, which indicated that the chain-looked structure was really normal among the entire PPI network; on the contrary, ID 31710 was more special in the original network structure than randomly showing that the combination of triangle and square subnetwork was enriched among PPI network. Similarly, motifs ID 4382, recurring much more often than random chosen subnetwork, were with negative Z-Score and large P value; ID 13278, ID 4958, and ID 27030 motif structures were also enriched from randomized network and obtain positive Z-Score and relatively smaller P value. The whole results were listed in the supplementary. The outcome suggests loop-structural motifs, that is, shapes like triangle, spoon, and square, which are special patterns with high occurrences in protein interactions network.

20 signaling pathways derived from KEGG [11] were analyzed for motifs distribution. Results in Figure 3 showed that proteins on looping motifs are mainly from cancer and correlated signal pathways, in other words, motifs with loop structure are enriched in 14 types of carcinomatosis and related signaling pathways, such as cell cycle signaling pathway and immune system signaling pathways.

2.4. Motifs Clustering and Enrichment in Cancer Related Signaling Pathways. Loop-shaped motifs with no more than four vertices have the only two specific possibilities—triangle and square. Motifs Cluster (MC) is defined as converged cyclic motifs that share at least one protein. The common protein is called Center Point (CP), which is the identifier for distinguishing different motifs clusters. The toy model was showed in Figure 4.

Since looping motifs were proved to be occurring much more often in cancer and related signaling pathways, they can be treated as a bridge to link up cancer and its kinship pathways, which would provide a characteristic group of candidate protein interactions for future unclear links forecast.

Let P_1 be a chosen cancer signaling pathway, and let P_2 be a cancer-related signaling pathway. $\{\mathrm{MC}_1^{P_1P_2}, \mathrm{MC}_2^{P_1P_2}, \ldots, \mathrm{MC}_n^{P_1P_2}\}$ are the total n motif clusters between P_1 and P_2, and thus $|\mathrm{MC}^{P_1P_2}| = n$ by virtue of the number of identifiers $|\mathrm{CP}^{P_1P_2}| = n$, where $|\cdot|$ denotes the number of elements in a set. In order to evaluate the enrichment of MCs lying between P_1 and P_2, P value was introduced as the probability of obtaining a larger number of MCs for a pair of randomly chosen protein sets, keeping the same sizes with P_1 and P_2 and the capacity of intersection, than for P_1 and P_2. Consider the following:

$$p = \mathrm{prob}\left\{n' > n \mid n' = \left|\mathrm{MC}^{S_1S_2}\right|, n = \left|\mathrm{MC}^{P_1P_2}\right|\right\}, \quad (1)$$

where S_1 and S_2 are random protein sets picked out from entire proteins of HPRD database with the same size as P_1 and P_2; that is, $|S_1| = |P_1|$ and $|S_2| = |P_2|$, and satisfy $|S_1 \bigcap S_2| = |P_1 \bigcap P_2|$. Repeating the sampling for 1000 times, a random distribution f for the 1000 numbers of MCs can be generated. The complementary set of cumulative probability density function $F(|\mathrm{MC}^{P_1P_2}|)$ interprets the chances that a stochastic pair of protein sets has quantity of MCs which are being the bridges between them rather than the two chosen pathways, which is indeed motif clusters' P value. Consider the following:

$$p = F'\left(\left|\mathrm{MC}^{P_1P_2}\right|\right) = 1 - F\left(\left|\mathrm{MC}^{P_1P_2}\right|\right). \quad (2)$$

MCs connecting P_1 and P_2 are enriched if the P value is tiny, indicating that the bridging MCs linking up two protein sets are the main substructure of cancer signaling path P_1 and cancer enrolled signaling path P_2. Comparing with searching the enormous and complex integrated PPI network, the enriched MCs bridge efficiently limits and specializes the

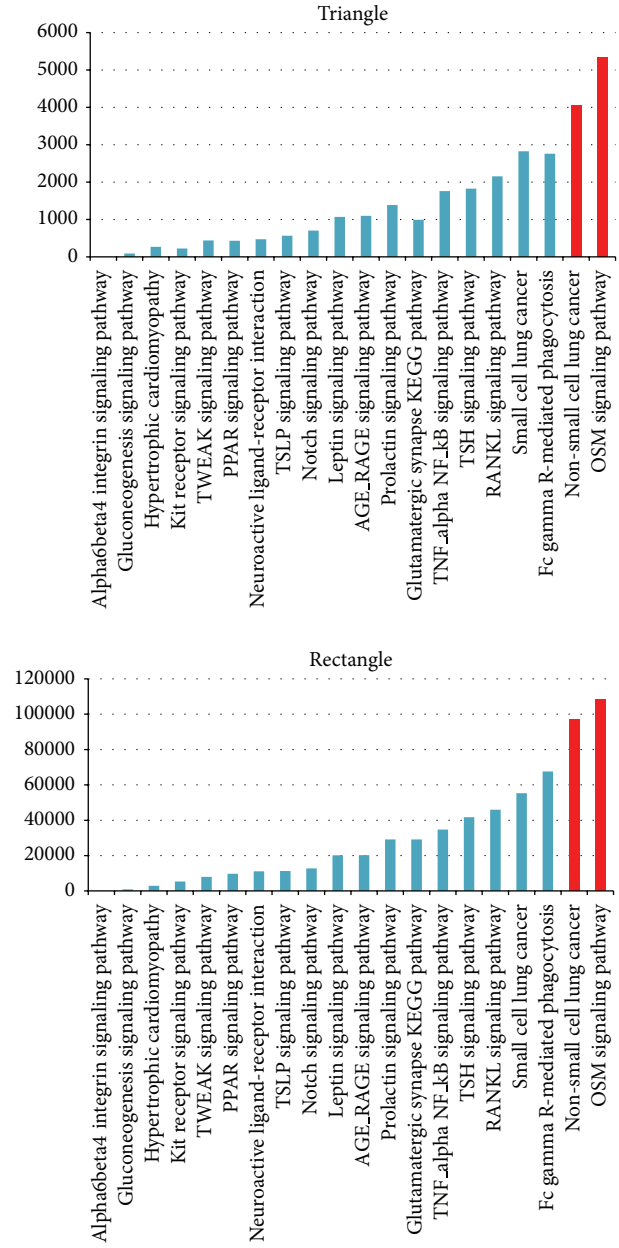


Figure 3

traversing range for forecasting uncertain protein paths, which increases the calculation speed thoroughly.

MCs' P value for combinations of different types and subtypes of carcinomatosis and involved signaling pathways were calculated and exampled in Figure 5. MCs bridges that have P value less than 0.01 were chosen to be the candidate subnetwork, from which ill-defined protein pathways would be selected.

2.5. Ill-Defined Protein Pathways Prediction. If changing condition of a protein A results in the upregulated or downregulated protein B, while A and B have neither direct interaction with each other nor indirect upstream and downstream relationship on any authentic signaling pathway, the underlying protein pathways for them could be detected on those MCs enriched pathway bridges. An optimization model, which is described as follows, was employed to acquire high-confidence potential protein pathways:

$$f(x) = \underset{\vec{x}\in\{MC_1^{P_1P_2}, MC_2^{P_1P_2}, \ldots, MC_n^{P_1P_2}\}}{\operatorname{argmin}} \sum_{i=1}^{N} x_i \cdot \mathrm{DES}_i + \lambda \sum_{i=1}^{N} x_i,$$

$$\text{s.t.} \quad \begin{cases} 2 \le \sum_{i=1}^{N} x_i \le 7, \\ i = 1, 2, \ldots, N, \\ \lambda \in |\Re|, \end{cases} \tag{3}$$

where N is the total number of proteins pertaining to MCs bridge for the selected pair P_1 and P_2. $\vec{x} = \{x_1, x_2, \ldots, x_N\}$ is protein path vector implying which element was contributed to the path—if protein i was taken count into the lurking protein path, then $x_i = 1$; otherwise, $x_i = 0$. Differential expression score (DES) was defined by each gene's P value from Student's t-test of gene expression experiment data in two conditions. The larger the P value, the larger the DES, and the less reliable the data. Thus, minimizing the first part of the objective function $\sum_{i=1}^{N} x_i \cdot \mathrm{DES}_i$ could ensure the maximization of the reliability of the predicted protein paths. The length of protein pathway; that is, $\sum_{i=1}^{N} x_i$, is an integer in the range of $[2, 7]$, which was decided by the fact that MCs were composed of looping structures up to 4 vertices. At this point, the latter part of the objective function took the responsibility of controlling the length of analyzed protein pathway with the aid of distinct settings of nonnegative parameter λ. $|\Re|$ is the absolute value of real number. Large λ was made for limited proteins and short connections, and optimization result was free to rope in proteins when $\lambda = 0$.

The optimization function was solved by the shortest path package in R, where Dijkstra's algorithm was employed and the length controlling parameter λ was set as zero to gain all the possible predictions. The solution of the optimization function was the optimal protein path vector $\vec{x}$ of N elements x_i, representing protein i was in the lowest-cost path or not.

2.6. Interacted Pairs Inference for Protein without PPI. For those proteins that have no canonical protein interaction supported, gene expression data was conduced to providing indistinct mutual effects and pointing out candidate proteins with which the separated proteins were closely bound up by the correlation between the pairs of genes.

2.7. Multiple Microarray Data Based Differential Expression Score. As a matter of fact, the P value of experimental gene expression data may vary a lot by different experiment designs and operators, and a good inferred protein path is the one which gets rid of the destabilizing factors. Thus, multiple microarray data sets were employed here for error deduction.

3. Results and Discussion

3.1. Dysregulation of OCIAD2 in Different Cancers and Its Induction by TGFβ in HCC and PC Cells. To determine the

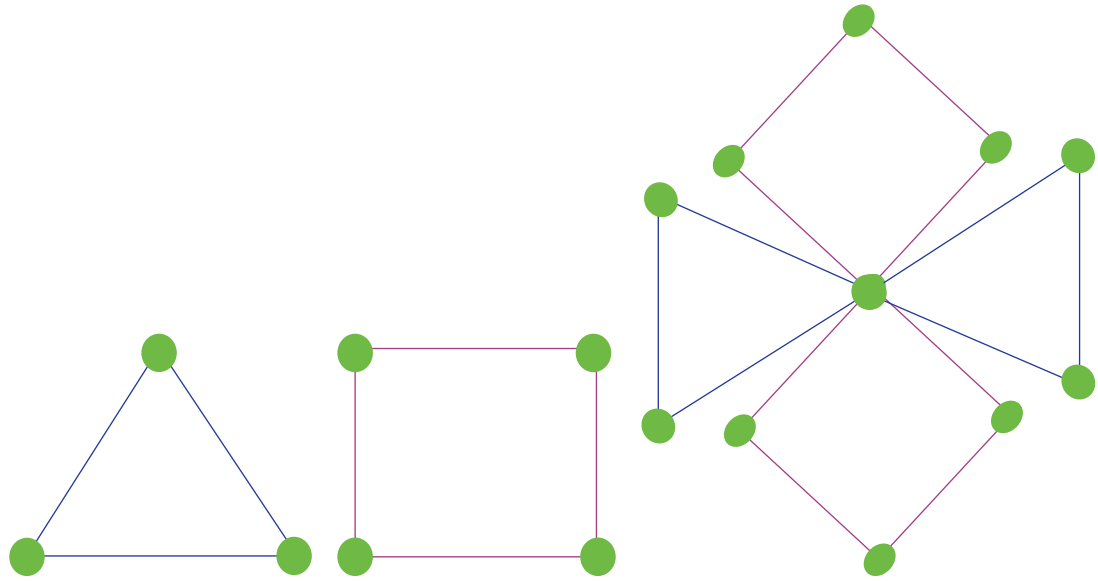


Figure 4

OCIAD2 expression, different available microarray studies were analyzed by the Oncomine database and GEO gene microarray data analysis tools. A significant downregulation of OCIAD2 mRNA expression was found in liver cancer and gastric stroma carcinoma tissues ($P < 0.001$ in both cases) (Figure 6; Left) based on Oncomine database analysis. The result indicated that OCIAD2 expression is in metastatic prostate tissues, but not in primary tumor tissues, which is clearly lower than normal prostate gland ($P = 0.009$) (Figure 6; Right; GSE6919). Frequently downregulated OCIAD2 expressions are also observed in CLL and malignant pleural mesothelioma [5, 6]. In glioblastoma, OCIAD2 expression is being silenced via DNA methylation mechanism [7]. With the suggestion of the fact that OCIAD2 was substantially unregulated in TGFβ1 treated pancreatic (GDS4106), lung (GSE17708), and ovarian (GSE6653) cancer cells, we have tested the possibility of OCIAD2 expression inducted by TGFβ in HCC and PC cells. Human HCC and PC cell lines, Hep-3B and Du-145, were treated with 5 ng/mL TGFβ1, 2 μM LY2109761, and combination for 24 h in serum free media, and OCIAD2 mRNA levels were determined by quantitative real-time RT-PCR analysis. OCIAD2 mRNA has increased 2.5- and 4.6-fold in Hep-3B and Du-145 cells by TGFβ1 treatment, respectively. This induction was totally suppressed by TGFβ1 receptor inhibitor LY2109761 (Figure 6(b)).

3.2. Potential Protein Pathway Prediction. Remarkable gene array profiles from GEO database indicated the expression of OCIAD2 in several kinds of cancer; that is, GDS3634 showed that OCIAD2 was obviously unregulated in prostate cancer cell line transfected with 20 nM miRNA Presursor Molecules miR-205. MiR-205 is selectively downregulated in metastatic breast and prostate cancer and suppresses metastatic spread of a human breast cancer xenograft in nude mice. In addition to its function in the regulation of EMT, the loss of miR-205 in prostate cancer also reduced some tumor suppressor genes' expression.

Mesenchymal stem cells (MSC), like other bone marrow-resident cells, have the capacity to differentiate into fibroblast-like cells that have been variably referred to as myofibroblasts, tumor-associated fibroblasts (TAF), fibrocytes, or pericytes within the tumor microenvironment [16]. Therefore, pbMOO approach was applied on both prostate cancer cell line GDS3634 and liver cancer associated mesenchymal stem cells GSE42357 gene expression data to study the possible molecular path involved in OCIAD2 by TGFβ stimulation.

3.2.1. Pathways in Prostate Cancer. Based on experimental data GDS3634, miR-205 expression effect on prostate cancer cell line, from NCBI public database browser, P value of 8 samples of Student's t-test, microarray data was applied as the link cost for predicted paths. Searched on the pathway bridge that has been enriched through prostate cancer pathways, the top 10 out of 88 forecasted paths were picked for further biological meaning verification (Supplementary Table 1.2). One reasonable forecast shown in Figure 7 was miR205-PRKCE- CNA13- CDH1- PTPN14-OCIAD2. High correlated genes with OCIAD2 were distinguished as dashed lines (red for positive correlation and green for negative correlation). PRKCE, as one of the six target genes of miR205, was marked as the green box. Among the pathway bridge that consisted of protein nodes (dashed circles) and proteins' interactions (green lines), a shortest protein path with least DES cost was emphasized by red lines.

Suggested by the significant association between TGFβ1 and CDH1, pbMOO approach was employed again aiming at finding out how TGFβ affects OCIAD2 across CDH1 in prostate cancer. Interestingly, "TGFβ1 influenced CDH1 across SMADs" was observed after filtering the predicted protein paths and verified by [17].

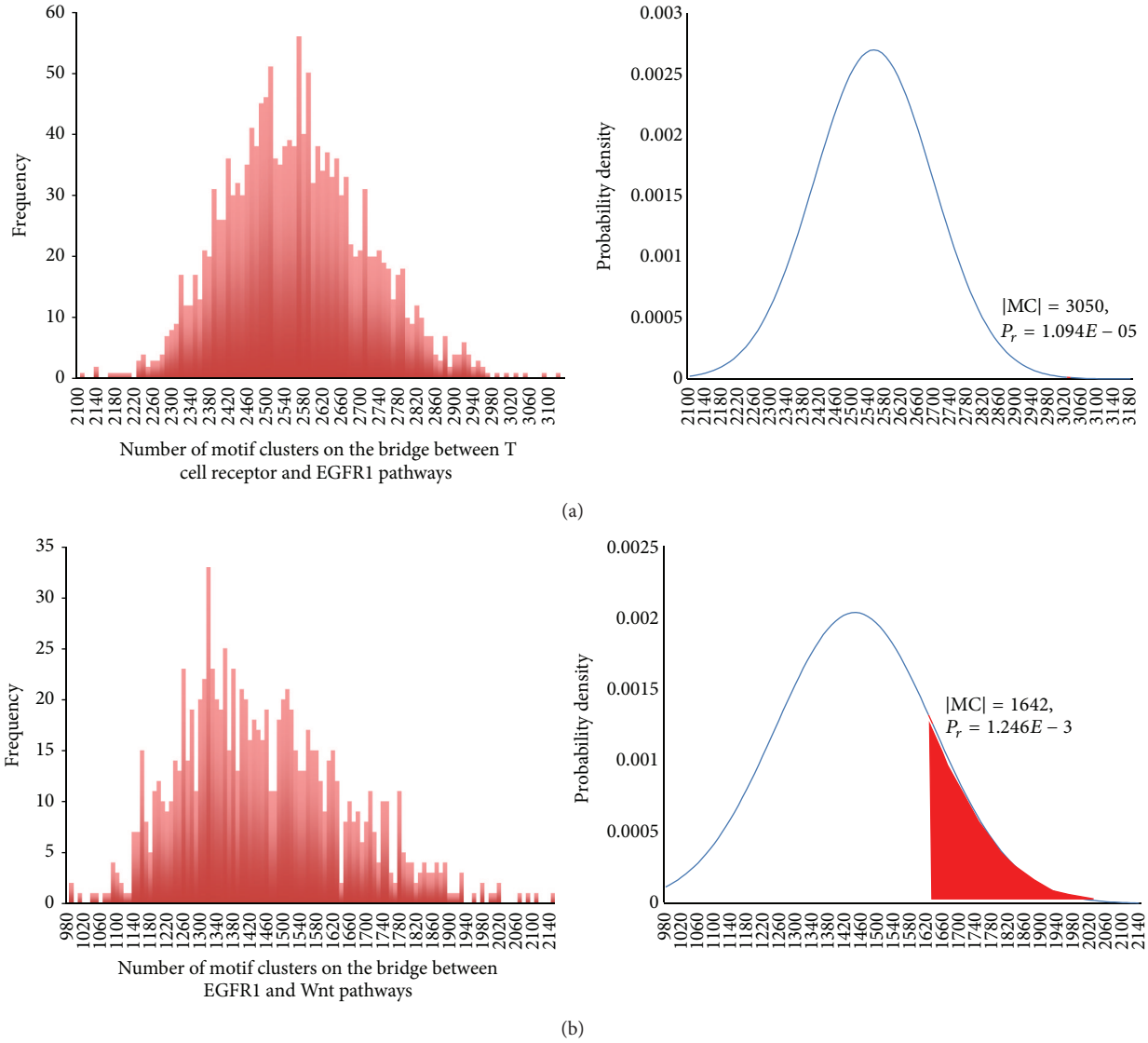


Figure 5: (a) Enriched Motif Clusters between EGFR1 and T cell receptor pathways. (b) Less Enriched Motif Clusters between EGFR1 and Wnt pathways.

3.2.2. Pathways in Liver Cancer. By the analysis of GSE42357 gene expression data, genes like C5, AG7, SDC2, and FHL2 have been suggested to be the candidates of OCIAD2 by their tight correlations with it. Significantly, those candidate genes all play important roles in cancer related processes, for instant, C5 takes the responsibility in inflammatory and cell killing processes [18] and FHL2 acts as both tumor-promoter or tumor-suppressor depending on different types of cancer [19]. The calculated results of the approach were paths with credibility cost, that is, TGFβ1- TGFβR1- CLU- C7-C5-OCIAD2 (cost 0.233181). This pathway was fully explained by the fact that CLU is a modulator of TGFβ1 signaling pathway by regulating Smad2/3 proteins [20] and the well-known protein interactions CLU-C7 and C7-C5.

3.3. Modular Mechanism Exploration. If a signaling transmission process, from extracellular through cytoplasm to nucleus, results in upregulation or downregulation of genes in the cell, then transcription factor (TF) usually plays the downstream role in this signaling flow. Since OCIAD2 was differentially expressed in prostate cancer cell line, liver cancer was associated with mesenchymal stem cells, and especially in TGFβ treated Panc-1 pancreatic adenocarcinoma cell line, and the question how is OCIAD2 activated by TGFβ signaling was solved by studying the probable transcription factor of OCIAD2, which also acting as the downstream of TGFβ signal.

Among all the 30981 genes from Transcriptional Regulatory Element Database [9], 177 transcription factors of homo sapiens were picked out as background human transcription factors library.. The algorithm to find the possible transcription factors of OCIAD2 in TGFβ treated signaling was divided into three main steps: first, pbMOO approach was employed to calculate the costs of all the shortest paths

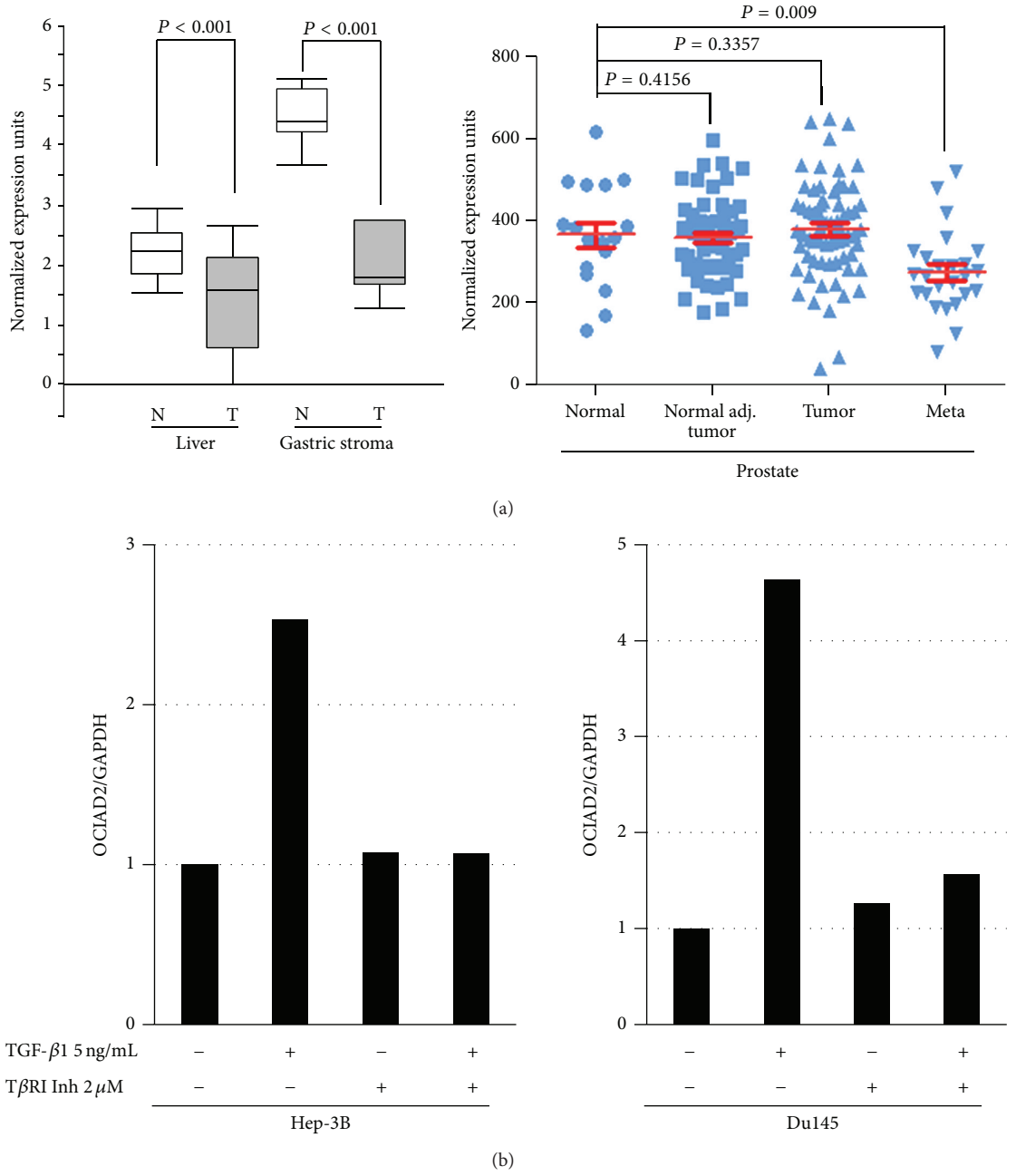


Figure 6: Expression of OCIAD2 and its induction by TGF-β.

between TGFβ and human transcription factors; then the ones with the least costs and high correlations with OCIAD2 in gene expression data were filtered out and selected as candidate transcription factors for OCIAD2; finally, biological TGFβ induced OCIAD2's differential expression mechanism which was concluded with literature verification.

3.3.1. Speculation of Human TF Enrolled in TGFβ Signal. As a fresh gene with rare reported property, the discovery of transcription factor in TGFβ1 signal is the main issue in OCIAD2 study. Among those 177 human transcription factors, the ones with the least pathway cost, which was defined by the sum of gene expression experimental P value of proteins on the pathway, are the most credible TFs for OCIAD2. The start point of the pathway was chosen as TGFβ1, and the end point was OCIAD2. All the pathway costs for those passing through TFs were calculated by applying pbMOO approach and the top of them were listed in the Supplementary Table 1.3.

3.3.2. Feasible TF of OCIAD2 in Cancer Cell Line. In this part, verification of the observation that AR might be the transcription factor of OCIAD2 in TGFβ1 signal, and SMAD

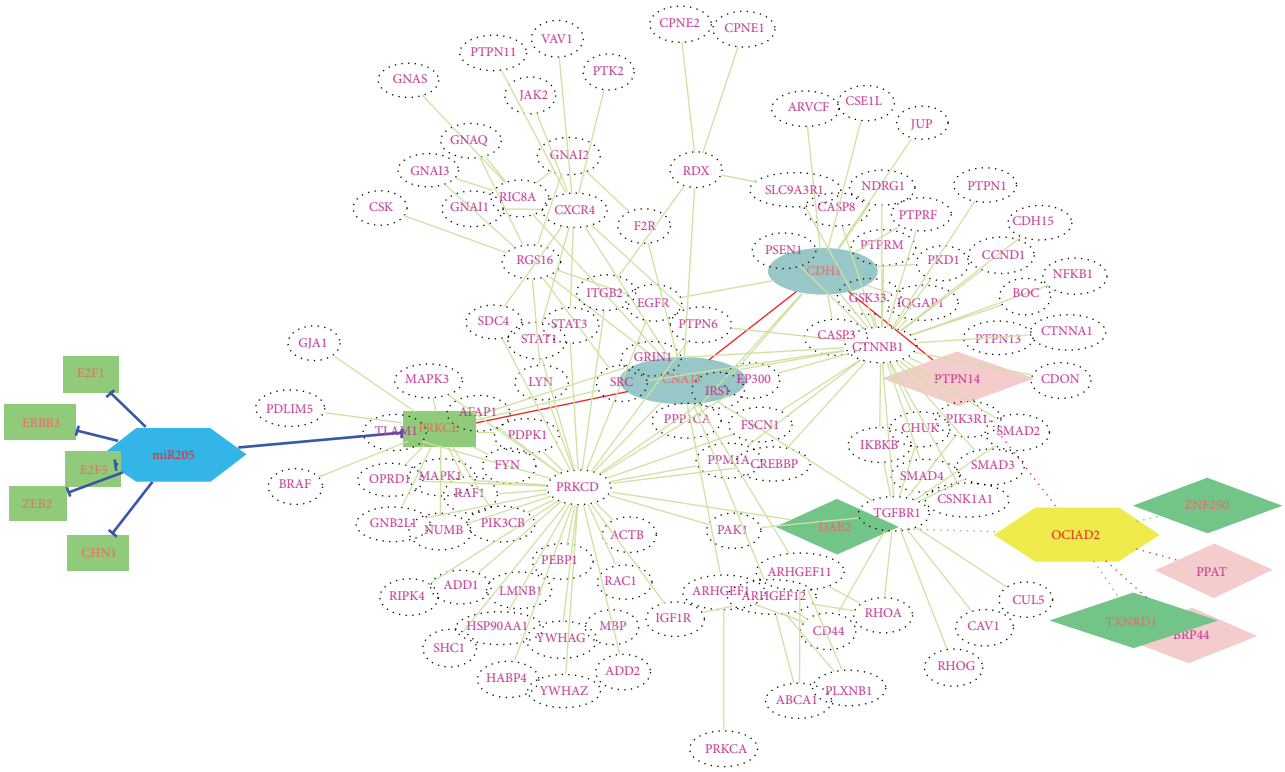
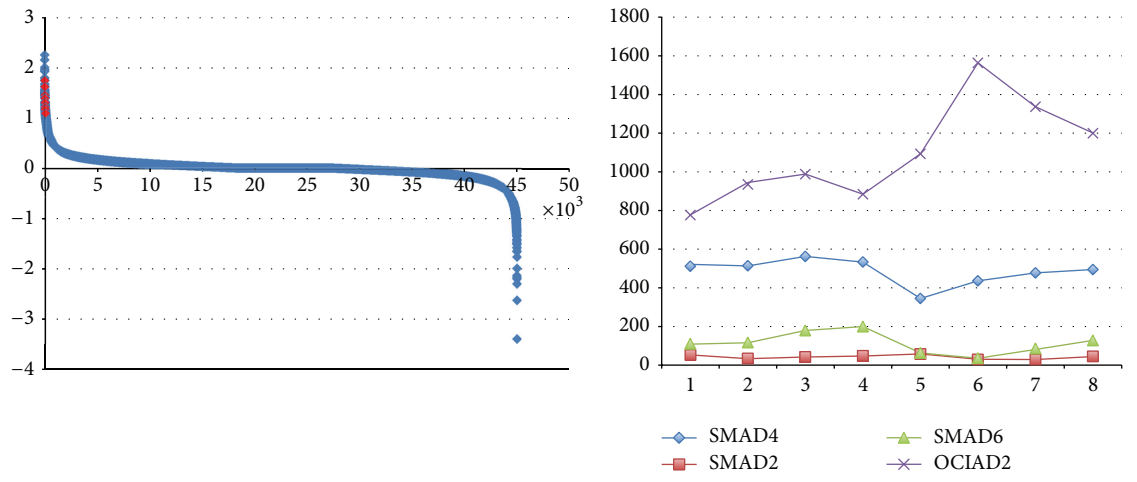


Figure 7



(a) (b)

Figure 8

group might enroll this process, was made in the light of gene expression data from both HCC and prostate cancer cell lines.

Due to the fact that GSE42357 gene expression data was the comparison between liver cancer associated mesenchymal stem cells (LC8-MSC) and normal ones (LN8-MSC) from the same patient, genes like OCIAD2 had only two experimental values—one for condition LC8-MSC and one for control LN8-MSC. The sample space was too tiny for Pearson Correlation calculation. For better results, the distribution of the fold change of each gene was plotted as the following Figure 8, and evidently, AR, which had ten pairs of experiment data in the range [1.09315, 1.74845], highly differentially expressed in HCC microenvironment. The fold change value of OCIAD2 in the same array data is −0.377255, which implied that AR must have negative effect on OCIAD2, in the other words, it should be the inhibitor of OCIAD2. Not like the top obviously expressed genes with fold changes close to 3.0, the SMAD group showed the relatively lower differential expression—most of them had slight positive changes less than 0.1. However, SMAD4 with fold change

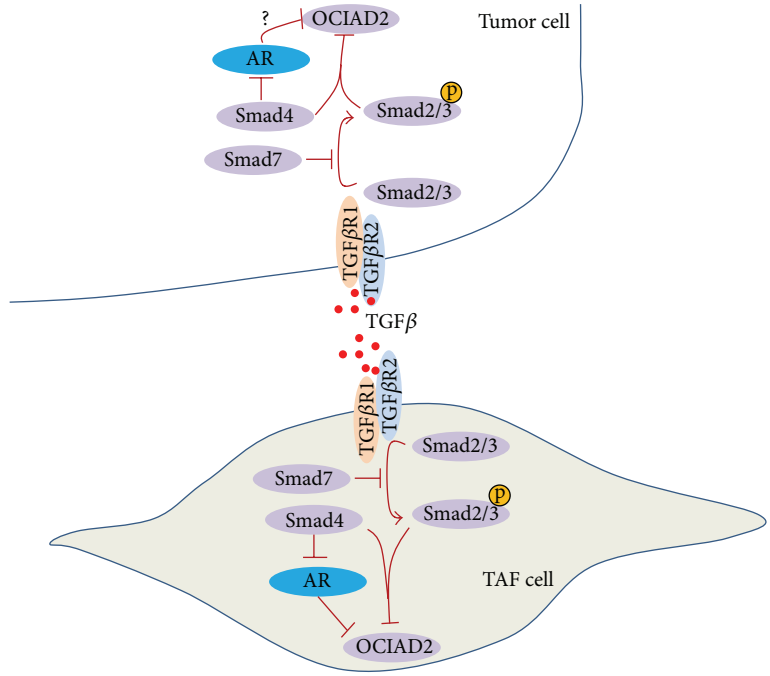


Figure 9

0.06034, SMAD2 0.4908, and SMAD3 0.02102 still survived as pathway proteins in pbMOO predictions, which were ignored by other pathway analysis methods.

Analyzing the results (top 30 were detailed in Supplementary Table 1.3), two interesting facts were observed: AR was shown with the highest frequency as the transcription factor of OCIAD2 in 17 pathways out of the top 50, while STAT5A was the second recurrent one that was the transcription factor of 10 pathways; AR appeared 34 times, and SMAD group proteins appeared 16 times in the top 50 pathways with lowest cost, in which other transcription factors had less occurrences. The observation insinuated that AR was the most reliable transcription factor of TGFβ1 signal induced OCIAD2, and SMAD group proteins had the closest relationships with this signaling process.

In DU145 prostate cancer cell line with restored miR-205 expression, unfortunately no data mapping with AR was found. However, these 8 samples of experimental data were still powerful to analysis how SMADs enrolled in OCIAD2 expression. As the figure showed, SMAD4 had the largest expression value as well as the highest negative correlation −0.696476 with differential expressed OCIAD2 among SMADs, followed by SMAD2 with correlation value −0.595238 and SMAD6 −0.571429.

3.3.3. Mechanism of TGFβ Induced OCIAD2's Expression. Analyzing the observations comprehensively on modular study and referring to related literature, the signaling pathway from TGFβ1 targeting OCIAD2 was concluded as shown in Figure 9: the signaling transmits from TGFβ1- TGFβR1- AR-OCIAD2 in liver cancer mesenchymal stem cell, then differentiates into Tumor-Associated-Fibroblasts (TAFs) in tumor stroma. As the only known mammalian coSMAD, SMAD4 transferred signaling from cytoplasm to TGFB signal. AR, the symbol of androgen receptor, mainly functioned as a DNA-binding transcription factor that regulates target gene expression from cytoplasm into nucleus.

Loss of cell adhesions or polarity is widely associated with CDH1 (E-cadherin). This process, referred to as EMT, enhances motility and invasiveness of many cell types and is often considered as a prerequisite for tumor infiltration and migration. TGFβ mediated induction of EMT processes is associated with specific stages of morphogenesis and during tumorigenesis by activating downstream signaling pathways in both Smad-dependent and Smad-independent mechanisms. The upregulation of OCIAD2 expression by TGFβ stimulation, and downregulated OCIAD2 expression in metastatic prostate tissues, revealed that OCIAD2 played roles in TGFβ promoted tumor cell migration, invasion, and mobility. The discovery that OCIAD2 has been enrolled in TGFβ signal across CDH1 powerfully testified our postulation—OCIAD2 could act as a downstream effector of TGFβ signals. In our predicted path, SMAD4 had the largest expression value as well as the highest negative correlation −0.696476 with differential expressed OCIAD2 among SMADs families. Previous study reported that Smad3/4 cooperated with Snail1 which acted as corepressors of CDH1 in the EMT process [21]. Due to the lack of information on protein interaction with OCIAD2, future biological assay needs to investigate the potential relationships between CDH1 and OCIAD2 in tumor EMT. In addition, smad signaling is required to maintain epigenetic silencing of some key EMT related proteins in breast cancer progression [22]. Because OCIAD2 frequently methylated in some kinds of cancers [5, 6, 23], we speculate that activated TGFβ-Smad signaling provides an epigenetic memory to maintain silencing of

OCIAD2 in EMT as well. Thus, disruption of TGFβ-Smad4-OCIAD2 signaling may be a useful therapeutic strategy to target tumor progression.

Specifically, the predicted path TGFB1- TGFBR1-SMAD2/3-SMAD4 was verified by [24, 25], and TGFB1's influence on AR with SMAD3 was also proved in [26].

4. Conclusions

In this study, a bioinformatics approach was developed, and it demonstrated that the function-unknown protein ovarian carcinoma immunoreactive antigen-like protein 2 (OCIAD2) is probably regulated by TGFβ and AR signals in the tumor EMT process. OCIAD2 is an immunoreactive antigen, which functions, involved pathways, and molecular mechanisms have never been reported. Current popular signaling analysis tools like IPA [12] are focusing on the highest differentially expressed genes with sufficient literature supported, ignoring signatures such as OCIAD2. Moreover, as the comprehensive analysis of wide-field database, the output pathways of those tools can hardly be specific for an interesting stimulation or disease. Overcoming the insufficiency of the knowledge on the new gene OCIAD2 and studying the modular signaling mechanism from the given ligand to the pointed signature, the pbMOO approach successfully answered the question "how did the observed signature gene OCIAD2 get involved in ligand TGFβ stimulation signal," and detailed the predicted pathways into the tumor microenvironment.

With pbMOO approach, a new pathway "TGFβ1-TGFβR1- AR-OCIAD2" in liver cancer mesenchymal stem cell was predicted, which will differentiate into Tumor-Associated-Fibroblasts (TAFs), one of the major components of tumor stroma. Stromal-epithelial crosstalk regulates all phases of cancer metastasis. In prostate cancer, androgen signaling is central to stromal-epithelial cross-talk in tumor progression. Tissue-based studies of human prostate cancer have shown that stromal AR expression and transcriptional activity downstream of the AR are lower in stromal cells which are derived from carcinomas. Androgen Receptor (AR) may promote hepatocarcinogenesis or suppress HCC metastasis. These opposite roles of AR also occur in prostate cancer [27]. The potential mechanisms for the AR dual roles are possibly caused by the differential AR signals in different cellular types having an oncogenic role in stroma and epithelial cells, but a suppressive role in basal intermediate epithelial cells. As DU-145 is an AR-independent cell lacking AR protein expression, the predicted path from AR to OCIAD2 in prostate cancer needs more support of more biological experiments. Further biological experiments are still needed to explore the existence of pathway TGFβ-Smad4-OCIAD2 signaling in AR-dependent cell models as well.

The signal from TGFβ, via the AR, played a critical role in the deregulation of TGFβ signaling in prostate and/or liver tumorigenesis, and those TGFβ effectors (Smads 3 and 4) serving as negative regulators of AR-mediated transcription in cancer cells have been established by several investigations [28]. With pbMOO approach, the functional unknown protein OCIAD2 was also enrolled into a signal pathway "TGFβ1- TGFβR1- SMAD2/3- SMAD4- AR-OCIAD2" in tumor and adjacent microenvironment. Currently, clinical studies using antiandrogens had disappointing results, few beneficial effects on patients, or even less survivals. Understanding the molecular mechanisms of AR in tumor microenvironment will undoubtedly further improve the results obtained with antitumor therapeutic strategies.

Conflict of Interests

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

Authors' Contribution

Rengjing Zhang carried out the approach and solved the optimization problem for predicting the lurking pathways, participated in the modular mechanism studies, and drafted the paper. Chen Zhao carried out the problem setup, participated in predicted results analysis and verification, and helped to draft the paper. Zixiang Xiong participated in the study design and coordination. Xiaobo Zhou conceived of the study, participated in the approach algorithm design and coordination, and helped to draft the paper. All authors read and approved the final paper.

Funding

This work was supported by funding: NIH R01LM010185 (Zhou), NIH U01HL111560 (Zhou), U01 CA166886-01 (Zhou).

Acknowledgment

The authors would like to thank Dr. Guangxu Jing for his early work on pathway modeling and the discussion with the members of Bioinformatics and Systems Biology.

References

[1] I. P. Witz and O. Levy-Nissenbaum, "The tumor microenvironment in the post-PAGET era," *Cancer Letters*, vol. 242, no. 1, pp. 1–10, 2006.

[2] M. Zeisberg, J. I. Hanai, H. Sugimoto et al., "BMP-7 counteracts TGF-β1-induced epithelial-to-mesenchymal transition and reverses chronic renal injury," *Nature Medicine*, vol. 9, no. 7, pp. 964–968, 2003.

[3] Y. Kojima, A. Acar, E. N. Eaton et al., "Autocrine TGF-β and stromal cell-derived factor-1 (SDF-1) signaling drives the evolution of tumor-promoting mammary stromal myofibroblasts," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 46, pp. 20009–20014, 2010.

[4] L. Y. Luo, A. Soosaipillai, and E. P. Diamandis, "Molecular cloning of a novel human gene on chromosome 4p11 by immunoscreening of an ovarian carcinoma cDNA library," *Biochemical and Biophysical Research Communications*, vol. 280, no. 1, pp. 401–406, 2001.

[5] F. Gueugnon, S. Leclercq, C. Blanquart et al., "Identification of novel markers for the diagnosis of malignant pleural mesothelioma," *The American Journal of Pathology*, vol. 178, no. 3, pp. 1033–1042, 2011.

[6] M. Kulis, S. Heath, M. Bibikova et al., "Epigenomic analysis detects widespread gene-body dna hypomethylation in chronic lymphocytic leukemia," *Nature Genetics*, vol. 44, no. 11, pp. 1236–1242, 2012.

[7] H. Noushmehr, D. J. Weisenberger, K. Diefes et al., "Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma," *Cancer Cell*, vol. 17, no. 5, pp. 510–522, 2010.

[8] H. Kim, J. Watkinson, V. Varadan, and D. Anastassiou, "Multicancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1," *BMC Medical Genomics*, vol. 3, article 51, 2010.

[9] C. Jiang, Z. Xuan, F. Zhao, and M. Q. Zhang, "TRED: a transcriptional regulatory element database, new entries and other development," *Nucleic Acids Research*, vol. 35, no. 1, pp. 137–140, 2007.

[10] S. Peri, J. D. Navarro, T. Z. Kristiansen et al., "Human protein reference database as a discovery resource for proteomics," *Nucleic Acids Research*, vol. 32, pp. 497–501, 2004.

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

[12] Ingenuity Systems, http://www.ingenuity.com/.

[13] S. Wernicke and F. Rasche, "FANMOD: A tool for fast network motif detection," *Bioinformatics*, vol. 22, no. 9, pp. 1152–1153, 2006.

[14] G. Jin, K. Cui, X. Zhou, and S. T. C. Wong, "Unraveling the signal-transduction networks in cancer metastasis," *IEEE Signal Processing Magazine*, vol. 26, no. 5, pp. 129–132, 2009.

[15] F. Schreiber and H. Schwöbbermeyer, "Motifs in biological networks," 2008.

[16] M. Ogawa, A. C. LaRue, and C. J. Drake, "Hematopoietic origin of fibroblasts/myofibroblasts: its pathophysiologic implications," *Blood*, vol. 108, no. 9, pp. 2893–2896, 2006.

[17] J. Stegmüller, M. A. Huynh, Z. Yuan, Y. Konishi, and A. Bonni, "TGFβ-Smad2 signaling regulates the Cdh1-APC/SnoN pathway of axonal morphogenesis," *Journal of Neuroscience*, vol. 28, no. 8, pp. 1961–1969, 2008.

[18] J. Varani, M. J. Bendelow, D. E. Sealey et al., "Tumor necrosis factor enhances susceptibility of vascular endothelial cells to neutrophil-mediated killing," *Laboratory Investigation*, vol. 59, no. 2, pp. 292–295, 1988.

[19] T. Amann, Y. Egle, A.-K. Bosserhoff, and C. Hellerbrand, "FHL2 suppresses growth and differentiation of the colon cancer cell line HT-29," *Oncology Reports*, vol. 23, no. 6, pp. 1669–1674, 2010.

[20] K. B. Lee, J. H. Jeon, I. Choi, O. Y. Kwon, K. Yu, and K. H. You, "Clusterin, a novel modulator of TGF-β signaling, is involved in Smad2/3 stability," *Biochemical and Biophysical Research Communications*, vol. 366, no. 4, pp. 905–909, 2008.

[21] T. Vincent, E. P. A. Neve, J. R. Johnson et al., "A Snail1-Smad3/4 transcriptional repressor complex promotes TGF-β mediated epithelial-mesenchymal transition," *Nature Cell Biology*, vol. 11, no. 8, pp. 943–950, 2009.

[22] P. Papageorgis, A. W. Lambert, S. Ozturk et al., "Smad signaling is required to maintain epigenetic silencing during breast cancer progression," *Cancer Research*, vol. 70, no. 3, pp. 968–978, 2010.

[23] S. Matsumura, I. Imoto, K. Kozaki et al., "Integrative array-based approach identifies mzb1 as a frequently methylated putative tumor suppressor in hepatocellular carcinoma," *Clinical Cancer Research*, vol. 18, no. 13, pp. 3541–3551, 2012.

[24] S. I. Berndt, W. Y. Huang, N. Chatterjee et al., "Transforming growth factor beta 1 (TGFB1) gene polymorphisms and risk of advanced colorectal adenoma," *Carcinogenesis*, vol. 28, no. 9, pp. 1965–1970, 2007.

[25] H. Y. Lan, "Diverse roles of TGF-β/smads in renal fibrosis and inflammation," *International Journal of Biological Sciences*, vol. 7, no. 7, pp. 1056–1067, 2011.

[26] H. Y. Kang, H. K. Lin, Y. C. Hu, S. Yeh, K. E. Huang, and C. Chang, "From transforming growth factor-β signaling to androgen action: Identification of Smad3 as an androgen receptor coregulator in prostate cancer cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 6, pp. 3018–3023, 2001.

[27] Y. Niu, S. Altuwaijri, K. P. Lai et al., "Androgen receptor is a tumor suppressor and proliferator in prostate cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 34, pp. 12182–12187, 2008.

[28] H. Y. Kang, K. E. Huang, S. Y. Chang, W. L. Ma, W. J. Lin, and C. Chang, "Differential modulation of androgen receptor-mediated transactivation by smad3 and tumor suppressor smad4," *Journal of Biological Chemistry*, vol. 277, no. 46, pp. 43749–43756, 2002.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 396727, 9 pages
http://dx.doi.org/10.1155/2014/396727

Research Article

Gender-Specific DNA Methylome Analysis of a Han Chinese Longevity Population

Liang Sun,[1] Jie Lin,[2,3] Hongwu Du,[4] Caiyou Hu,[5] Zezhi Huang,[6] Zeping Lv,[5] Chenguang Zheng,[7] Xiaohong Shi,[1] Yan Zhang,[2] and Ze Yang[1]

[1] *The Key Laboratory of Geriatrics, Beijing Hospital and Beijing Institute of Geriatrics, Ministry of Health, Beijing 100730, China*
[2] *Key Laboratory of Nutrition and Metabolism, Institute for Nutritional Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai 200031, China*
[3] *Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai 200031, China*
[4] *University of Science and Technology Beijing, Beijing 100083, China*
[5] *Department of Neurology, Jiangbin Hospital, Nanning, Guangxi 530021, China*
[6] *Yongfu Committee of the Chinese People's Political Consultative Conference, Yongfu, Guangxi 541800, China*
[7] *Department of Cardiothoracic Surgery, Guangxi Maternal and Child Health Hospital, Nanning, Guangxi 530003, China*

Correspondence should be addressed to Yan Zhang; yanzhang01@sibs.ac.cn and Ze Yang; yang.ze@live.cn

Received 28 November 2013; Accepted 28 February 2014; Published 14 April 2014

Academic Editor: Jean X. Gao

Human longevity is always a biological hotspot and so much effort has been devoted to identifying genes and genetic variations associated with longer lives. Most of the demographic studies have highlighted that females have a longer life span than males. The reasons for this are not entirely clear. In this study, we carried out a pool-based, epigenome-wide investigation of DNA methylation profiles in male and female nonagenarians/centenarians using the Illumina 450 K Methylation Beadchip assays. Although no significant difference was detected for the average methylation levels of examined CpGs (or probes) between male and female samples, a significant number of differentially methylated probes (DMPs) were identified, which appeared to be enriched in certain chromosome regions and certain parts of genes. Further analysis of DMP-containing genes (named DMGs) revealed that almost all of them are solely hypermethylated or hypomethylated. Functional enrichment analysis of these DMGs indicated that DNA hypermethylation and hypomethylation may regulate genes involved in different biological processes, such as hormone regulation, neuron projection, and disease-related pathways. This is the first effort to explore the gender-based methylome difference in nonagenarians/centenarians, which may provide new insights into the complex mechanism of longevity gender gap of human beings.

1. Introduction

Over the last 100 years, humans experienced a huge increase of life expectancy. These advances were largely driven by extrinsic improvements of their living environment (such as diet and disease prevalence) as well as genetic variations (such as polymorphism and DNA methylation). Since human aging and longevity is a very complex trait where environmental, genetic, and stochastic factors are involved, it has largely aroused the attention of scientists around the world. A great number of studies have been carried out to investigate the mechanisms and key factors that may influence human mortality, aging, and lifespan [1–5].

As specific cohorts, nonagenarians and centenarians are always considered as the most valuable models to study the mechanisms involved in human aging and longevity [6]. They are considered to have reached the extreme limits of human life span but still show relatively good health conditions to maintain physiological function and escape the common fatal diseases [7]. Despite the increasing numbers of very old

people worldwide, both nonagenarians and centenarians are still few from a demographic point of view. For example, in USA, nonagenarians now represent ~4.7% of the 65-and-older population whereas centenarians account for 17.3 per 100,000 people. Thus, it should be important to understand genetic and other factors, as well as the ways involved in healthy aging and longevity.

Currently, the majority of genome-based studies focused on the association between longevity and sequence variations including single nucleotide polymorphism or copy number variation [7–12]. In addition, epigenetic regulations, such as DNA methylation and histone modification, have emerged as a key driver of cell fate and their disruption could be related to a variety of human diseases [13–17]. Furthermore, with the development of genome-wide epigenetic analysis, much work has been carried out on epigenetic mechanisms of genome regulation during aging [18–20]. For example, heritable changes to the epigenome at both early and late life stages [21, 22], immune system/tissues specific variations [23–25], and dynamic epigenetic modifications through the lifespan [26, 27] have been reported to be responsible for many biological processes during healthy aging and longevity. Very recently, Heyn et al. found that the centenarian DNA had a genome-wide lower DNA methylation content and a reduced correlation in the methylation status of neighboring cytosine-phosphate-guanosine (CpG) sites in comparison with the newborn DNA [28]. This study demonstrated for the first time that the DNA methylomes at the two extremes of the human lifespan are distinct.

A significant trend observed in most parts of the world is that females have a longer life span than males. In particular, when nonagenarians and centenarians are considered, the male/female ratio has been reported to range between 1 : 4 and 1 : 7 [29]. Such a gender gap is quite remarkable, which has challenged scientists for decades to investigate possible reasons, such as better living conditions, specific biological advantages, and fewer behaviors that are bad for health compared to men [30–32]. A number of genome-based studies have been carried out to identify factors that may influence the gender difference based on animal models [33–35]. Recently, researchers have started to analyze gender-based genetic variations using human samples [36, 37]. It was suggested that the role of gender in the regulation of longevity may be linked to gender-specific genetic differences, such as the expression of sex hormone patterns and the changes in these patterns during lifetime [38]. However, so far it is difficult to collect enough samples to conduct a population-based longevity study. Moreover, gender-based DNA methylation analysis of the longevity population is not yet available, which may provide useful information with respect to epigenetic regulation of the longevity gender gap.

China has the largest population of adults aged 60+ years in the world [39]. In South China, there are several "longevity counties" due to the high number of nonagenarians and centenarians living there, such as Yongfu County, which has been qualified as the "Longevity Town" by Geriatric Society of China in 2007. In this study, a total of 200 Han nationality nonagenarian/centenarian participants (100 men and 100 women) from Yongfu County were recruited. We used a pool-based strategy to perform epigenome-wide investigation of DNA methylation profiles in male and female cohorts using the Illumina 450 K Methylation Beadchip. Differentially methylated CpGs between male and female samples and related genes were identified. To our knowledge, this is the first effort with such a large sample size of nonagenarians/centenarians to study the methylome difference that may contribute to the longevity gender gap.

2. Materials and Methods

2.1. Subjects. This project is an extension of the "Longevity and Health of Aging Population in Guangxi China" project conducted in 2008 and 2010 [40]. One hundred pairs of geography and nationality matched male and female volunteers aged 95+ years from urban and rural areas of Yongfu County, South China, were enrolled after exclusion of the subjects undertaking drug treatment. The male group (mean age 97.34 ± 2.66 years) was comprised of 94 nonagenarians (aged 95–99 years) and 6 centenarians (aged 100–105 years). The female group (mean age 99.14 ± 2.20 years) was comprised of 80 nonagenarians (aged 95–99 years) and 20 centenarians (aged 100–106 years). All subjects self-reported as Han nationality. The study was conducted according to the principles expressed in the Declaration of Helsinki. The Ethics Committee of Beijing Hospital, Ministry of Health, approved the study protocol. After the protocol was explained to the subjects, they provided written informed consent.

2.2. Genomic DNA Isolation and Pooling. Peripheral blood mononuclear cells (PBMCs) were isolated from whole blood for genomic DNA extraction using the Qiagen mini kit (Qiagen, Germany) following the manufacturer's protocol. DNA concentrations were determined by NanoDrop microvolume quantitation assay and 1% agarose electrophoresis. After validation of quality and integrity of individual genomic DNA, we equally pooled each sample into male and female groups, respectively.

2.3. Genome-Wide DNA Methylation Assay. The prepared genomic DNA (0.5 μg) was bisulfate-converted with the EZ DNA Methylation Gold kit (Zymo Research, USA). After bisulfite conversion, each pooled sample was whole-genome amplified, enzymatically fragmented, precipitated, resuspended, and hybridized at 48°C for 16 h to Illumina Human Methylation 450 K BeadChip containing 485,577 locus-specific oligonucleotide primers. The probes were distributed among 20,216 transcripts, potential transcripts, or isolated CpG islands (CGIs). IlluminaHiScan SQ scanner was used for detection by fluorescent single-base primer extension assay. The methylation score is represented as β-value, a continuous parameter between 0 and 1 to show the ratio of the methylated-probe signal to total locus signal intensity. CpGs with a detection P value (representing the measured signal compared to negative controls) >0.05 were removed from the raw data. Raw data were further normalized using Illumina's control probe scaling procedure and background subtraction [41].

2.4. Differential Methylation Analysis. As described above, the measurement of whole genome DNA methylation used a pool-based approach, in which the β-value of each probe represents the average methylation level among all samples in the pool. To avoid sex-biased DNA methylation differences, we excluded methylation data for the X and Y chromosomes (473,864 probes remained). To identify differentially methylated probes (DMPs), we first assumed that the methylation levels for the whole genome obey a Gaussian model which could be used to predict DMPs [42]. Two simulation data sets which follow the Gaussian model with the same mean, standard variation, and sample size for male and female were randomly created, respectively. Then we calculated the different degrees of methylation changes between the two data sets (Δ_{Me}, female-male). Since Δ_{Me} obeys the normal distribution, we built the normal distribution with the same mean and variance of Δ_{Me} and calculated the prediction intervals corresponding to a P value < 0.05. After repeating this process 1000 times, the cutoff of 95% confidence interval was 0.197 ± 0.067. Therefore, a threshold of a 0.20 of Δ_{Me} was finally used to identify DMPs.

2.5. Bioinformatics Analysis. Gene ontology (GO) and KEGG (Kyoto encyclopedia of genes and genomes) pathway enrichment analyses were conducted using the *R* (version 2.14.0) package GOstats (version 2.28.0) [43]. GO terms and KEGG information were downloaded from Bioconductor (http://www.bioconductor.org). The P value was initially calculated based on hypergeometric distribution and filtered by adjusted P value < 0.05. Multiple comparison adjustment was applied to get the adjusted P value using the false discovery rate (FDR) approach by *R* [44, 45].

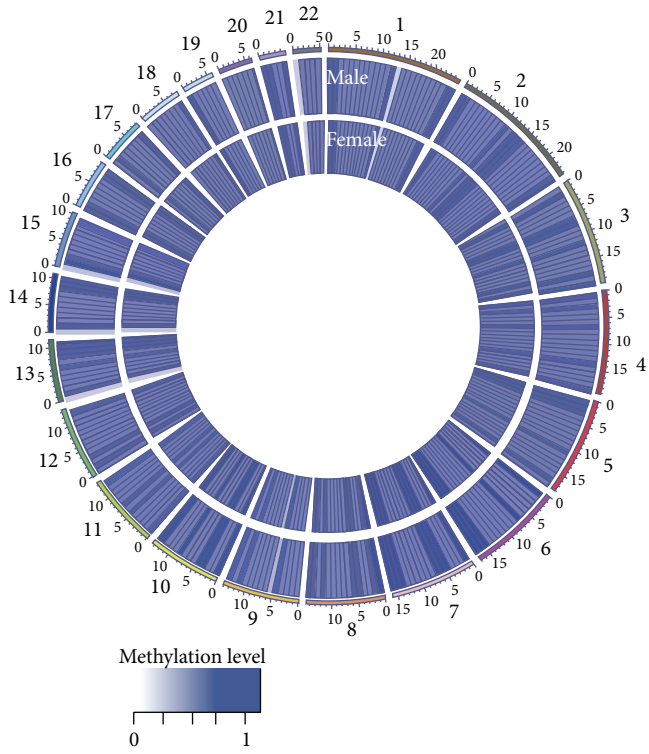


FIGURE 1: General view of DNA methylation level for autosomes of Chinese nonagenarians/centenarians. The average genome-wide DNA methylation levels in male and female samples are represented using Circos. The inner and outer tracks indicate the average methylation levels for female and male samples, respectively. All autosomes are represented via 10 Mbp-wide windows. The average methylation level in each region represents the average β-value (0-1) for all the probes in this region.

3. Results and Discussion

It has been suggested that DNA pooling allows accurate assessment of average DNA methylation in large groups of individual genomes [46, 47]. Here, we applied this strategy to compare genome-wide methylation patterns between male and female groups of Han Chinese nonagenarians/centenarians.

3.1. General Analysis of DNA Methylomes of the Chinese Longevity Population. A general view of whole genome methylation profiles in autosomes of male and female nonagenarians/centenarians from Yongfu County in China was shown in Figure 1, using Circos software [48]. It appeared that the majority of the methylated regions have quite similar methylation patterns between male and female samples, implying that DNA methylation-based epigenetic profiles might be mostly common and gender-independent in the longevity population. Further analysis of the average methylation level of all examined CpGs confirmed that there is no significant difference between male samples (0.4962) and female ones (0.4974) at the whole genome level (using t test, $P = 0.2452$). However, a significant number of gender-specific DNA methylation differences between male and female samples were identified, which might play a role in gender-specific life span extension, for example, different gene expression regulation.

3.2. Identification of Differentially Methylated Probes and Related Genes. The discrepancies of DNA methylomes between male and female nonagenarians/centenarians prompted us to search for particular DMPs. In this study, DMPs were predicted based on a Gaussian model with the same mean and variation of Δ_{Me} (see Section 2). Using the male samples as control, hypermethylated and hypomethylated probes in female samples were selected if $\Delta_{\text{Me}} > 0.2$ or < -0.2, respectively. Based on these criteria, we identified 850 DMPs (0.179% of all examined CpGs in autosomes), which are illustrated in Figure 2(a). These DMPs appeared to be enriched in certain genomic regions, especially in chromosome 17, which has been reported to contain many disease-associated genes [49]. This interesting finding implied that DMPs enriched in these chromosomal regions may play an important role in longevity gender gap.

We further examined the location of DMPs based on different parts of genes: 1500 bp above transcription start site (TSS1500), 200 bp above TSS (TSS200), 5′ untranslated region (5′-UTR), the 1st exon, gene body (other exons except

Table 1: GO analysis of hypomethylated and hypermethylated DMGs.

GO ID	Description	P value	FDR
	Hypomethylated genes		
Biological process			
GO:0016043	Cellular component organization	$5.03E-07$	$1.06E-03$
GO:0071840	Cellular response to vitamin A	$1.74E-06$	$1.83E-03$
GO:0071299	Cell surface receptor linked signaling pathway	$2.27E-04$	$2.36E-02$
GO:0007166	Regulation of hormone levels	$3.95E-04$	$1.36E-02$
GO:0010817	Cell projection organization	$4.08E-04$	$2.36E-02$
GO:0030030	Hormone secretion	$4.96E-04$	$2.36E-02$
GO:0046879	Cellular component organization at cellular level	$6.80E-04$	$4.36E-02$
GO:0071842	Cellular response to vitamin	$6.81E-04$	$4.36E-02$
GO:0071295	Regulation of transcription from RNA polymerase II promoter	$7.23E-04$	$4.36E-02$
GO:0006357	Epithelial cell development	$8.35E-04$	$4.59E-02$
GO:0002064	Hormone transport	$8.52E-04$	$4.59E-02$
GO:0009914	Production of molecular mediator involved in inflammatory response	$8.54E-04$	$4.59E-02$
GO:0002532	Cell morphogenesis involved in differentiation	$8.77E-04$	$4.59E-02$
GO:0000904	Transmembrane receptor protein tyrosine kinase signaling pathway	$9.59E-04$	$4.94E-02$
GO:0007169	Wnt receptor signaling pathway	$9.71E-04$	$4.94E-02$
Cellular component			
GO:0015629	Actin cytoskeleton	$1.94E-03$	$4.99E-02$
	Hypermethylated genes		
Biological process			
GO:0000902	Cell morphogenesis	$3.23E-05$	$3.54E-02$
GO:0021955	Central nervous system neuron axonogenesis	$3.39E-05$	$3.54E-02$
GO:0048869	cellular developmental process	$4.21E-05$	$3.54E-02$
GO:0032989	cellular component morphogenesis	$7.81E-05$	$3.73E-02$
GO:0048858	cell projection morphogenesis	$8.50E-05$	$3.73E-02$
GO:0032990	cell part morphogenesis	$1.01E-04$	$3.73E-02$
GO:0051179	localization	$1.04E-04$	$3.73E-02$
GO:0048667	cell morphogenesis involved in neuron differentiation	$1.26E-04$	$3.96E-02$
GO:0030154	cell differentiation	$1.71E-04$	$4.80E-02$
Cellular component			
GO:0044459	plasma membrane part	$1.05E-05$	$1.99E-03$
GO:0016020	membrane	$3.19E-05$	$1.99E-03$
GO:0044425	membrane part	$4.55E-05$	$1.99E-03$
GO:0005911	cell-cell junction	$4.67E-05$	$2.16E-03$
GO:0030054	cell junction	$6.34E-05$	$4.65E-03$
Molecular function			
GO:0015108	chloride transmembrane transporter activity	$3.46E-05$	$1.99E-03$
GO:0015103	inorganic anion transmembrane transporter activity	$6.03E-05$	$2.16E-03$
GO:0015296	anion:cationsymporter activity	$1.17E-04$	$2.65E-02$
GO:0004714	glycoprotein binding	$1.71E-04$	$4.80E-02$

Table 2: KEGG pathway enrichment analysis of hypomethylated and hypermethylated DMGs.

KEGG ID	Description	P value	FDR
Hypomethylated genes			
KEGG:04512	ECM-receptor interaction	$2.50E-02$	$7.50E-03$
Hypermethylated genes			
KEGG:04360	Axon guidance	$3.7E-3$	$1.48E-02$
KEGG:04514	Cell adhesion molecules (CAMs)	$1.6E-2$	$3.20E-02$

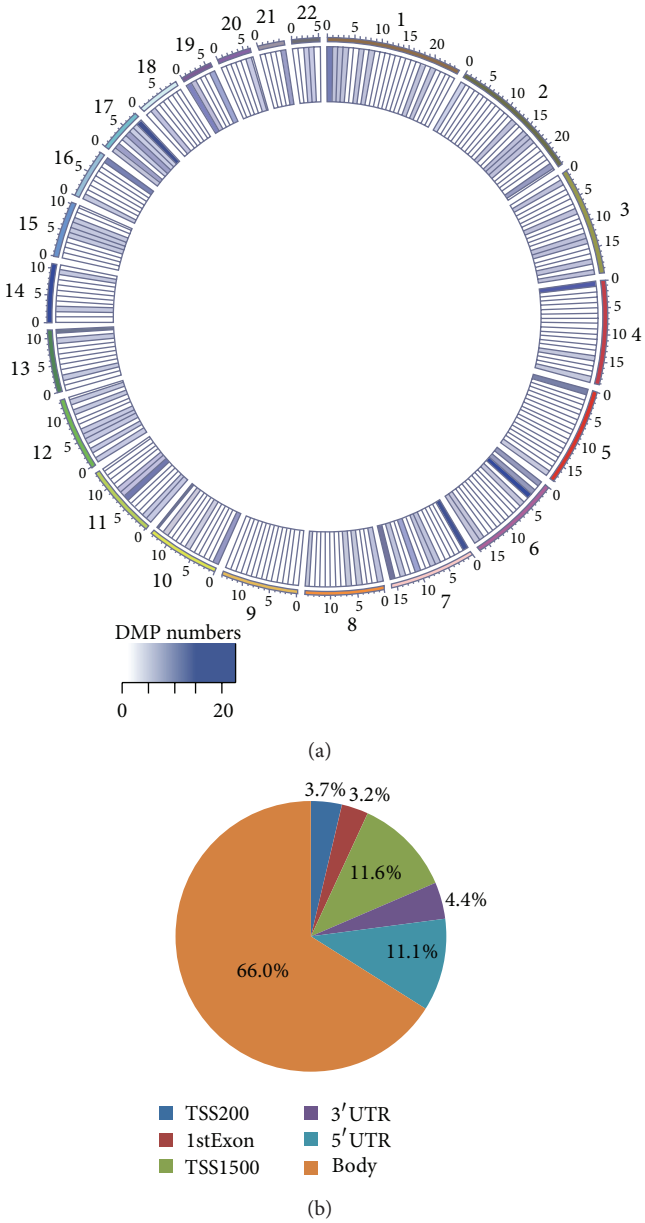


Figure 2: Distribution of DMPs between male and female samples. (a) Circos representation of the total number of DMPs in each region. The number is calculated using 10 Mbp-wide windows for each autosome. (b) Distribution of DMPs according to different regions of genes.

the 1st exon), and 3′-UTR (Figure 2(b)). Most DMPs were enriched in gene body (66.0%). Although it has been reported that the methylation level of CpGs in coding region may regulate gene transcriptional activity [50], it is unclear whether DMPs detected in this study could affect the expression of corresponding genes. On the other hand, 15.4% DMPs were observed in the potential promoter regions (TSS1500 + TSS200) of genes. Thus, it is possible that some of these DMPs may be related to distinct expression difference of certain genes between men and women.

We also analyzed the trend of methylation changes of DMPs. The majority of DMPs (54.5%) were hypermethylated in female compared to those in male samples (Figure 3(a)). Further analysis of different parts of genes revealed that, except the first exon, there were more hypermethylated DMPs than hypomethylated DMPs in all parts of genes (Figure 3(b)). These results implied that a more significant trend of DNA hypermethylation in females may be related to the gender gap in life expectancy.

To investigate the potential relationship between DMPs and genes, all DMPs were mapped to 564 genes (named differential methylated genes or DMGs; see Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/396727). Here, a hypermethylated DMG was defined if it only contains hypermethylated DMPs. Similarly, a hypomethylated gene was defined if only hypomethylated DMPs were detected. If a gene contains both hypermethylated and hypomethylated DMPs, it was considered as a "mixed" DMG. In this study, 54.4%, 42.6%, and 3.0% of DMGs were found to belong to hypermethylated, hypomethylated, and mixed DMG groups. Thus, it appears that almost all DMGs have remained with a consistent trend of methylation changes.

It is known that epigenetic changes may affect the aging process and may be one of the central mechanisms of many age-related diseases [51]. In addition, it has also been reported that human disease genes are much closer to aging genes than expected by chance [52]. To investigate the potential relationship between DMGs detected in this study and aging or disease genes, we compared DMGs with known human aging genes and disease genes (provided in [52]), respectively. Few common genes could be found for both aging and disease genes (Tables S2 and S3), suggesting that the longevity gender gap might be unrelated to known aging or disease-related genes or processes. In other words, male and female longevities may share similar antiaging or antidisease mechanisms.

3.3. GO and KEGG Functional Enrichment Analysis of Differentially Methylated Genes. To extrapolate the biological processes of DMGs, a *R* package GOstats [43] was used to perform GO term and KEGG pathway enrichment analyses. Interestingly, no significant overlaps of GO terms could be found between hypomethylated and hypermethylated DMGs (Table 1).

Hypomethylated DMGs were mainly enriched in cellular component organization, cell surface receptor signaling, hormone regulation, and some disease-related pathways (such as Wnt receptor signaling pathway). It has been known for a long time that Wnt signaling pathway may lead to tumor development [53–55] and ROS-induced damage [56]. Some of the DMGs, such as chloride channel 7 (CLCN7), alpha-1 type I collagen (COL1A1), and estrogen receptor 1 (Esr1), are known to be associated with osteoporosis and fractures that are more common in women [57, 58]. Thus, hypomethylation of these genes may help extend the life span of women. In addition, some of these DMGs are involved in maintenance of cellular homeostasis (such as GO:0071840,

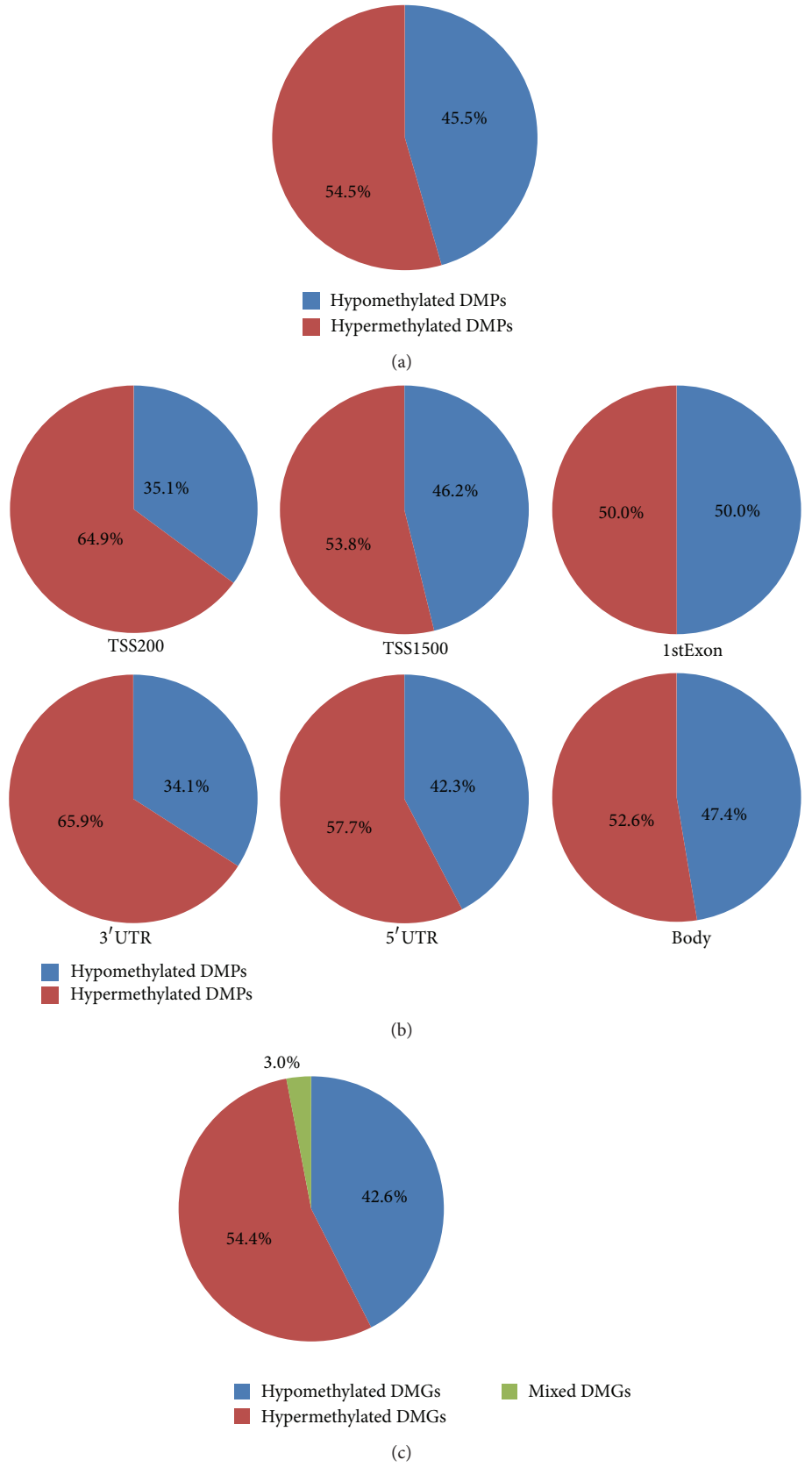


Figure 3: Distribution of hypermethylated and hypomethylated DMPs and DMGs. (a) Distribution of hypermethylated and hypomethylated DMPs. (b) Distribution of hypermethylated and hypomethylated DMPs in different parts of genes. (c) Distribution of hypermethylated, hypomethylated, and mixed DMGs.

GO:0071299, GO:0071842, and GO:0000904). It has been reported that a reduced homeostatic ability in response to internal or external stimuli may increase the occurrence of many diseases even death [59].

On the other hand, hypermethylated DMGs were found to be enriched in cell morphogenesis, cell-cell junction, and cell projection. Some of the enriched biological processes are known to be involved in neuron projection and central nervous system development [60]. In addition, the insulin-like growth factor 2 receptor (IGF2R) enriched in GO:0004714 (glycoprotein binding) could reduce the insulin resistance, which has been reported to increase life span more in females than in males [61, 62]. Our results suggested that DNA hypomethylation and hypermethylation may regulate corresponding genes involved in different processes. Therefore, identification of gender-specific methylation patterns may provide important information regarding the possible epigenetic mechanisms of longevity gender gap.

KEGG analysis showed that very few pathways could be significantly enriched for either hypomethylated or hypermethylated DMGs (Table 2). Hypomethylated DMGs were only enriched in extracellular matrix- (ECM-) receptor interaction (KEGG:04512), whereas hypermethylated DMGs were enriched in axon guidance (KEGG:04360) and cell adhesion molecules (CAMs) (KEGG:04514). The relationship between these pathways and gender-specific longevity is not clear.

3.4. Investigation of Hypomethylation Status of X Chromosome. Although the *X* chromosome contributes to the gender-specific methylation discrepancies, we could not analyze the complete methylation data for this chromosome because the "silent" *X* chromosome may cause bias when analyzing the DNA methylation level in women using our criteria. However, we could still identify its hypomethylated DMPs. In this study, we found 185 hypomethylated DMPs that correspond to 95 hypomethylated DMGs. Most of the hypomethylated DMPs were enriched in gene body (34.0%) and 5′-UTR (36.1%). Some DMGs either are known to be associated with hormonal effects (such as androgen receptor, AR) or have been considered as age-related genes in human cerebral cortex [63]. This observation suggested that *X*-linked hypomethylated DMGs may contribute to gender gap of human longevity, which is consistent with the hypothesis that gender-specific regulation of longevity may be related to the expression of sex hormone patterns [38]. Similar to autosomes, few common genes could be found while comparing the DMGs in *X* chromosome with known aging and disease genes (Tables S2 and S4).

3.5. Ongoing Work: Comparative DNA Methylome Analysis of Adults and Nonagenarians/Centenarians. We are collecting samples for examining the DNA methylomes of male and female adults from Yongfu County to identify epigenetic patterns that may be related to the mechanisms of longevity in the Han Chinese population. Based on our preliminary data, less DMPs in autosomes and more DMPs in sex chromosomes were observed in adults than in nonagenarians/centenarians (unpublished data). The average methylation level was significantly higher in both male and female adult samples compared to those in nonagenarians/centenarians, implying that hypomethylation in certain genomic regions may be related to longer life span. Hormone regulation and cell morphogenesis regulation appeared to be important for gender-specific longevity. These findings may help us figure out the difference of aging process between male and female. However, as longevity is a very complex trait, it should be understood that reliance on a single aspect of genetics has its limitations. In the future, by increasing the sample size, generating different levels of data, and developing more reliable methods, these defects may be rectified, providing scientists with more opportunities to explore in detail the role of DNA methylation and other factors involved in gender-specific longevity.

4. Conclusions

In summary, we are unique in reporting a comprehensive comparison of DNA methylome between male and female nonagenarians/centenarians in a Han Chinese population. The average methylation level in female samples was similar to that in male samples in spite of the fact that a significant number of DMPs were identified. These DMPs prefer to be enriched in certain chromosome regions. Further analysis of DMPs in different parts of genes revealed that most of them are located in gene body regions. Almost all of the DMGs are solely hypermethylated or hypomethylated. Functional enrichment analysis of these genes revealed that DNA hypermethylation and hypomethylation may regulate genes involved in different processes or pathways, some of which may contribute to the gender gap of life span. In addition, identification of *X*-based hypomethylated probes and genes could provide evidence for better understanding of the mechanism of longer lives in females.

Conflict of Interests

The authors have declared that no competing interests exist.

Authors' Contribution

Liang Sun and Jie Lin contributed equally to the work.

Acknowledgments

This work was supported by Beijing Nova program (Z121107002512058), National Natural Science Foundation of China (no. 81370445, 31171233), and a grant from Chinese Academy of Sciences (CAS) (2012OHTP10).

References

[1] T. Iannitti and B. Palmieri, "Inflammation and genetics: an insight in the centenarian model," *Human Biology*, vol. 83, no. 4, pp. 531–559, 2011.

[2] S. Salvioli, F. Olivieri, F. Marchegiani et al., "Genes, ageing and longevity in humans: problems, advantages and perspectives," *Free Radical Research*, vol. 40, no. 12, pp. 1303–1323, 2006.

[3] M. Capri, S. Salvioli, F. Sevini et al., "The genetics of human longevity," *Annals of the New York Academy of Sciences*, vol. 1067, no. 1, pp. 252–263, 2006.

[4] S. Salvioli, M. Capri, A. Santoro et al., "The impact of mitochondrial DNA on human lifespan: a view from studies on centenarians," *Biotechnology Journal*, vol. 3, no. 6, pp. 740–749, 2008.

[5] J. Vijg and Y. Suh, "Genome instability and aging," *Annual Review of Physiology*, vol. 75, pp. 645–668, 2013.

[6] C. Franceschi and M. Bonafè, "Centenarians as a model for healthy aging," *Biochemical Society Transactions*, vol. 31, no. 2, pp. 457–461, 2003.

[7] A. R. Brooks-Wilson, "Genetics of healthy aging and longevity," *Human Genetics*, vol. 132, no. 12, pp. 1323–1338, 2013.

[8] W.-H. Chung, R.-L. Dao, L.-K. Chen, and S.-I. Hung, "The role of genetic variants in human longevity," *Ageing Research Reviews*, vol. 9, supplement, pp. S67–S78, 2010.

[9] G. M. Martin, A. Bergman, and N. Barzilai, "Genetic determinants of human health span and life span: progress and new opportunities," *PLoS Genetics*, vol. 3, no. 7, p. e125, 2007.

[10] J. M. Murabito, R. Yuan, and K. L. Lunetta, "The search for longevity and healthy aging genes: insights from epidemiological studies and samples of long-lived individuals," *Journals of Gerontology A: Biological Sciences and Medical Sciences*, vol. 67, no. 5, pp. 470–479, 2012.

[11] P. Sebastiani, N. Solovieff, A. Puca et al., "Genetic signatures of exceptional longevity in humans," *Science*, vol. 2010, 2010.

[12] M. Kuningas, K. Estrada, Y.-H. Hsu et al., "Large common deletions associate with mortality at old age," *Human Molecular Genetics*, vol. 20, no. 21, Article ID ddr340, pp. 4290–4296, 2011.

[13] K. D. Robertson, "DNA methylation and human disease," *Nature Reviews Genetics*, vol. 6, no. 8, pp. 597–610, 2005.

[14] D. D. de Carvalho, J. S. You, and P. A. Jones, "DNA methylation and cellular reprogramming," *Trends in Cell Biology*, vol. 20, no. 10, pp. 609–617, 2010.

[15] M. L. Suvà, N. Riggi, and B. E. Bernstein, "Epigenetic reprogramming in cancer," *Science*, vol. 339, no. 6127, pp. 1567–1570, 2013.

[16] V. K. Rakyan, T. A. Down, D. J. Balding, and S. Beck, "Epigenome-wide association studies for common human diseases," *Nature Reviews Genetics*, vol. 12, no. 8, pp. 529–541, 2011.

[17] Z. Wen, Z. P. Liu, Z. Liu, Y. Zhang, and L. Chen, "An integrated approach to identify causal network modules of complex diseases with application to colorectal cancer," *Journal of the American Medical Informatics Association*, vol. 20, no. 4, pp. 659–667, 2013.

[18] A. R. Mendelsohn and J. W. Larrick, "The DNA methylome as a biomarker for epigenetic instability and human aging," *Rejuvenation Research*, vol. 16, no. 1, pp. 74–77, 2013.

[19] S. Han and A. Brunet, "Histone methylation makes its mark on longevity," *Trends in Cell Biology*, vol. 22, no. 1, pp. 42–49, 2012.

[20] J. D. Boyd-Kirkup, C. D. Green, G. Wu, D. Wang, and J. D. Han, "Epigenomics and the regulation of aging," *Epigenomics*, vol. 5, no. 2, pp. 205–227, 2013.

[21] H. Cedar and Y. Bergman, "Programming of DNA methylation patterns," *Annual Review of Biochemistry*, vol. 81, pp. 97–117, 2012.

[22] P. D'Aquila, G. Rose, D. Bellizzi, and G. Passarino, "Epigenetics and aging," *Maturitas*, vol. 74, no. 2, pp. 130–136, 2013.

[23] A. Saini, S. Mastana, F. Myers, and M. P. Lewis, "'From death, lead me to immortality'—mantra of ageing skeletal muscle," *Current Genomics*, vol. 14, no. 4, pp. 256–267, 2013.

[24] D. G. Hernandez, M. A. Nalls, J. R. Gibbs et al., "Distinct DNA methylation changes highly correlated with chronological age in the human brain," *Human Molecular Genetics*, vol. 20, no. 6, pp. 1164–1172, 2011.

[25] C. R. Balistreri, G. Candore, G. Accardi et al., "Genetics of longevity. Data from the studies on Sicilian centenarians," *Immunity & Ageing*, vol. 9, no. 1, p. 8, 2012.

[26] D. Gentilini, D. Mari, D. Castaldi et al., "Role of epigenetics in human aging and longevity: genome-wide DNA methylation profile in centenarians and centenarians' offspring," *Age*, vol. 35, no. 5, pp. 1961–1973, 2013.

[27] Y. Bergman and H. Cedar, "DNA methylation dynamics in health and disease," *Nature Structural & Molecular Biology*, vol. 20, no. 3, pp. 274–281, 2013.

[28] H. Heyn, N. Li, H. J. Ferreira et al., "Distinct DNA methylomes of newborns and centenarians," *Proceedings of the National Academy of Sciences USA*, vol. 109, no. 26, pp. 10522–10527, 2012.

[29] G. Passarino, C. Calignano, A. Vallone et al., "Male/female ratio in centenarians: a possible role played by population genetic structure," *Experimental Gerontology*, vol. 37, no. 10-11, pp. 1283–1289, 2002.

[30] J. Viña and C. Borrás, "Women live longer than men: understanding molecular mechanisms offers opportunities to intervene by using estrogenic compounds," *Antioxidants & Redox Signaling*, vol. 13, no. 3, pp. 269–278, 2010.

[31] E. L. B. Barrett and D. S. Richardson, "Sex differences in telomeres and lifespan," *Aging Cell*, vol. 10, no. 6, pp. 913–921, 2011.

[32] R. C. May, "Gender, immunity and the regulation of longevity," *BioEssays*, vol. 29, no. 8, pp. 795–802, 2007.

[33] J. Tower, "Sex-specific regulation of aging and apoptosis," *Mechanisms of Ageing and Development*, vol. 127, no. 9, pp. 705–718, 2006.

[34] Y.-F. Chen, C.-Y. Wu, C.-H. Kao, and T.-F. Tsai, "Longevity and lifespan control in mammals: lessons from the mouse," *Ageing Research Reviews*, vol. 9, supplement, pp. S28–S35, 2010.

[35] A. U. Jackson, A. T. Galecki, D. T. Burke, and R. A. Miller, "Mouse loci associated with life span exhibit sex-specific and epistatic effects," *Journals of Gerontology A: Biological Sciences and Medical Sciences*, vol. 57, no. 1, pp. B9–B15, 2002.

[36] E. Ziętkiewicz, A. Wojda, and M. Witt, "Cytogenetic perspective of ageing and longevity in men and women," *Journal of Applied Genetics*, vol. 50, no. 3, pp. 261–273, 2009.

[37] G. de los Campos, Y. C. Klimentidis, A. I. Vazquez, and D. B. Allison, "Prediction of expected years of life using whole-genome markers," *PLoS ONE*, vol. 7, no. 7, Article ID e40964, 2012.

[38] Z. Pan and C. Chang, "Gender and the regulation of longevity: implications for autoimmunity," *Autoimmunity Reviews*, vol. 11, no. 6-7, pp. A393–A403, 2012.

[39] S. Bennett, X. Song, A. Mitnitski, and K. Rockwood, "A limit to frailty in very old, community-dwelling people: a secondary analysis of the Chinese longitudinal health and longevity study," *Age and Ageing*, vol. 42, no. 3, pp. 372–377, 2013.

[40] L. Sun, C. Y. Hu, X. H. Shi et al., "Trans-ethnical shift of the risk genotype in the CETP I405V with longevity: a Chinese case-control study and meta-analysis," *PLoS ONE*, vol. 8, no. 8, Article ID e72537, 2013.

[41] Y. Liu, M. J. Aryee, L. Padyukov et al., "Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis," *Nature Biotechnology*, vol. 31, no. 2, pp. 142–147, 2013.

[42] J. U. Guo, D. K. Ma, H. Mo et al., "Neuronal activity modifies the DNA methylation landscape in the adult brain," *Nature Neuroscience*, vol. 14, no. 10, pp. 1345–1351, 2011.

[43] S. Falcon and R. Gentleman, "Using GOstats to test gene lists for GO term association," *Bioinformatics*, vol. 23, no. 2, pp. 257–258, 2007.

[44] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, vol. 57, pp. 289–300, 1995.

[45] R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 299–314, 1996.

[46] G. Toperoff, D. Aran, J. D. Kark et al., "Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood," *Human Molecular Genetics*, vol. 21, no. 2, Article ID ddr472, pp. 371–383, 2012.

[47] S. J. Docherty, O. S. Davis, C. M. Haworth et al., "Bisulfite-based epityping on pooled genomic DNA provides an accurate estimate of average group DNA methylation," *Epigenetics & Chromatin*, vol. 2, no. 1, article 3, 2009.

[48] M. Krzywinski, J. Schein, I. Birol et al., "Circos: an information aesthetic for comparative genomics," *Genome Research*, vol. 19, no. 9, pp. 1639–1645, 2009.

[49] F. Gilbert, "Disease genes and chromosomes: disease maps of the human genome. Chromosome 17," *Genetic testing*, vol. 2, no. 4, pp. 357–381, 1998.

[50] Y. Ding, F. He, H. Wen et al., "DNA methylation status of cyp17-II gene correlated with its expression pattern and reproductive endocrinology during ovarian development stages of Japanese flounder (Paralichthys olivaceus)," *Gene*, vol. 527, no. 1, pp. 82–88, 2013.

[51] D. Ben-Avraham, R. H. Muzumdar, and G. Atzmon, "Epigenetic genome-wide association methylation in aging and longevity," *Epigenomics*, vol. 4, no. 5, pp. 503–509, 2012.

[52] J. Wang, S. Zhang, Y. Wang, L. Chen, and X.-S. Zhang, "Disease-aging network reveals significant roles of aging genes in connecting genetic diseases," *PLoS Computational Biology*, vol. 5, no. 9, Article ID e1000521, 2009.

[53] R. Nusse and H. Varmus, "Three decades of Wnts: a personal perspective on how a scientific field developed," *The EMBO Journal*, vol. 31, no. 12, pp. 2670–2684, 2012.

[54] 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.

[55] J. M. Devaney, S. Wang, S. Funda et al., "Identification of novel DNA-methylated genes that correlate withhuman prostate cancer and high-grade prostatic intraepithelialneoplasia," *Prostate Cancer and Prostatic Diseases*, vol. 16, no. 4, pp. 292–300, 2013.

[56] J. C. Yoon, A. Ng, B. H. Kim, A. Bianco, R. J. Xavier, and S. J. Elledge, "Wnt signaling regulates mitochondrial physiology and insulin sensitivity," *Genes & Development*, vol. 24, no. 14, pp. 1507–1518, 2010.

[57] P. M. Cawthon, "Gender differences in osteoporosis and fractures," *Clinical Orthopaedics and Related Research*, vol. 469, no. 7, pp. 1900–1905, 2011.

[58] W.-F. Li, S.-X. Hou, B. Yu, M.-M. Li, C. Férec, and J.-M. Chen, "Genetics of osteoporosis: accelerating pace in gene identification and validation," *Human Genetics*, vol. 127, no. 3, pp. 249–285, 2010.

[59] C. Franceschi, M. Bonafè, S. Valensin et al., "Inflamm-aging. An evolutionary perspective on immunosenescence," *Annals of the New York Academy of Sciences*, vol. 908, pp. 244–254, 2000.

[60] T. A. Christensen and J. G. Hildebrand, "Male-specific, sex pheromone-selective projection neurons in the antennal lobes of the moth Manduca sexta," *Journal of Comparative Physiology A*, vol. 160, no. 5, pp. 553–569, 1987.

[61] C. Franceschi, L. Motta, M. Motta et al., "The extreme longevity: the state of the art in Italy," *Experimental Gerontology*, vol. 43, no. 2, pp. 45–52, 2008.

[62] J. Tower and M. Arbeitman, "The genetics of gender and life span," *Journal of Biology*, vol. 8, no. 4, p. 38, 2009.

[63] K. D. Siegmund, C. M. Connor, M. Campan et al., "DNA methylation in the human cerebral cortex is dynamically regulated throughout the life span and involves differentiated neurons," *PLoS ONE*, vol. 2, no. 9, p. e895, 2007.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 305629, 13 pages
http://dx.doi.org/10.1155/2014/305629

Research Article

Augmenting Multi-Instance Multilabel Learning with Sparse Bayesian Models for Skin Biopsy Image Analysis

Gang Zhang,[1,2] Jian Yin,[1] Xiangyang Su,[3] Yongjing Huang,[4] Yingrong Lao,[4] Zhaohui Liang,[4] Shanxing Ou,[5] and Honglai Zhang[4]

[1] *School of Information Science and Technology, Sun Yat-Sen University, Guangzhou 510275, China*
[2] *School of Automation, Guangdong University of Technology, Guangzhou 510006, China*
[3] *Department of Dermatology and Venerology, The 3rd Affiliated Hospital of Sun Yat-Sen University, Guangzhou 510630, China*
[4] *The 2nd Affiliated Hospital, Guangzhou University of Chinese Medicine, Guangzhou 510405, China*
[5] *Department of Radiology, Guangzhou General Hospital of Guangzhou Military Command, Guangzhou 510010, China*

Correspondence should be addressed to Honglai Zhang; kjkfk@gzucm.edu.cn

Received 18 January 2014; Accepted 3 February 2014; Published 7 April 2014

Academic Editor: Xing-Ming Zhao

Skin biopsy images can reveal causes and severity of many skin diseases, which is a significant complement for skin surface inspection. Automatic annotation of skin biopsy image is an important problem for increasing efficiency and reducing the subjectiveness in diagnosis. However it is challenging particularly when there exists indirect relationship between annotation terms and local regions of a biopsy image, as well as local structures with different textures. In this paper, a novel method based on a recent proposed machine learning model, named multi-instance multilabel (MIML), is proposed to model the potential knowledge and experience of doctors on skin biopsy image annotation. We first show that the problem of skin biopsy image annotation can naturally be expressed as a MIML problem and then propose an image representation method that can capture both region structure and texture features, and a sparse Bayesian MIML algorithm which can produce probabilities indicating the confidence of annotation. The proposed algorithm framework is evaluated on a real clinical dataset containing 12,700 skin biopsy images. The results show that it is effective and prominent.

1. Introduction

Skin diseases are common in our daily life. Most of the skin diseases are not harmful to our health, while some kinds of them would lead to serious problems for our health. For example, malignant melanoma is a highly aggressive skin cancer which looks just like some harmless nevi in some cases. Pemphigus mostly characterized by the development of blisters on the skin is a rare skin disorder that leads to severe infection without effective treatment. Consequently, rapid recognition and correct diagnosis are important to the grave skin diseases as well as neoplasms, bullous dermatoses, sexually transmitted diseases (STD), and so forth. However, it is a great challenge for doctors specializing in dermatology since there are more than 3,000 kinds of diseases in this field, and what is worse is that the number of patients in dermatology is increasing rapidly [1], leading to great burden for doctors to precisely inspect large amount of cases every day.

Generally there are two categories of skin imaging inspection methods. The first is skin surface imaging. A doctor could be confident of making a diagnosis through observation and routine examination on the skin surface in some cases. However, in many other cases, especially in cases of skin cancer, a doctor is not easy to make a diagnosis decision when only skin surface information is available. The second is skin biopsy imaging, which is the imaging of slice of skin tissue under microscope. Skin biopsy images reflect the pathological changes behind skin lesions at a microscopic level. It is widely accepted that histopathology is the gold standard of diagnosing a skin disease [2]. Skin biopsy imaging can provide valuable information of what happens under skin surface. To reach correct annotation or diagnosis, a doctor needs not only professional knowledge and rich experience

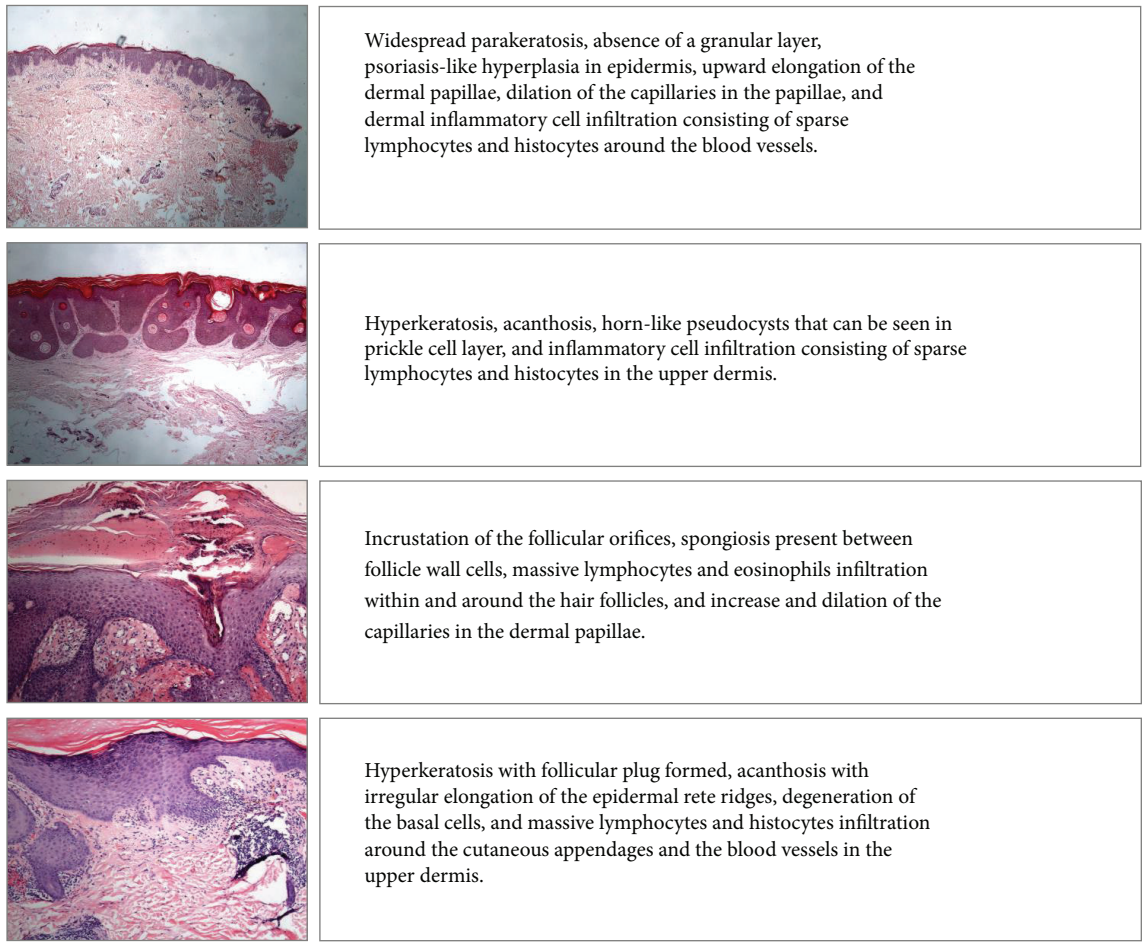


Figure 1: Example of skin biopsy images and their corresponding description in plain text.

in inspecting skin lesions, but also deep understanding of skin histopathological imaging. While analyzing skin biopsy images consumes more time and requires more skills, differentiating normal/lesion regions or similar skin diseases becomes great challenges for doctors. Meanwhile, current skin biopsy image inspection is heavily relied on experience and professional knowledge of histopathological laboratory experts, which are subjective and unstable. To obtain a stable and reproducible diagnosis result, a computer-aid diagnosis (CAD) system is necessary.

Hence it is meaningful to develop computational methods for automatic feature recognition and annotation of skin biopsy images. However, there are some significant challenges due to the complex structures and textures of biopsy images and indirect relationship between historic diagnosis records and images. First of all, in dermatological practice, when annotating biopsy skin images, doctors only give plain text description for a patient attached to several skin biopsy images. The plain text description involves a set of standard dermatological annotation terms and some linked words to show key features reflected by the biopsy images, as shown in Figure 1. However, in fact, the dermatological terms only reflect certain local regions instead of the whole image. See Figure 2 for details. Only one or more small local regions is responsible for a certain dermatological term. However, the correspondence between dermatological terms and local regions is unknown in current datasets. Thus we cannot model this correspondence directly.

Another challenge is that, even for the same term, its corresponding local regions may be significantly varied in size, shape, texture, lightening, inner structure, or the relation between local regions with different terms. In addition, we should be aware of the fact that sublayers of a skin tissue are strictly ordered, leading to some correlations between local visual regions as well as the corresponding features [3]. All these challenges make the task more difficult to tackle compared with traditional machine learning ones.

Several attempts have been reported publicly to build models or classifiers for skin image automatic annotation or recognition. A portion of them have attempted to design different color space-based feature extraction methods and to apply different machine learning models to achieve good performance for different kinds of skin diseases [4–6]. However, a large amount of these methods have to face the problem of manually labeling lesion regions. In order to build a training dataset comprising both normal and lesion skin images, we are required to pick out normal and lesion regions for each skin image. Meanwhile, a large number

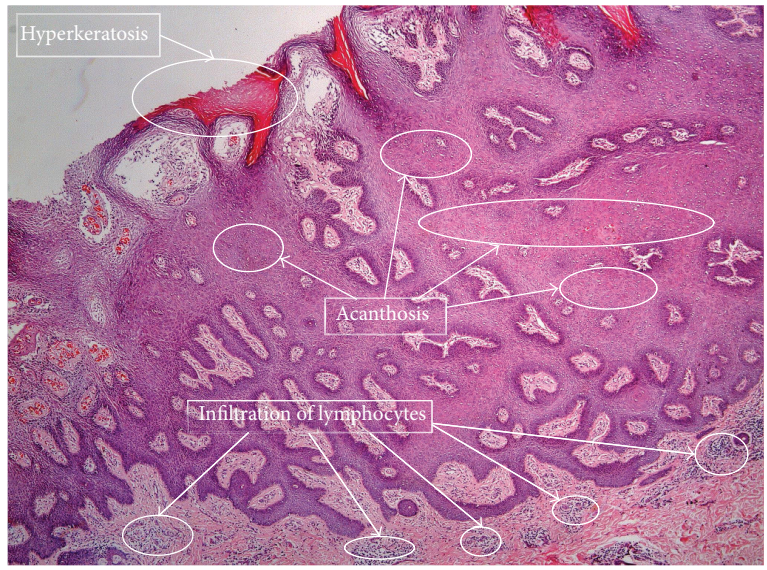


Figure 2: Dermatological terms and their corresponding regions.

of histopathological image analysis methods have also been reported for classification or grading of biopsy images [7–10]. But few of them attempted to model the indirect relationship between histopathological features and parts of a biopsy image.

Moreover, many previous methods required specialized knowledge to choose a proper color space representation and a model, which is not feasible in most cases. Recently, Bunte et al. [11] proposed a machine learning framework to combine several color space representation methods through a weighting procedure. Zhang et al. [12] proposed to convert the skin biopsy image feature recognition problem into a multi-instance (MI) learning problem and then solve it by current well-studied MI algorithms, which is the first attempt to tackle the skin biopsy image annotation problem within machine learning framework. In their paper, they applied a famous graph cutting algorithm, named nnormalized cut [13], to generate visual disjoint regions and then apply image feature extraction algorithm for each local region, so as to turn each image into a MI sample. However, they simply trained an individual MI learner for each target feature to be recognized, discarding the correlation between target features, which is not sufficient from a medical point of view.

In this paper, we attempt to tackle the skin biopsy image feature extraction problem under a recently proposed machine learning framework, multi-instance multi-label (MIML) learning. We first show that the problem is naturally a MIML learning problem. Then we propose a sparse Bayesian MIML learning algorithm with a Gaussian prior as the main model, which is able to model a posterior distribution of the target features giving images as input. We evaluate the proposed algorithm on a real dataset from the department of dermatology and venereology of a large local hospital. The evaluation results show that the proposed algorithm framework can effectively annotate the concerning terms of skin biopsy images superior to existing methods.

Table 1: 15 considered annotation terms and their occurence frequency.

Number	Name	Rate
$T1$	Retraction space	28.65%
$T2$	Papillomatosis	22.71%
$T3$	Follicular plug	1.8%
$T4$	Hypergranulosis	32.15%
$T5$	Horn cyst	4.14%
$T6$	Basal cell liquefaction degeneration	6.48%
$T7$	Thin prickle cell layer	2.61%
$T8$	Infiltration of lymphocytes	9.12%
$T9$	Hyperpigmentation of Basal cell layer	36.99%
$T10$	Nevocytic nests	18.56%
$T11$	Munro microabscess	7.72%
$T12$	Acanthosis	19.05%
$T13$	Absent granular cell layer	23.24%
$T14$	Parakeratosis	6.81%
$T15$	Hyperkeratosis	11.30%

2. Materials and Methods

2.1. Materials. We aim at building a machine learning model for annotating a given skin biopsy image with a set of standard dermatology terms. The skin biopsy images are digitally stored. The size of each image is 2048×1536 pixels with $24k$ colored. The image files are fed to the model that outputs a binary vector to indicate whether the terms are annotated. We consider totally 15 annotation terms which appeared in the electronic records and regarded important for diagnosis in this study. Table 1 lists 15 terms and their occurrence ratios in the whole evaluation dataset.

In our evaluation dataset, each patient has at least one skin biopsy image of the target skin tissue, associated

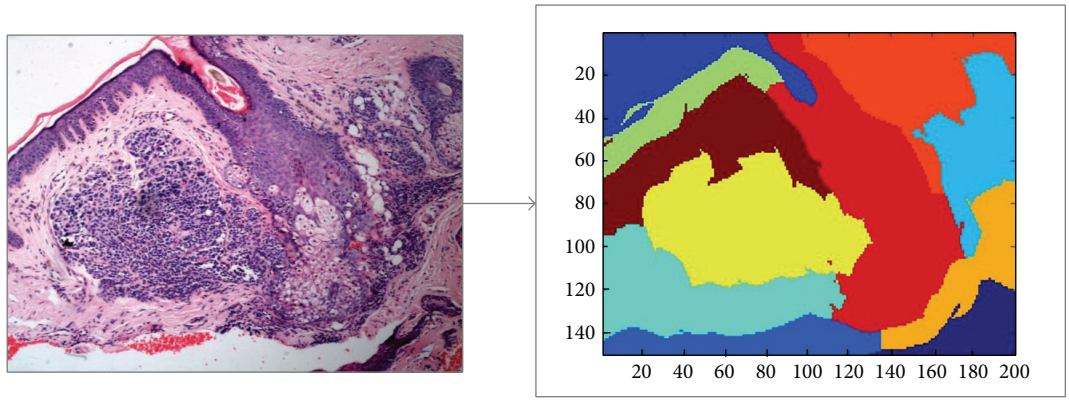


Figure 3: Normalized cut with $k = 11$.

with a plain text description given by an expert. We only select an image for each patient and assume that each selected image contains all terms in the text description. Then we can convert the text description into a binary vector through simple word-matching procedure. Thus the original problem becomes a multilabel binary classification problem.

We further formally define the problem as follows. Let $D = \{(X_1, T_1), \ldots, (X_n, T_n), X_i \in I, T_i \in W\}$ be a set of images associated with the annotated terms, where X_i is an image, $T_i = \{t_1, \ldots, t_{m_i}\}$ is a set of terms associated with the image, and I, W stand for the whole set of images and terms, respectively. The problem is to learn a function $f : I \to W$ with a training image set D such that when given a test image X_t it can give the posterior probability of each term in W to be annotated to X_t.

To represent the key features of a given image, different feature extraction methods have been proposed and developed and in various fields of image understanding research [7]. However, a large body of feature extraction methods previously applied in histopathological image analysis, which extract global features, is not suitable for our biopsy image annotation task. Because in our problem there are m to n relationships between notation terms and local regions within images, methods extracting global features are not able to express local features corresponding to each region of interest.

If a given image can be cut properly to generate meaningful regions, the above correspondence can be directly modeled. The proper cutting of a given image should generate regions attached with terms as few as possible. Such regions are relatively simple and easy to be described. In histopathological image analysis, several image cutting methods have been applied in different tasks. Caicedo et al. [4] proposed a bag-of-words approach for histopathological image annotation. They divided an image into blocks of equal size to generate a codebook for feature representation. Ji et al. [14] and Li et al. [15] applied the almost same block-cutting method to generate MI samples from given images. Another region generating method that should be mentioned is based on block clustering proposed by Chen and Wang [16]. They generated regions by clustering 2D waveform transformation coefficiencies of each block. Thus similar blocks can be gathered into a single cluster. In their work clusters were regarded as regions and it generated discontiguous regions, not regions in common sense.

However, such cutting approaches cannot generate regions of medical meaning as we need. As shown in our previous work [12], the model that is built upon such region generating methods cannot properly capture the direct medical knowledge and experience for annotating biopsy images. An experienced doctor would annotate an image by directly inspecting some local visual disjoint regions within the image. Following this observation, we apply the same idea to cut a given image into k visual disjoint regions through the normalized cut algorithm proposed by Shi and Malik [13]. The number of regions should be set before running the algorithm. Figure 3 shows the result of normalized cut for an skin biopsy image with $k = 11$.

It should be noted that there is not any optimal k for the annotation problem, since the concept of local region is not an actual cutting of an image. A smaller k leads to larger regions, which may contain more than one term, while fragment regions may be generated if k is large. Hence we add a region size constraint when running the cutting algorithm. A generated region should contain at least 1500 pixels to avoid too much fragments, along with a relatively large k. Thus we can get as much as possible regions but avoiding too much fragments.

To further express each generated region as a vectorial representation, we propose a feature representation method that can capture both texture and structure features of regions. The method combined the features extracted through the method introduced in our previous work [8, 12] and features from a graph view of the image. Briefly saying, for the first part of the features, the method performs a waveform transformation for each equal-sized block within each region and combines the waveform transformation coefficiencies to form a 9-ary real vector for each region. To make the paper self-contained, we present some details of the extracted features. The first three features f_1, f_2, f_3 are means of L, U, V values of all pixels within a region. The next three features f_4, f_5, f_6 are mean DWT coefficients HH, HL and LH of all blocks. The last three features are

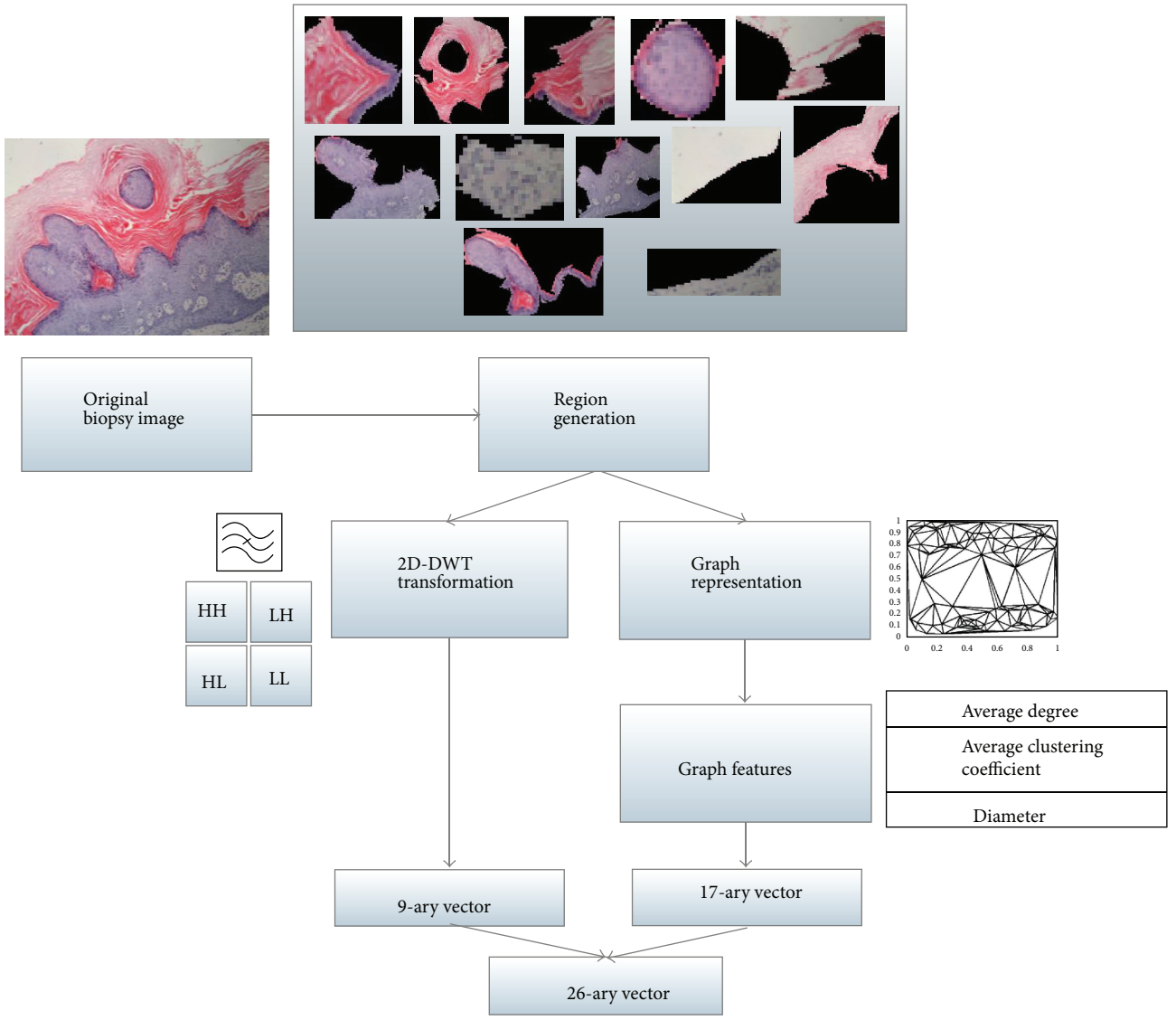


Figure 4: Feature extraction for local regions.

the 1st, 2nd, and 3rd order normalized criteria [17] of the whole region.

For the second part of the features, we represent a region as a graph in which nodes are centroids of clusters of pixels and edges are the relationship between nodes with real weights. We apply a heuristic algorithm [5] to seek the centroids of local similar pixels. Then a Delaunay triangulation method [18] is applied to the set of centroids to add edges. Graph representation methods are widely used in histopathological image analysis for it is able to capture the structure of a tissue [7, 9, 10, 19]. Figure 4 illustrates the main steps of our region feature extraction procedure.

There are three types of graph features considered in our feature representation. The first is average degree of nodes belonging to each cluster in the graph. It can be simply obtained by averaging the degrees of all nodes belonging to the same cluster. The degree of a node is the number of edges. The second is average clustering coefficient (ACC) [20], which measures the average connectivity of a node and its neighbors. The ACC for node i is defined as

$$\mathrm{ACC}_i = \frac{2C_i}{d_i\left(d_i - 1\right)}. \tag{1}$$

In (1), C_i is the number of edges between node i and its neighbors and d_i is the degree of node i. The neighborhood between each pair of nodes is measured by the Euclidean distance. We calculate the values of ACC for nodes belonging to different clusters. We compute the average ACC of all nodes in the graph and nodes in the same cluster. Hence there are $p + 1$ average ACC where p is the number of clusters. The third is the diameter of the graph, which is defined as the shortest path of the longest path between pair of nodes on the graph. In our work $p = 4$, there are 4 average degrees, 4×3 different types of node connection, which results in 12 ACCs, and finally a diameter value of the whole graph. Totally we get a 17-ary feature vector.

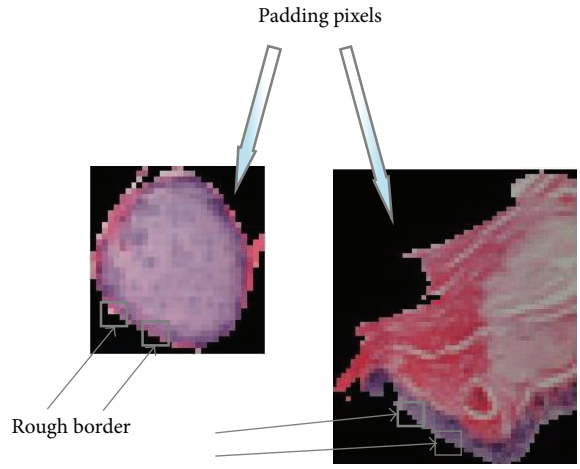


Figure 5: Padding pixels.

Since the generated regions are irregular in shape, padding pixels (in black) must be excluded from our feature extraction procedure. To do this, for the texture features, blocks that have at least one black pixel are discarded. Since the block in our method is of 4×4 pixels, it leads to a rough border of the original region which would not significantly affect the texture features. For graph features, it is not a problem since the black pixels would of course be clustered into a single cluster. Thus we can simply discard such black cluster to get rid of padding pixels. Details of the above idea were presented in our recent work [8]. Figure 5 illustrates the processing of padding pixels in our feature extraction procedure.

Thus, a skin biopsy image is decomposed into a MI example (bag), in which visual disjoint regions are instances. Moreover, we can define a binary vector to indicate whether an annotation term is associated with a given image. An annotation term can be regarded as a label associated with an image. Hence the biopsy image annotation problem can be naturally considered as a multi-instance multilabel (MIML) problem. Based on the relationship between regions and terms from clinical experience, we tackle the problem under the standard MI assumption which was firstly introduced by Dietterich et al. [21], assuming that a sample was labeled positively if at least one instance in it is positive and negative otherwise. The standard MI assumption has been widely used in bioinformatics study [22] and it is also suitable for this work.

2.2. Methods

2.2.1. Sparse Bayesian MIML Learning Framework. In the previous subsection, we have shown that the problem is naturally a MIML problem. Now we propose a novel algorithm to solve this problem effectively. The general idea is that we first randomly construct a set of basic MIML learners and then learn a sparse weights vector under the relevant vector machine (RVM) [23] framework to combine the basic learners together. The learning framework prunes off many learners by automatically driving the corresponding weights to zero so as to get a sparse solution. The motivation of this work is the consideration of time complexity of building a good MIML learner. A weighted ensemble method is adopted, and the weights are determined by RVM method. The method does not require basic learners of good quality. It can find an optimal combination of learners of low quality at relatively low cost.

2.2.2. Generating Basic Learners. We make use of a recently proposed Bayesian MIML learning model [24] for the generation of MIML basic learners. The method directly models a predictive distribution of terms conditioning on training data with a Gaussian process (GP) prior. We introduce a set of unobserved real-value functions $f = \{f_1, \ldots, f_s\}$ ranging from $[0, 1]$, where s is the number of target labels. The value of f for a given instance (region) indicates to which extent it should be annotated with the s concerning terms. Under the standard MI assumption, the bag label can be determined by a max or soft max function over f_i on all instances in the bag [25].

We formally describe the procedure of basic learner construction as follows. The goal is to model the predictive probability of the concerning annotation terms T, giving the training set D, a prior K^{GP}, and a test sample x, which can be expressed as $p(T \mid D, x, K^{\mathrm{GP}})$. The prior K^{GP} can be given by a kernel function through a Gaussian process. The likelihood function associated with latent functions f on D can be expressed as

$$p(T \mid F) = \prod_{i=1}^{s} \prod_{j=1}^{n} p\left(t_i \mid F_{ij}\right), \tag{2}$$

where F_{ij} is the value of applying f_i to all instances in bag x_j and F is a matrix containing all values of applying all f on D.

Since F is unknown, we impose a prior for F to avoid overfitting when evaluating it. Following Bonilla et al.'s work [26], a Gaussian prior for F with zero mean and covariance is defined as follows:

$$p(F) = N\left(F \mid 0, K^{\mathrm{GP}} \otimes K\right). \tag{3}$$

In (3), K stands for the gram matrix for some kernel functions (e.g., RBF or poly kernel) in instance space and K^{GP} in fact indicates the relationship between terms to be annotated. In [24], they adopted a marginal likelihood maximization method to find the optimal K^{GP}, which is expensive. In this work, we do not directly work out the optimal solution for K^{GP}. On the contrary, we randomly generate K^{GP} Q times and then learn a vector of weights to obtain an optimal combination.

With K^{GP}, we can further derive the posterior distribution given a training dataset D as

$$p(F \mid D, T) = \frac{p(T \mid F)\, p(F)}{\int p(T \mid F)\, p(F)\, dF}. \tag{4}$$

Notice that the second $p(T) = \int p(T \mid F) p(F) dF$ is a constant value since T is constant and F is integrated out. Thus it can

be ignored. Because $p(T \mid F)p(F)$ is not a Gaussian [26], we use some approximation methods to evaluate it. Following Nickisch and Rasmussen's work [27], we apply the Laplace approximation to convert $p(T \mid F)$ into a Gaussian near its true mode. According to [26, 27], we can directly write down the mean and variance of the approximation distribution for $p(T \mid F)$. Meanwhile we notice that $p(F)$ is also a Gaussian, which leads to a Gaussian distribution for $p(F \mid D, T)$.

The predictive probability can then be derived from the likelihood, prior, and posterior distribution aforementioned. We have

$$p(t_i \mid D, T, x) = \int \max(F_x)\, p(F_x \mid D, T, x)\, dF_x, \quad (5)$$

where x is a test bag (image) and F_x is a vector of applying all f to all instances in x. The first term on the right-hand side reflects the standard MI assumption, meaning that the largest value among f determines the probability to be annotated with the corresponding term. For computational convenience, we often use soft max function instead of max in (5), given by $\ln \sum_i e^{a_i}$. The predictive distribution is also a Gaussian and can be solved directly as follows:

$$p(t_i = \text{true} \mid D, T, x) = \int \ln\left(\frac{\sum_j F_{xj}}{|F_x|}\right) p(F_x \mid D, T, x)\, dF_x. \quad (6)$$

The right-hand side of (6) is a Gaussian, which can be determined through a EM-like procedure [27]. An important thing should be noticed is that (6) has a parameter matrix K^{GP} that controls the relationship between terms.

The time complexity of the above procedure can be analysed as follows. Suppose we generate a set of Q basic learners and $|T|$ annotation terms. For each learner, there is a random sampling procedure for K^{GP} which requires $O(|T|^2)$ operations; training a MIML learner requires $O(|T| \times |D|^2)$, where $|D|$ denotes the number of instances in training dataset.

2.2.3. Sparse Bayesian Ensemble. Since the cost of calculating the optimal K^{GP} is very high, we randomly set them Q times to obtain a set of different learners and then apply a weighted ensemble procedure as follows:

$$f_{\mathrm{ens}}(x) = \sum_{i=1}^{Q} f_i(x). \quad (7)$$

A RVM-like algorithm [23] is adopted to find the optimal weights to combine them. The main reason for using RVM is twofold. On one hand it is purely based on Bayesian theory which is consistent with our basis learner. On the other hand, RVM can give a sparse solution which is preferred in large data analysis and fast annotation. Figure 6 shows the main steps of the proposed algorithm framework.

The target model is a weighted ensemble of a set of basic learners. To get a sparse representation, we impose an ARD prior [28] on the weights w which is a Gaussian with zero mean and different variances α_i for each weight w_i. In RVM's optimization procedure [23], a large body of variances would be driven to infinity leading the corresponding weights to zero. Hence a large body of weights would be pruned off from the model and final a sparse model is obtained. Formally, let $w = \{w_1, \ldots, w_Q\}$ be a set of weights associated with Q learners. A Gaussian prior with zero mean and different variances is imposed on w. Tipping's work [23] indicated that when applying a maximum a posterior (MAP) learner to learn an optimal w, a large body of w would be driven to zero. Following this idea, we apply RVM algorithm on w given the training dataset D.

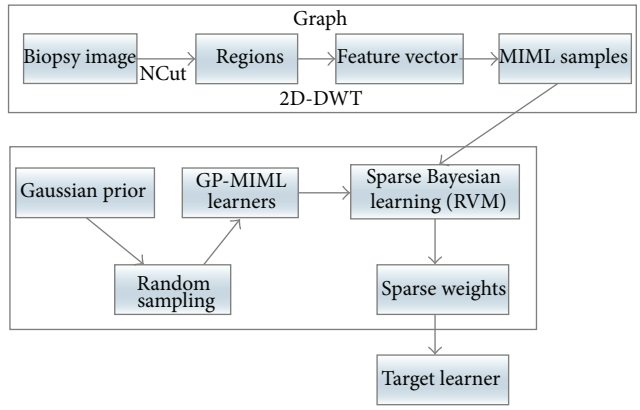


Figure 6: Main steps of the proposed algorithm.

Please note that the weighted ensemble may not follow a Gaussian distribution. This is because $\sum_i w_i$ is not guaranteed to be 1. A normalization procedure should be applied to obtain a normalized combination

$$w_i = \frac{w_i}{\sum_j w_j}. \quad (8)$$

By applying RVM, a smooth learner can be obtained which captures the general features of the whole training dataset. RVM adopts an iterative procedure to find optimal weights.

3. Results and Discussion

3.1. Results. We present the evaluation result of the proposed algorithm on a real dataset gathered from a large local hospital. The setting of basic learner generation is the same as [24] and the setting of RVM follows Tipping's original implementation [23]. The proposed method is compared with some existing approaches in histopathological image analysis. Since some of them are not consistent with the MIML setting in our work, we would implement them on a more general foundation for image analysis.

3.1.1. Dataset and Data Preprocessing. The evaluation was carried out on a real skin disease clinical dataset from a large local hospital. The dataset has been reconstructed to get rid of irregular patient information and low quality biopsy images.

The biopsy images in the evaluation dataset are taken by a Leica DFC290 digital camera with 20x, 40x, and 100x

microscope objective lenses. The images are taken in RGB color space and stored in JPEG format. For convenience, we only keep images at 40x magnification ratio. It contains 4,123 patients with 12,700 images. The images are 2048 × 1536 pixels with 24k colors. For computational efficacy, they are rescaled to 800×600 pixels. There are three 40x biopsy images for each patient on average. We consider 15 features to be annotated, corresponding to 15 standard terms, as shown in Table 1, and then convert the plain-text description into a 15-ary binary vector in which each element indicates whether the corresponding term exists in the diagnosis record in plain text, as shown in Figure 1. Since most doctors use standard terms and link words in their description, training dataset of good quality can be obtained in this way.

Each image associated with a patient is converted into a bag through normalized cut and then a feature extraction method combined with waveform transformation and graph representation. For normalized cut, the number of regions k must be set manually. In our evaluation we set $k = 11$ which means an image would be converted into a bag consisting of 11 instances. A further discussion on the setting strategy of k is presented in the next section. Different images of the same patient are associated with the 15-ary binary of the patient. We denote the dataset generated through the above procedure as $D1$. For waveform transformation, each region should be divided into blocks of size 4 × 4 pixels. Blocks containing at least one black pixel would be discarded. For graph representation, the number of clusters p is set to 5, assuming that there are 5 different tissues in each image on average. In node identification algorithm, circles containing less than 20 pixels would not be taken into account.

Since there are other compared methods that are not consistent with the MI setting, we generate another three data representations, namely, $D2$, $D3$, and $D4$, for these methods. Data representation $D2$ is based on the equal-sized block cutting method proposed in [4]. We first cut each image into 4 × 4 blocks and apply a scale-invariant feature transform (SIFT) descriptor [29] to extract features and use histogram to express it as a feature vector. $D2$ is a bag-of-words [14] image representation which is widely used in image understanding. Dataset $D3$ is an equal-sized block MI sample representation proposed in [15]. The main procedure is similar to $D2$, but it directly regards each block with SIFT representation as an instance. Hence in $D3$ there are totally 30,000 instances in each bag. Finally dataset $D4$ is a clustering based representation. It clusters equal-sized blocks represented in real value vector and regards each cluster as an instance. Details of this method can be found in [16]. Table 2 lists the above data representation and their consistent methods for comparison.

Note that these datasets are only different in their preprocessing steps. In Table 2 we can see that method $M4$ can be fed with $D1$ and $D3$, for $M4$ is a MIML learning algorithm naturally consistent with MI data representation. However $D3$ cannot be fed to $M1$ because the idea of $M1$ is to regard each visual disjoint region instead of equal-sized block, as an instance. Different definitions of instance are originated from the difference of underlying idea of the problem. A single block may not contain medically acceptable features, which is not consistent with our MI framework.

Table 2: Data representation and their consistent methods.

Method	Reference	Dataset
$M1$: our method	This work	$D1$
$M2$: MIBiopsy	Zhang et al. [12]	$D1$
$M3$: bag of features	Caicedo et al. [4]	$D2$
$M4$: MIMLSVM	Li et al. [15]	$D1$, $D3$
$M5$: DDSVM	Chen and Wang [16]	$D4$

Table 3: Evaluation criteria for multilabel learning.

Name	Equation
hloss	Evaluate the number of misclassified label pairs
one-error	Evaluate the portion that a label of highest probability is not a correct label
coverage	Evaluate the average distance to go down to find the proper label for a given image
rloss	Evaluate the average fraction of label pair that are misordered in the ranking list

3.1.2. Evaluation Criteria. We adopt five different criteria to evaluate the performance of the proposed method and the compared methods. The first is accuracy, a zero-one loss function evaluating whether a single term is correctly annotated. It can be applied to evaluate the performance of methods that annotate only one term each time. Since the proposed method is a MIML one, it can be regarded as a multilabel learner. Several evaluation criteria have been proposed in multilabel learning and MIML learning study [30]. Introducing such criteria is necessary for our evaluation. Formal definition of the four multilabel evaluation criteria can be found in [30]. Table 3 lists five criteria used in our evaluation.

3.1.3. Evaluation Result. For the methods shown in Table 2, we use the same setting for evaluation. The evaluation is launched through a supervised learning manner. The whole dataset (with 12,700 images) is divided into training set and test set at a ratio 3 : 7. To avoid learning bias, the occurrence ratios of the concerning terms in Table 1 were kept the same as the training set. For method $M1$, we use a modified GPMIL and RVM implementation which were originally proposed by Kim et al. [31] and Tipping [23].

The first evaluation focuses on the annotation accuracy. Recall that we have 15 concerning annotation terms. Table 4 gives the results of annotating each term by different methods in Table 2.

It should be noted that the output of method $M1$ is a 15-ary real vector indicating the probabilities of annotating 15 terms. In this part of evaluation, we simply use an indicator function which outputs 1 if the probability is not less than 0.5 and 0 otherwise. Figure 7 shows some outputs of $M1$ and $M5$, in which the probabilities of the concerning terms are shown, as well as the groundtruth annotation terms.

Table 4: Annotation result evaluated by accuracy.

Term	$M1$	$M2$	$M3$	$M4$	$M5$
$T1$	**78.2%**	76.1%	70.6%	75.9%	68.3%
$T2$	**80.3%**	75.9%	76.1%	74.5%	73.8%
$T3$	77.7%	**79.5%**	77.8%	76.2%	68.5%
$T4$	81.3%	81.2%	80.5%	**82.4%**	81.2%
$T5$	69.3%	66.5%	67.9%	**70.1%**	67.4%
$T6$	**76.3%**	75.0%	71.7%	74.2%	72.3%
$T7$	**77.8%**	77.4%	76.5%	75.8%	75.9%
$T8$	85.1%	**85.2%**	84.6%	83.8%	80.9%
$T9$	**87.3%**	86.8%	81.4%	83.0%	78.2%
$T10$	**75.9%**	75.4%	74.5%	73.8%	72.0%
$T11$	69.9%	**71.5%**	68.9%	70.7%	69.6%
$T12$	**78.0%**	76.1%	73.2%	75.8%	73.2%
$T13$	79.2%	**80.1%**	77.2%	78.8%	72.5%
$T14$	80.6%	81.2	77.2%	**81.9%**	73.5%
$T15$	**87.9%**	86.4%	82.6%	83.1%	80.2%

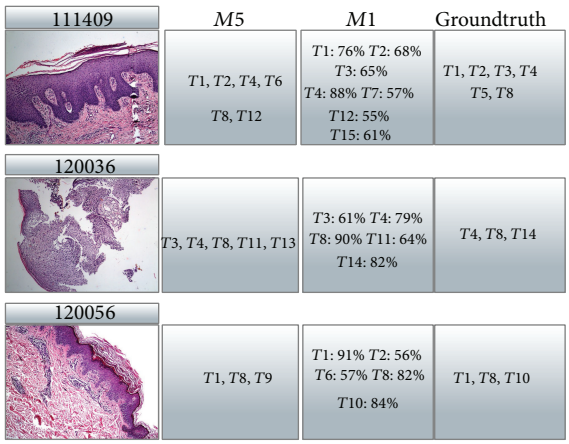


Figure 7: Sample outputs of methods $M1$ and $M5$.

For each row in Table 4, the best accuracy is highlighted. It can be seen that $M1$ achieved the best performance in annotating most terms, which shows the effectiveness of our method. However, for some terms, for example $T4$ and $T5$, method $M4$ performed better than $M1$ and $M2$. We think this is because our graph cutting representation is not consistent with these terms, while the more general grid cutting representation is better.

The second evaluation focuses on the performance of annotating several terms simultaneously. Note that, in previous part of evaluation, accuracy of annotation was evaluated term by term; hence the overall accuracy of annotating all concerning terms may not be as high as the individual ones. We adopt four criteria listed in Table 3 to show the performance of annotation of all terms at the same time. Some criteria rely on the ranking of terms. We can get a natural ranking for the proposed method since it gives the probabilities for all terms. For other methods to be compared in our evaluation, we use the ranking strategy similar to [30]. Note that methods $M2$, $M3$, and $M5$ are not multilabel classifiers. Hence we only compare $M1$, $M4$ with $D1$ and $M4$ with $D3$. Figure 8 shows the performance evaluated by the above four criteria.

According to Table 3, the smaller results of the four criteria indicate the better performance. From Figure 8, it can be seen that method $M1$ achieved best performance compared to other methods in a multilabel classification setting at different training data ratios. For method $M4$, different data representations $D1$ and $D3$ lead to different performances. It can be seen that $D1$ is better than $D3$ in most cases. Since the intuition of $D1$ and $D3$ is totally different, it may be concluded that the representation $D1$ is more consistent with the term set and the models.

Finally we evaluate the sparsity of the proposed model. We vary the ratios between training data and test data and plot them with the nonzero-weighted basic learners after RVM procedure. In this case the set of basic learners contains 200 learners; that is, $Q = 200$. Figure 9 shows the result.

From Figure 9 we can see that RVM procedure can prune off about 2/3 learners, which yields a sparse ensemble learner. Figure 10 shows the corresponding annotation accuracy of different training set sizes. It can be seen that large training set would lead to high accuracy. Figure 9 indicates that the number of nonzero-weighted learners is stable at different training set sizes. The performance of the proposed method obeys the basic principle of machine learning; that is, more training data means model of high accuracy. For illustration, Figure 10 shows the relationship between accuracy and the size of training set for terms $T1$, $T6$, and $T9$.

3.2. Discussions. Some important issues are worth addressing here. First, we must answer why MIML rather than MI framework is consistent with our task. MIML learning problem can be decomposed into several MI learning problems if we assume labels are independent of each other. When coming to our annotation problem, it is observed that there are correlations between annotation terms, including the cooccurence of some terms or the absence of other terms. Furthermore, some annotation terms may appear at the same time for some diseases. To capture the correlations mentioned above, MI learning framework which regards each annotation term independently is not sufficient. However, MIML learning framework is able to capture the relationship between annotation terms, as well as regions, which is superior to MI framework.

Second, our proposed regions generating method is based on normalized cut, which generates visual disjoint regions for a given image. The number of regions generated by normalized cut must be manually set. A small k would lead to large regions that may contain different terms. A large k would lead to fragment regions associated with the same term, as shown in Figure 11. However, in either case, MIML learning framework works according to the standard MI assumption [21, 32]. The former case is equivalent to an instance corresponding to more than one term. The latter case is equivalent to several instances corresponding to the

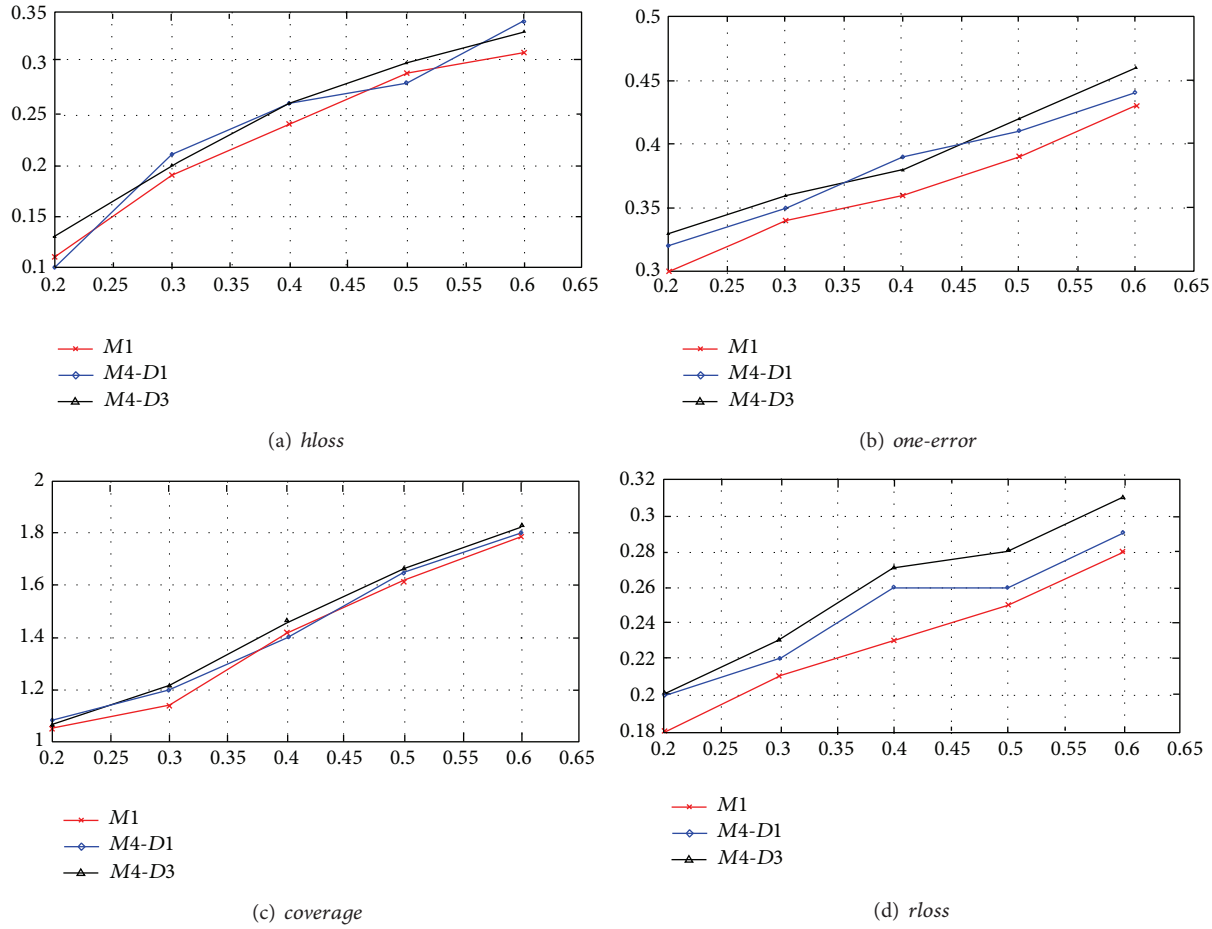


Figure 8: Evaluation result of four criteria.

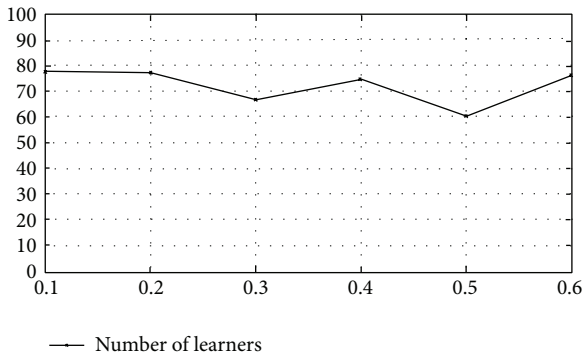


Figure 9: Sparsity and number of basic learners.

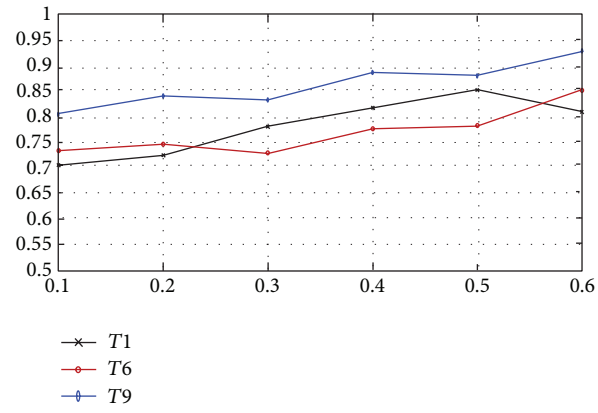


Figure 10: Accuracy and the size of training set.

same term. Though the quantity of k would not affect the effectiveness of MIML, too small k would affect the effect of feature extraction. A region contains different terms cannot be expressed as a real feature vector distinguishing between each term at the same time. Hence, in our work, we use a relative large k according to medical experience to avoid a region containing more than one term and too much fragments.

Third, a Bayesian model can generate probability for each concerning annotation term, which makes it available to build a more powerful model for automated skin disease diagnosis. Annotation terms can be regarded as latent variables between skin biopsy images and diseases, meaning that $p(w \mid I) = \sum_{t \in T} p(w \mid t) p(t \mid I)$ for independent and identically distributed (i.i.d.) terms, where I, t, T, respectively, stand

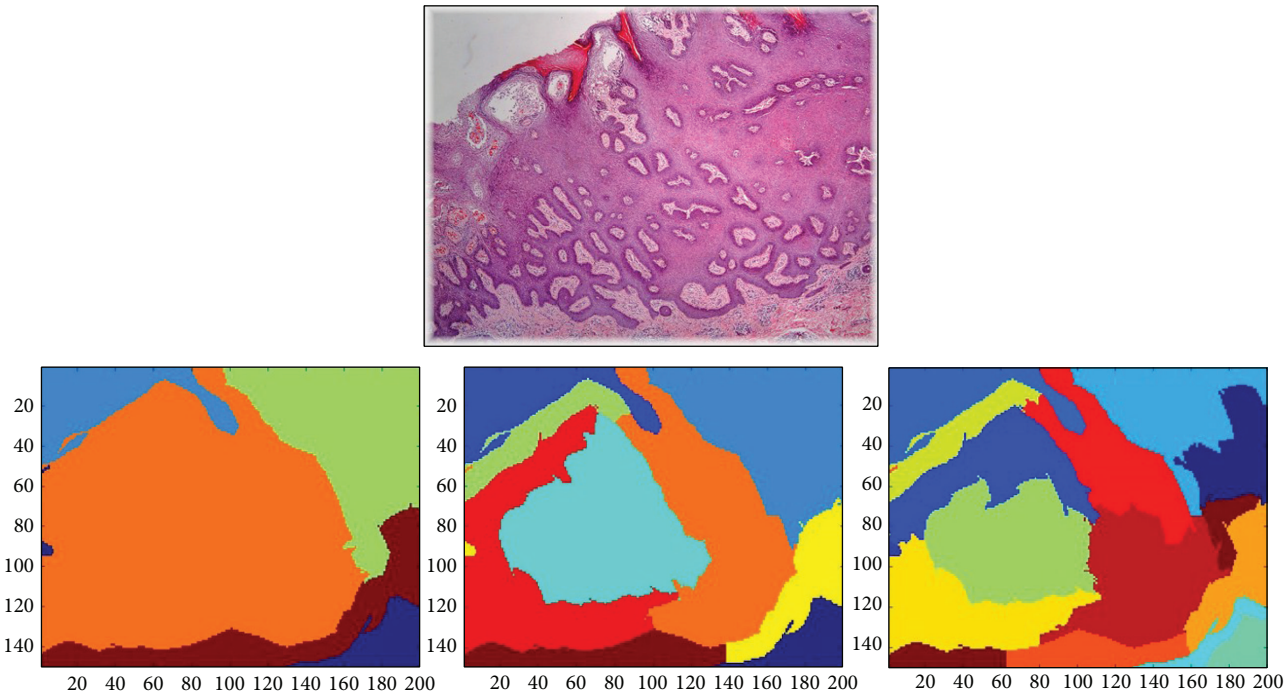


Figure 11: The result of normalized cut with different settings of k.

for diseases, images, a certain term, and the set of terms. And for non-i.i.d. terms, we can separate the terms into dependent term groups and apply almost the same equation as in the i.i.d. case. The method proposed in this paper can effectively evaluate $p(t \mid I)$, and $p(w \mid t)$ can be obtained directly from clinical experience. Hence, it is meaningful in CAD system design and implementation.

Finally, we discuss the multi-instance assumption implied in this work. We use the standard MI assumption [21] when considering the relationship between regions and terms. The standard MI assumption does not directly consider the impact of the number of regions and the relationship between regions to the terms. From clinical observation, most annotation terms can mainly be determined by a single region if the generated regions are not too small. Large region may contain more than one term, but it is also consistent with the standard MI assumption and this can be solved due to the power of MIML models. Though our proposed MIML model in fact considers such relationship, a simple assumption of the problem may lead to simple model.

4. Conclusions

In this paper we propose a MIML framework for skin biopsy image annotation. We adopt a famous graph cutting algorithm named normalized cut to transfer a biopsy image into a MI sample, in which each region is regarded as an instance. To effectively express features of biopsy images, each region is expressed as a 9-ary real vector. To reduce the model complexity and training time, we propose a novel sparse Bayesian MIML learning model, which applies a RVM-like algorithm to obtain a sparse weighted combination for a set of basic learners. We also make use of the well-studied Bayesian MIML learner as basic learners. Evaluation of a real clinical dataset shows that the proposed model can achieve good performance and reach a medical acceptable result. We have achieved an annotation accuracy up to 85% in our evaluation dataset.

The proposed annotation framework directly models doctor's experience of annotation biopsy images. Different from previous work, it is explicable since it can give the correspondence between local visual disjoint regions and the terms associated with them. Future work will focus on studying the relationship between biopsy images and the final diagnosis given the annotation term set as latent variables. And the feature fusion algorithm towards an effective feature representation is another research direction.

Conflict of Interests

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

Authors' Contribution

Gang Zhang and Yongjing Huang contributed equally to this work. This paper is an extended version based on "A sparse Bayesian multi-instance multilabel model for skin biopsy image analysis, " by Gang Zhang, Xiangyang Su, Yongjing Huang, Yingrong Lao, Zhaohui Liang, Shanxing Ou, and Jimmy Huang which appeared in Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference.

Acknowledgments

The authors would like to thank Yunting Liang for her professional advice for this paper. This work is supported by the National Natural Science Foundation of China (nos. 61273249, 81373883, 81274003, 61033010 and 61272065), Science and Technology Project of Guangdong Province (no. 2011B080701036), Natural Science Foundation of Guangdong Province (nos. S2011020001182, S2012010009311), Research Foundation of Science and Technology Plan Project in Guangdong Province (nos. 2011B040200007, 2012A010701013), Zhaoyang Personnel Training Plan of Guangdong Provincial Hospital of Chinese Medicine (no. 2013KT1067), Research Grant of Guangdong Medical Foundation (no. A2012215), Research Grant of Guangdong Administration of Chinese Medicine (no. 2010144), and the Open Foundation of the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University (ESSCKF201401).

References

[1] K. Sellheyer and W. F. Bergfeld, "A retrospective biopsy study of the clinical diagnostic accuracy of common skin diseases by different specialties compared with dermatology," *Journal of the American Academy of Dermatology*, vol. 52, no. 5, pp. 823–830, 2005.

[2] A. Fogelberg, M. Ioffreda, and K. F. Helm, "The utility of digital clinical photographs in dermatopathology," *Journal of Cutaneous Medicine and Surgery*, vol. 8, no. 2, pp. 116–121, 2004.

[3] D. C. Fernandez, R. Bhargava, S. M. Hewitt, and I. W. Levin, "Infrared spectroscopic imaging for histopathologic recognition," *Nature Biotechnology*, vol. 23, no. 4, pp. 469–474, 2005.

[4] J. C. Caicedo, A. Cruz-Roa, and F. A. González, "Histopathology image classification using bag of features and kernel functions," in *Proceedings of the 12th Conference on Artificial Intelligence in Medicine (AIME '09)*, C. Combi, Y. Shahar, and A. Abu-Hanna, Eds., vol. 5651 of *Lecture Notes in Computer Science*, pp. 126–135, Verona, Italy, 2009.

[5] A. B. Tosun, M. Kandemir, C. Sokmensuer, and C. Gunduz-Demir, "Object-oriented texture analysis for the unsupervised segmentation of biopsy images for cancer detection," *Pattern Recognition*, vol. 42, no. 6, pp. 1104–1112, 2009.

[6] O. Sertel, J. Kong, U. V. Catalyurek, G. Lozanski, J. H. Saltz, and M. N. Gurcan, "Histopathological image analysis using model-based intermediate representations and color texture: follicular lymphoma grading," *Journal of Signal Processing Systems*, vol. 55, no. 1–3, pp. 169–183, 2009.

[7] E. Ozdemir and C. Gunduz-Demir, "A hybrid classification model for digital pathology using structural and statistical pattern recognition," *IEEE Transactions on Medical Imaging*, vol. 32, no. 2, pp. 474–483, 2013.

[8] G. Zhang, J. Yin, Z. Li, X. Su, G. Li, and H. Zhang, "Automated skin biopsy histopathological image annotation using multi-instance representation and learning," *BMC Medical Genomics*, vol. 6, supplement 3, article S10, pp. 1–14, 2013.

[9] D. Altunbay, C. Cigir, C. Sokmensuer, and C. Gunduz-Demir, "Color graphs for automated cancer diagnosis and grading," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 665–674, 2010.

[10] C. Gunduz-Demir, M. Kandemir, A. B. Tosun, and C. Sokmensuer, "Automatic segmentation of colon glands using object-graphs," *Medical Image Analysis*, vol. 14, no. 1, pp. 1–12, 2010.

[11] K. Bunte, M. Biehl, M. F. Jonkman, and N. Petkov, "Learning effective color features for content based image retrieval in dermatology," *Pattern Recognition*, vol. 44, no. 9, pp. 1892–1902, 2011.

[12] G. Zhang, X. Shu, Z. Liang, Y. Liang, S. Chen, and J. Yin, "Multi-instance learning for skin biopsy image features recognition," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '12)*, pp. 1–6, Philadelphia, Pa, USA.

[13] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[14] S. Ji, Y.-X. Li, Z.-H. Zhou, S. Kumar, and J. Ye, "A bag-of-words approach for *Drosophila* gene expression pattern annotation," *BMC Bioinformatics*, vol. 10, article 119, 2009.

[15] Y.-X. Li, S. Ji, S. Kumar, J. Ye, and Z.-H. Zhou, "*Drosophila* gene expression pattern annotation through multi-instance multi-label learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 1, pp. 98–112, 2012.

[16] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *The Journal of Machine Learning Research*, vol. 5, pp. 913–939, 2004.

[17] A. Gersho, "Asymptotically optimal block quantization," *IEEE Transactions on Information Theory*, vol. 25, no. 4, pp. 373–380, 1979.

[18] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Transactions on Mathematical Software*, vol. 22, no. 4, pp. 469–483, 1996.

[19] C. Demir, S. H. Gultekin, and B. Yener, "Augmented cell-graphs for automated cancer diagnosis," *Bioinformatics*, vol. 21, supplement 2, pp. ii7–ii12, 2005.

[20] S. N. Dorogovtsev and J. F. F. Mendes, "Evolution of networks," *Advances in Physics*, vol. 51, no. 4, pp. 1079–1187, 2002.

[21] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[22] X. Wang and G. Z. Li, "Multilabel learning via random label selection for protein subcellular multilocations prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 2, pp. 436–446, 2013.

[23] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, no. 3, pp. 211–244, 2001.

[24] J. He, H. Gu, and Z. Wang, "Bayesian multi-instance multi-label learning using Gaussian process prior," *Machine Learning*, vol. 88, no. 1-2, pp. 273–295, 2012.

[25] M.-L. Zhang, "Generalized multi-instance learning: problems, algorithms and data sets," in *Proceedings of the WRI Global Congress on Intelligent Systems (GCIS '09)*, vol. 3, pp. 539–543, Xiamen, China, May 2009.

[26] E. V. Bonilla, K. M. Chai, and C. K. I. Williams, "Multi-task Gaussian process prediction," in *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., MIT Press, Cambridge, Mass, USA, 2008.

[27] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, Mass, USA, 2006.

[28] R. M. Neal, *Bayesian Learning for Neural Networks*, Springer, Secaucus, NJ, USA, 1996.

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

[30] R. E. Schapire and Y. Singer, "BoosTexter: a boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2-3, pp. 135–168, 2000.

[31] M. Kim and F. de la Torre, "Gaussian processes multiple instance learning," in *Proceedings of the 27th International Conference on Machine Learning (ICML '10)*, J. Fürnkranz and J. T. Omnipress, Eds., pp. 535–542, Haifa, Israel, June 2010.

[32] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *Knowledge Engineering Review*, vol. 25, no. 1, pp. 1–25, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 678971, 11 pages
http://dx.doi.org/10.1155/2014/678971

Research Article

Identification of Simple Sequence Repeat Biomarkers through Cross-Species Comparison in a Tag Cloud Representation

Jhen-Li Huang,[1] Hao-Teng Chang,[2,3] Ronshan Cheng,[4] Hui-Huang Hsu,[5] and Tun-Wen Pai[1]

[1] *Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung 20224, Taiwan*
[2] *Graduate Institute of Basic Medical Science, China Medical University, Taichung City 40402, Taiwan*
[3] *Department of Computer Science and Information Engineering, Asia University, Taichung City 41354, Taiwan*
[4] *Department of Aquaculture, National Taiwan Ocean University, Keelung 20224, Taiwan*
[5] *Department of Computer Science and Information Engineering, Tamkang University, New Taipei City 25137, Taiwan*

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

Received 22 November 2013; Revised 27 February 2014; Accepted 27 February 2014; Published 31 March 2014

Academic Editor: Jose C. Nacher

Simple sequence repeats (SSRs) are not only applied as genetic markers in evolutionary studies but they also play an important role in gene regulatory activities. Efficient identification of conserved and exclusive SSRs through cross-species comparison is helpful for understanding the evolutionary mechanisms and associations between specific gene groups and SSR motifs. In this paper, we developed an online cross-species comparative system and integrated it with a tag cloud visualization technique for identifying potential SSR biomarkers within fourteen frequently used model species. Ultraconserved or exclusive SSRs among cross-species orthologous genes could be effectively retrieved and displayed through a friendly interface design. Four different types of testing cases were applied to demonstrate and verify the retrieved SSR biomarker candidates. Through statistical analysis and enhanced tag cloud representation on defined functional related genes and cross-species clusters, the proposed system can correctly represent the patterns, loci, colors, and sizes of identified SSRs in accordance with gene functions, pattern qualities, and conserved characteristics among species.

1. Introduction

Simple sequence repeats (SSRs) are nonrandom distributed nucleotides in genomes of different organisms with repeated basic patterns of lengths from mononucleotide to hexanucleotide [1]. SSRs have been demonstrated as important motifs involved within various biological events including evolutionary processes, gene expression, genetic disease, chromatin organization, and DNA metabolic processes [2–4]. For example, dysplasia disease is a genetic disorder of abnormal cellular development due to imperfect polyalanine expansions (GCC repeats) on *RUNX2* (*CBFA1*) [5]. Another example of Huntington's disease (HD) was found as an irregular distribution of polyglutamine expansions (CAG repeats) located within the coding regions of Huntingtin (HTT) gene, and the excessive repeat number caused the symptoms of genetic neurological disease which appeared at an earlier stage [6]. In addition to illustrate the effects of mutations and expansions of SSR repeats on diseases, another example to demonstrate the function of SSR motifs is the insulin-like growth factor 1 (*IGF1*) which was confirmed as one of the growth control genes. The *IGF1* gene contains "AC" repeats located within the upstream regions and is a major determinant of small body size in dogs [7–9]. From previous reports, evidences show that SSR regulation relies on pattern of repeat unit, repeat length, and genetic location in the target genes [2]. These features are fundamental parameters for identifying functional SSRs under various biological applications. However, due to abundant amount of SSRs distributed within genome sequences, it is yet challenging to select significant SSR biomarkers or gene regulation related SSRs automatically from limited information. Therefore, identifying highly conserved SSRs through cross-species comparison may provide an alternative approach to recognize significant biomarkers or discover putative gene regulatory SSR motifs from enormous gene candidates under the assumption of

natural long-term evolutionary processes. On the other hand, discovering exclusive SSR motifs among different species clusters could be applied as species-specific genetic markers or provide unique genetic functions which were developed after species differentiation events. The comparison of SSR motifs across different species clusters may provide important clues and evidences to further understand evolutionary development.

To efficiently identify SSR biomarkers from large amount of genes in different species, considering a few interested genes at a time provides an intuitive and effective approach. One possible approach of selecting interested gene groups from gene ontology (GO) terms was employed in this study. The GO is a set of structured vocabularies defined by Gene Ontology Consortium [10], which is aimed to provide a universal standard of functional annotation for gene products. All GO terms are connected with each other by directed acyclic graphs with hierarchy relationship. Each term belongs to one of the three independent ontologies: biological process (BP), molecular function (MF), and cellular component (CC) and represent different aspects of gene in temporal, functional, and spatial domains, respectively. In this study, the query biological keywords associated with corresponding GO terms could provide a set of functional-associated gene set for SSR biomarker analysis. Recently, several gene sequence studies associated with GO analysis have been reported, such as the Gene Ontology SSR Hierarchy (GOSH) system which adopts GO terms to reveal prominent orthologous SSR patterns [11], FatiGO which is a web tool for finding significant associations of Gene Ontology terms with groups of genes [2], and Goblet system which performs automatic GO term annotation on anonymous sequences [12].

To enhance the ranking and readability of identified SSR motifs, a tag cloud technique was adopted to display the comparative results of cross-species SSRs. Tag cloud representation is a widespread visualization technology which provides users with an informative image from a designated set of data. Tags of different phases or short sentences represent key information of each entry in the dataset. Multiple tags for various data entries could be displayed in an image simultaneously, and which are manually assigned by users or automatically generated by computer algorithms. Each tag cloud could be shown with different visual attributes such as different sizes or colors. In tradition, different sizes of tags are designed to indicate various levels of representativeness of tags within the dataset [12]. Currently, tag clouds have been widely used in several different types of websites including photo albums, bookmarks, and blogs. It is also used in some tag-based biomedical datasets to help users to rapidly understand the representative information from a complex dataset. For example, the iHOPerator system employed tag cloud technique with related functions for genes analysis [13], INTERFEROME applied tag cloud visualization on gene ontology databases for interferon regulated genes [14], and REVIGO used tag cloud approach to summarize and visualize long lists of gene ontology terms [15]. All these examples have shown that tag cloud visualization techniques could be applied to strengthen key information from complex biological datasets.

In this study, we have collected complete genome sequences of 14 model species as the fundamental dataset. All SSR motifs in each gene were extracted and saved in the designed database in advance. Users can prepare a set of genes or assign keywords to defined query genes and then choose model species of interest for cross-species cluster comparison. Model species of interest could be manually clustered or automatically categorized into two groups of mammal and marine species clusters. SSR retrieval and distribution analysis for single species is also available from the developed system. Once all parameters have been settled, the system will perform online comparison and display all grouped SSR motifs in a tag cloud visualization approach. All significantly conserved or exclusive SSR motifs located within the specified gene sets from two species clusters will be efficiently identified and displayed. In addition, all retrieved SSR biomarker candidates will be shown in a tag cloud representation with occurrence frequency, conserved ratio, gene annotation, sequence contents, and corresponding translated proteins through a fast, responsive, and user-friendly web page design.

2. Materials and Methods

2.1. System Configuration. In this study, there are fourteen initially selected genomes obtained from Ensembl database [16], and all collected gene sequences with their corresponding gene coordinates and annotations were downloaded for cross-species comparison in next modules. Each gene sequence including upstream and downstream regions was scanned and all perfect/imperfect SSR patterns under different parameter settings were extracted from collected genes and saved in a newly created SSR database. According to gene coordinate information, the developed system determined the corresponding genetic regions for each SSR motif and all related annotations were saved in the same entry. Accordingly, the analytical module utilized cross-species comparison techniques between two assigned species clusters, and all statistically conserved and/or exclusive SSR patterns could be shown under a tag cloud representation technique. All details are introduced in the next two sections.

2.2. Genome Sequences. To obtain genome sequences of various organisms, the developed system employed Ensembl release 65 as the major data resource. Ensembl database provides complete genome information on multiple eukaryotic model organisms including whole genome sequence, gene annotations, and molecular functions. To lay emphasis on identification of consensus and unique features of SSRs among different species, two species clusters including fishery and mammal species were initially selected for comparison. Since there were only 6 fishery species that could be collected from Ensembl release 65, we therefore selected another 6 popular mammal species for equivalent status. Besides, two famous research organisms in experimental studies were also included in our database. These intentionally selected

model species are zebrafish (*Danio rerio*), stickleback (*Gasterosteus aculeatus*), medaka (*Oryzias latipes*), fugu (*Takifugu rubripes*), tetraodon (*Tetraodon nigroviridis*), and cod (*Gadus morhua*) as fishery species; human (*Homo sapiens*), gorilla (*Gorilla gorilla*), macaque (*Macaca mulatta*), mouse (*Mus musculus*), cow (*Bos taurus*), and dog (*Canis familiaris*) as mammal species; roundworm (*Caenorhabditis elegans*) and fruit fly (*Drosophila melanogaster*) as extra two popular experimental species. These downloaded data include sequence contents, coordinates of exons/introns and UTRs for each gene, and upstream and downstream regions with a length of 2,000 nucleotides.

2.3. SSR Motif Database Construction. To accelerate searching speed in identifying all perfect and imperfect orthologous SSRs from a set of specified genes among fourteen species, we performed an autocorrelation based SSR discovery algorithm and constructed the SSR motif database in advance [17]. The autocorrelation algorithm could extract all candidate perfect/imperfect SSR motifs under different threshold parameters through an efficient and effectively approach. In this study, we only considered SSR motifs with nucleotide length longer than 20 nucleotides and the length of fundamental repeat unit ranging from 1 to 6 nucleotides. The SSR motif-searching algorithm also applied a proportional quality factor for defining SSR patterns of different degrees of noise. In this study, three different tolerant settings were tentatively applied for considering noisy patterns within multiscale SSR tag clouds in later presentation. The tolerant parameters were initially set as 0, 0.1, and 0.2 for representing 0, 10, and 20 percent of noisy contents within an SSR motif. The percentage of noise is defined as the ratio of the nonrepeated nucleotides within a total length of an identified SSR, which includes noise types of insertion, deletion, and substitution mutations. In other words, the zero percent noisy rate represents a perfect repeat segment without any tolerance. The formula for the tolerant percentage is shown in the following equation:

$$\text{Tolerant } (\%) = \frac{\text{nonrepeated nucleotides}}{\text{identified SSR length}} \times 100\%. \quad (1)$$

An SSR motif could locate in six different genetic regions of a specified gene including coding, intron, $5'$ UTR (untranslated region), $3'$ UTR, and upstream and downstream regions. In this system, the upstream and downstream regions are defined as an extended range of 2000 nucleotides from the start and end positions of transcription. In addition, according to the shifting mechanism of repeating segments and the complementary based-paired nature in DNA double-stranded helical structures, several possible combinations of SSR patterns could be considered as an identical SSR motif within genetic loci. For example, any rotation of a basic repeat pattern is considered as the same SSR element, such as a "TA" repeat could be also defined as an equivalent repeat motif as an "AT" repeat pattern through one nucleotide shifting. Another situation of an identical SSR motif with different appearance is through complementary based paring and inverse reading from the DNA sequences. For example, the repeat pattern of "AGC" would appear as "GCT" within the other complementary strands of DNA. Therefore, to enumerate all possible SSR patterns in all DNA sequences under these two constraints, there are exactly 501 fundamental basic SSR patterns from 1 to 6 nucleotides in length [18]. However, there is one special condition that should be carefully considered when an SSR motif occurs in coding regions. Since the translation processes convert an mRNA sequence into a string of amino acids through the codon table encoding processes, the equivalent status due to shifting mechanisms and complementary strand should be limited. Here we provide their true translated protein sequences from the locations of identified SSR motifs, and the in-frame information will be clearly annotated when the orthologous repeat motifs are found in coding regions. Finally, to distinguish different SSR patterns from extensive genomic resources, the system defines an identifier for an SSR motif by its basic pattern in accordance with its corresponding genetic location within the specified gene. For example, "AG@Coding" in Ensembl gene id "ENSG00000069329" represents a specific repeated pattern "AG" appearing within the coding region of "ENSG00000069329." According to prerunning processes under various parameter settings for identifying all possible SSR motifs, in accordance with both detailed coordinates and annotated information from Ensembl database, we constructed a comprehensive SSR motif database for all genes from any specified species. These identified SSR motifs from each gene would be recognized as "tag" items for the following cross-species comparison, and all retrieved SSR tags from the input gene set will be further compared based on occurrence rates and applied to construct a multiscale tag cloud representation.

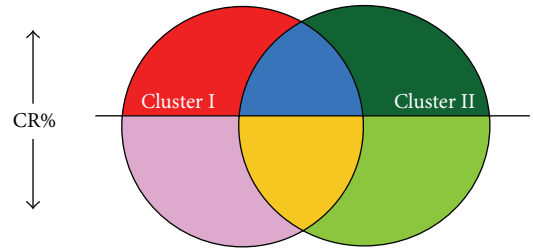
2.4. Grouped Species and Cross-Cluster Comparison. Due to tremendous amount of SSRs nonrandomly distributed in genome sequences, it is not an intuitive task to observe SSR biomarkers or identify gene regulatory related SSR motifs from an individual genome. Hence, we assume the conserved or exclusive SSR motifs providing important clues for identifying functional SSR motifs or representative biomarkers among various species. To emphasize the long-distance relationship from an evolutionary point of view, we have selected two groups of model vertebrate species for orthologous SSR motif comparison. The first group represents the mammalian species including *Bos taurus*, *Canis familiaris*, *Homo sapiens*, *Gorilla gorilla*, *Macaca mulatta*, and *Mus musculus*; the second group represents the fishery species including *Danio rerio*, *Gadus morhua*, *Gasterosteus aculeatus*, *Oryzias latipe*s, *Takifugu rubripes*, and *Tetraodon nigroviridis*. In addition to these twelve clustered species, we also included two widely used model organisms including *Drosophila melanogaster* and *Caenorhabditis elegans*. Nevertheless, in this developed system, users can either apply the previously defined two species groups or manually assign them into two clusters without any limitation. By integrating with cross-species comparison techniques and overrepresentation analysis from assigned gene sets, the SSR patterns with conserved and exclusive characteristics in selected genes between different species clusters can be recognized and treated. An identified conserved SSR motif would be initially defined as

an orthologous SSR motif if the conserved ratio meets the minimum threshold in an assigned species cluster. For example, a conserved ratio of 80% denotes the identified conserved SSR pattern that could be found in at least 80% of species in the assigned species cluster, which indicated that there are at least 5 ($6 * 80\% = 4.8$) different species possessing the orthologous gene(s) and holding the specific SSR pattern located within the same genetic region among all orthologous gene(s). Regarding the conditions of many-to-many orthologous genes, an SSR motif is defined as holding conserved feature as long as it could be detected in any one of its orthologous genes. The threshold level of conserved ratio can be assigned by users through interactive webpage settings.

Through cross-species comparison between two clustered groups, retrieving conserved or exclusive SSR motifs could help biologists in choosing significant biomarkers from a previously defined gene set before performing biological experiments. On the other hand, exclusive or common SSR motifs between two different species clusters might be regarded as important genetic markers under the evidences of biological evolution and functional conservation.

2.5. SSR Tag Cloud Visualization. Tag cloud visualization technique provides keyword representation of text data by showing each tag in various font sizes and colors. To enhance the importance of conserved and exclusive SSR motifs extracted from a set of specified homologous genes between two different species clusters, we adopted the tag cloud representation to display these identified SSR motifs according to their calculated weighting coefficients from query gene sets. In an SSR tag cloud, the tag size of each SSR motif not only indicates the conservation status of the motif among orthologous genes, but also displays the representativeness among different species clusters. A linear accumulation formula and normalization procedures for deciding SSR weighting coefficients were performed for tag size selection. This formula simply counts the number of occurrence times of each SSR motif found from each individual gene in different species clusters. According to the definitions of occurrence rate, if an identified SSR motif is well conserved in two different species clusters or highly represented in the specified gene set, the SSR tag will be assigned with a larger weighting coefficient. Accordingly, the SSR tag will be displayed with a bigger font size in the tag cloud.

In order to visually emphasize identified SSR motifs belonging to different species clusters, we applied different colors on SSR tags to distinguish the conserved and/or exclusive features of SSR biomarkers between two species clusters. In this study, red tags represent consensus SSR motifs for the first species cluster only and satisfy the conservation threshold in the first species cluster; pink tags are applied for representing consensus SSR motifs for the first species cluster only, but the conservation threshold is not satisfied; dark green tags represent consensus SSR motifs well conserved within the second species cluster only and these motifs also satisfied the conservation criterion in the second species cluster; light green tags denote consensus SSR motifs in the second species cluster only, but the conservation threshold is not satisfied; blue tags represent the identified SSR patterns well conserved in both species clusters and satisfy the species conservation percentage as well; yellow tags are applied to show identified SSR patterns conserved, but the species conservation criterion is not satisfied for the query gene set from both species clusters. The color-coded information in a resulting tag cloud is shown in Figure 1 and corresponding attributes are described in Table 1. The abbreviated term of CR% represents "conserved ratio" percentage of corresponding species clusters for each simulation.



*CR%: conserved ratio percentage

Figure 1: Color coded chart for tag cloud representation of identified SSR motifs between two species clusters and the criterion of conserved ratio.

Table 1: Relationship between colors, species clusters, and conserved ratios of detected SSR motifs.

Color	Species cluster	Conserved motif ratio
Red	I	≥CR%
Pink	I	<CR%
Blue	I and II	≥CR%
Yellow	I and II	<CR%
Dark green	II	≥CR%
Light green	II	<CR%

In the developed system, users can also try to identify imperfect SSR biomarkers by setting different tolerant levels, and the number of retrieved imperfect SSR motifs would be in accordance with the settings proportionally. Higher noisy rates allow more tolerant repeat patterns and reflect larger number of possible SSR motifs. Accordingly, the corresponding tag clouds could be depicted in multiscale representations under various noise threshold settings. In other words, different scales of tag clouds are composed of SSR motifs of different tolerant qualities. For instance, the highest quality of SSR tag cloud represents that all identified conserved SSR motifs are with perfect repeating patterns among different genes and group species. Contrarily, lower quality SSR tag clouds contain more tolerant SSR motifs within the tag image, and which may reflect evolutionary status due to gene specification and/or duplication events from either distant or close species. Multiscale tag clouds provide biologists with an easier way to compare and select suitable SSR candidate motifs as biomarkers through a progressive approach on different tolerance levels, which could be applied in various situations for further design of biological experiments.

Figure 2: Interface of the SSR tag cloud web system (http://ssrtc.cs.ntou.edu.tw/).

3. Results

3.1. SSR Tag Cloud Web System. In this study, we have developed an online web system (http://ssrtc.cs.ntou.edu.tw/) for identifying conserved and exclusive SSR biomarkers through cross-species cluster comparison. The main interface of the developed web system is shown in Figure 2. To discover significant SSR biomarker candidates from an automatically generated SSR tag cloud, a user is required to provide gene name(s) or keyword(s) of gene function and simply applies the default parameters for system prediction. In other words, a set of query genes could be defined at the first step by providing relevant EnsemblGene IDs, GO terms, or keywords. Besides, the thresholding settings of SSR feature parameters could also be assigned manually instead of default settings such as genetic region, length of basic pattern, minimum length of SSR motif, SSR quality, species cluster, and SSR motif conserved ratio. The genetic region and length of basic pattern are applied for distinguishing fundamental features of SSR motifs under cross-species cluster comparison. A *minimum SSR length* is applied to define the minimal length for identification of SSR motifs. The *SSR quality factor* represents a tolerance threshold for allowing imperfect SSRs as candidate biomarkers. The developed system initially provides three available settings for efficient identification: 1.0 for perfect SSRs, 0.8 and 0.9 for imperfect SSRs with 20% and 10% tolerant percentages for an identified SSR motif. The function of species cluster assignment is provided for cross-species comparison by classifying species of interest into two clusters. The parameter of *motif conserved ratio* is designed as the percentage of qualified species within a cluster possessing the conserved SSR motif within a target gene. Two different operation modes were designed for the *motif conserved ratio*. If a user chooses the condition of larger than or equal to *motif conserved ratio*, the system will display a resulting SSR tag cloud in 6 colors; otherwise an SSR tag cloud will appear in 3 colors only. Different color modes of an SSR tag cloud are defined in the previous section. Once all parameters and operation modes are defined, the system performs SSR biomarker evaluation automatically and generates a final SSR tag cloud for visualization. The font color of each SSR tag is mainly decided by the *motif conserved ratio* parameter, and the font size depends only on the occurrence frequency of an SSR element. Users can move the mouse device over any SSR item within the resulting tag clouds, and a total appearance number and conserved ratio of the selected SSR motif from the target genes of assigned species cluster will be displayed. The detailed information of each SSR tag is also available in a floating dialog box by clicking on it, which includes Ensembl gene ID, transcript ID of the specified gene possessing the target SSR motif, species name, coordinates in genomes, and DNA sequence contents. Additionally, if an SSR appears within coding regions, then its corresponding protein sequences could be recalled from Ensembl database and shown in an additional window.

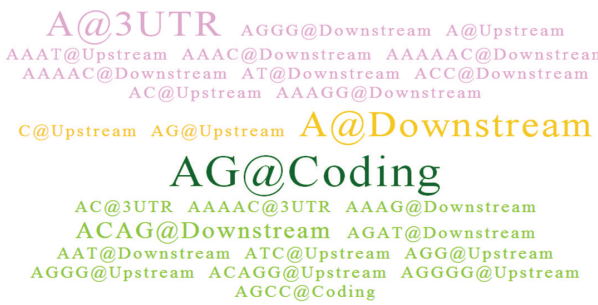


Figure 3: An SSR tag cloud example for ENSG00000069329 (*VPS35*) between two 6-species clusters.

3.2. SSR Biomarkers for Orthologous Genes. To demonstrate system performance, we have selected all orthologous genes from twelve vertebrate model species (except fruit fly and roundworm). All selected genes possess sequence identities higher than 80% compared to human genome individually. Under this criterion, there are totally 162 orthologous genes selected for the first testing case. If these twelve vertebrate species were classified into two species clusters including mammal and fishery species clusters for comparison, the conserved and exclusive SSR motifs for each gene could be successfully identified and significant SSR biomarker candidates for each individual gene were included in the Supplementary Material available online at http://dx.doi.org/10.1155/2014/678971. Here we only illustrate two genes of ENSG00000069329 and ENSG00000108883 as examples, and all conserved SSR motifs were carefully verified within all orthologous genes from twelve model species.

3.2.1. Case Study of ENSG00000069329 (VPS35). The Ensembl gene ID of ENSG00000069329 is a vacuolar protein sorting gene (*VPS35*) which possesses an average sequence identity of 80% by taking pairwise alignment between human and the other eleven model species. The resulting SSR tag cloud for *VPS35* was shown in Figure 3 by setting *SSR quality* of 80%, *minimum SSR length* of 20 nucleotides, and *motif conserved ratio* of 60% (i.e., required at least

4 species possessing identical SSR motifs in each species cluster). The first *species cluster* was assigned as the mammal group including human, macaque, mouse, cow, dog, and gorilla, and the second *species cluster* was assigned as the fishery group including zebrafish, stickleback, medaka, fugu, tetraodon, and cod. The *genetic region* parameters were set as searching for all regions except introns, and the length of *basic pattern* was selected from 1 to 6 nucleotides for comprehensive representation.

According to Figure 1 for SSR color codes, users can quickly observe that only three coconserved SSR motifs of "C@Upstream," "AG@Upstream," and "A@Downstream" in yellow were found between two species clusters. However, in this case, there is not any blue coded SSR tag in this experiment and which implies no coconserved SSR motif existing for at least 4 model species in each species cluster simultaneously. These three yellow color coded SSR tags were found due to their appearance in both species clusters but not well conserved with respect to the assigned conserved ratio. The dark green SSR tag of "AG@Coding" represented the consensus SSR motif could be found only in the second cluster of fishery species with more than 4 fishery species containing the SSR motif at coding region, but this motif pattern at coding region was not found in any mammal species from the first cluster. The light green SSR tags represented consensus SSR motifs which were found only in the fishery group but do not satisfied the *motif conserved ratio* requirement of 80%; that is, these light green coded SSR patterns were only found with less than 4 fishery species. On the other hand, the pink coded SSR tags represented consensus SSR motifs found only in the mammal species cluster exclusively with less than 4 mammal species. In addition, the dark green SSR tag of "AG@Coding" with the biggest font size implied this SSR holding as the most representative and exclusive feature for fishery species compared to mammal species.

3.2.2. Case Study of ENSG00000108883 (EFTUD2). The Ensemble gene ID of ENSG00000108883 is an elongation factor Tu GTP binding domain (*EFTUD2*) which possesses an average sequence identity of 80% by taking pairwise alignment between human species and other 11 model species individually. The resulting SSR tag cloud for *EFTUD2* was shown in Figure 4 by setting exactly the same parameters as the previous example. According to the resulting tag cloud, users can immediately identified that only one coconserved SSR tag of "ATC@Coding" could be found as a notable biomarker between two species clusters and it was well conserved across at least 4 species in each species cluster. Hence, the SSR tag was indicated by blue. Furthermore, one red coded SSR tag of "A@Downstream" represented the consensus SSR motifs found only in the first mammal species cluster and more than 4 species containing the SSR motif at coding region. However, this motif could not be found in any fishery species. The pink SSR tags represented all conserved SSR motifs found only in the mammal group but not satisfied the requirement of *Motif Conserved Ratio*. Similarly, the light green coded SSR tags represented consensus SSR motifs only found in the fishery species cluster exclusively with less than 4 fishery species. In addition, the red SSR tag of "A@Downstream" was shown with the biggest font size, which implied the SSR holding as the most representative and exclusive for mammal species compared to all other SSR candidates.

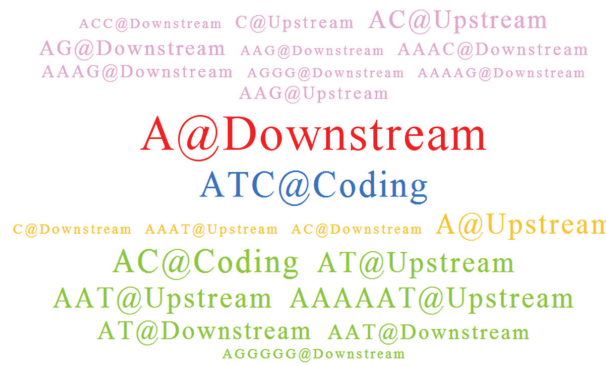


Figure 4: An SSR tag cloud example for ENSG00000108883 (*EFTUD2*) between two 6-species clusters.

Interestingly, the first gene, *VPS35* (ENSG00000069329), is associated with "Parkinson's disease (PD)" [19], and the second gene, *EFTUD2* (ENSG00000108883), causes "mandibulofacial dysostosis with microcephaly" [20]. In both cases, so far, scientists have only demonstrated that both diseases were caused by some gene mutations. Through *in silico* SSR biomarker detection by our proposed system, we could efficiently identify many important conserved and exclusive SSRs between two grouped species as biomarkers. However, without experimental verification, we could not make sure whether both diseases possess a true correlation with identified SSR motifs. To gain more confidence on the proposed system, we verified on some disease genes which were known to be associated with some specific SSR biomarkers. If a genetic disease is indeed caused by abnormal distributions of SSR motifs, we expect that our proposed SSR tag cloud representation system could identify those significant SSR biomarkers in an efficient and effective way.

3.3. Case Study of a Set of Skeletal Development Genes. To demonstrate functionally related SSR motifs, we have selected a gene set containing specific function of skeletal development. A total of 17 genes associated with such function are selected and these genes are *HOXA11*, *ZIC2*, *ALX4*, *HOXA2*, *DLX2*, *HOXA7*, *TWIST1*, *HOXC13*, *RUNX2*, *SOX9*, *HOXD11*, *HOXD13*, *GDF11*, *HLX*, *SIX3*, *HOXD8*, and *HOXA10* [21]. In this example, we have shown that the detailed information of each SSR tag is available in a floating dialog by clicking on it, and the appearance number and conserved ratio of a selected SSR motif from the target genes can be viewed by moving mouse cursor over the SSR tag.

The resulting SSR tag clouds from different combinatorial settings for 17 skeletal development related genes were shown in Figure 5. In Figure 5(a), the parameter settings were defined as follows: *SSR quality* of 90% for perfect SSR patterns, minimum *SSR length* of 20 nucleotides, *motif conserved ratio* of 80% (i.e., at least 5 species possessing

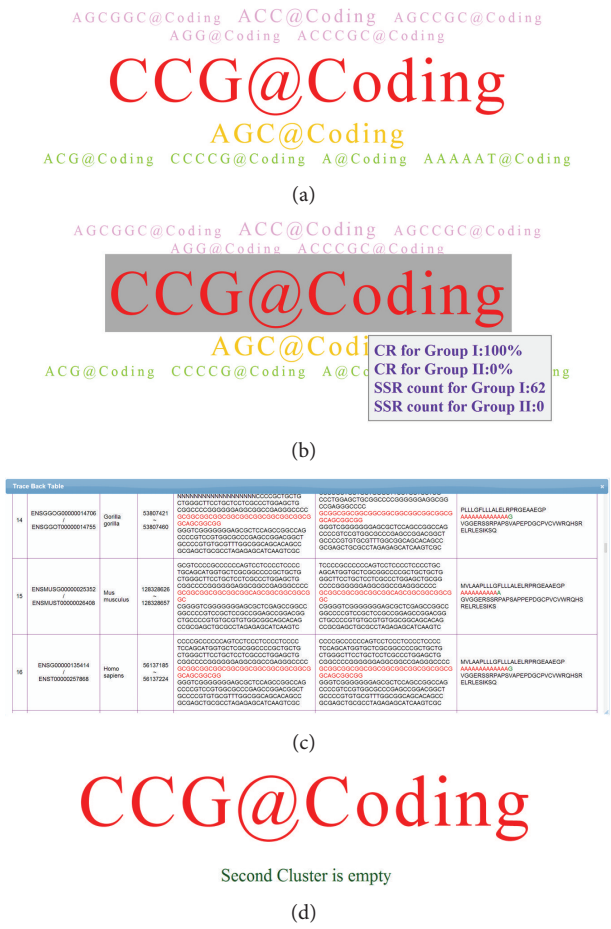


FIGURE 5: (a) SSR tag cloud for 17 skeletal development related genes constrained to coding regions; (b) results of moving a mouse device over the SSR tag of "CCG@Coding"; (c) detailed information of the SSR tag "CCG@Coding" in a floating dialog; (d) an SSR tag cloud for 17 skeletal development related genes by showing SSRs possessing high conserved ratios only.

identical SSR motifs in each species cluster), and all possible SSR candidates were shown. The first *species cluster* was assigned as the mammal group including human, macaque, mouse, cow, dog, and gorilla; the second *species cluster* was assigned as the fishery group including zebrafish, stickleback, medaka, fugu, tetraodon, and cod. The filter of *genetic region* was selected for coding region only, and the length of *basic pattern* was selected from 1 to 6 nucleotides for comprehensive representation. According to these settings, the simulated results were shown in Figure 5: the red coded SSR tag of "CCG@Coding" represented the only exclusive SSR motifs well conserved in mammal species. This tag could be found from at least 5 species within the mammal group, and it is highly correlated to the skeletal development related genes. Users can move a mouse device over the tag of "CCG@Coding," and the appearance number and conserved ratio of the selected SSR motif would be shown with a pop-up icon. In Figure 5(b), the CCG@Coding motifs appear in the mammal species cluster with a total of 62 times and a conserved ratio of 100%, while no such an SSR motif could be discovered from the skeletal development gene set within the fishery species cluster. If a user clicked on the tag of "CCG@Coding," detailed information of the SSR tag will be shown by a floating dialog with Ensembl gene ID, transcript ID, species name, coordinates in genomes, and DNA sequence contents. Particularly, if the SSRs appeared within coding regions, the table also provided the detailed information of cDNA sequence and its corresponding translated protein sequences. In Figure 5(c), the CCG repeated pattern in the last row of human's ENSG00000135414 (*GDF11*) gene is located at chromosome 12 and its coordinates are from 56137185 to 56137224. Since the CCG repeated pattern was found in coding regions, the table also provided the detailed information of DNA, cDNA, and corresponding protein sequence contents. Actually, this repeated pattern in *RUNX2* gene at coding region is a polyalanine peptide (GCC repeat in coding region), and it indeed plays a crucial role in cellular development function. Abnormal distribution of this polyalanine repeat biomarker might cause dysplasia disease, a genetic disorder of abnormal cellular development.

In Figure 5(d), most of parameters were set identically as Figure 5(a), except the display parameter was modified for showing highly conserved SSRs instead of showing all of identified SSRs. In the other words, tags with pink, light green, and yellow color codes would be hidden. The corresponding tag showed only one red coded tag of "CCG@Coding" existed under such high conservation requirements. Again, the SSR motif of "CCG@Coding" represented as a significant biomarker in mammal species highly correlated to the skeletal development related genes.

3.4. Case Study of Gene Ontology Term of "Embryonic Cranial Skeleton Morphogenesis". To demonstrate functionally related SSR motifs through GO term assignment, we selected a GO term of "embryonic cranial skeleton morphogenesis." The related genes annotated by this GO term include *TBX15, SIX4, DLX2, PRRX1, TWIST1, BMP4, SIX1, SMAD2, NIPBL, NODAL, WNT9B, TGFBR2, GAS1, SIX2, FOXC2, SMAD3, TBX1, TGFBR2, TBX15, GNAS, PRRX2, TGFBR1, TFAP2A, SMAD2, SETD2, BMP4, SMAD3, TWIST2, TFAP2A, SMAD3,* TGFBR1, and *BMP4*. To compare and show different results by various settings, we have tried several combinations of input parameters which were different from system default settings. In this case study, the parameter settings were defined as follows: *SSR quality* of 90% for perfect SSR patterns, *minimum SSR length* of 20 nucleotides, *motif conserved ratio* of 80% (i.e., at least 5 species possessing identical SSR motifs in each species cluster) and showed all possible SSR candidates. The first *species cluster* was assigned as mammal group and the second *species cluster* for fishery group as default settings. The filter of *genetic region* was selected for analyzing on coding regions only, and the length of *basic pattern* was selected from 1 to 6 nucleotides. According to these settings, the simulated results were shown in Figure 6(a). We could observe that there was only one red color coded SSR tag of "CCG@Coding" and which is the unique biomarker conserved in mammal species with

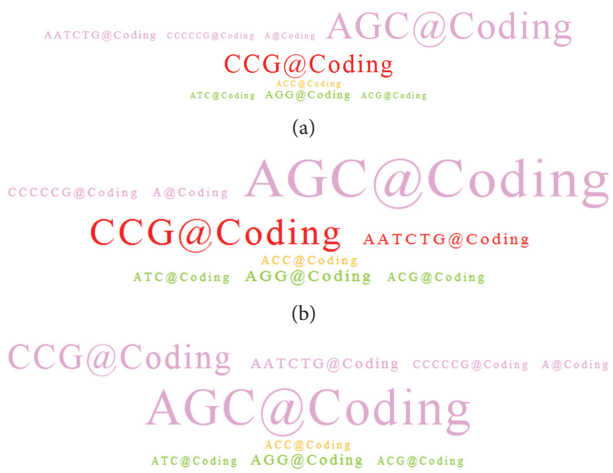


FIGURE 6: (a) SSR tag cloud for GO keyword "embryonic cranial skeleton morphogenesis" with motif conserved ratio of 80%; (b) motif conserved ratio of 60%; (c) motif conserved ratio of 100%.

respect to the embryonic cranial skeleton morphogenesis related genes.

Then, we lowered down the *motif conserved ratio* to 60% and the resulting SSR tag cloud was shown in Figure 6(b). We could observe that several tags were changed by their coded colors. Taking red color coded tags as an example, there was only one red tag "CCG@Coding" in previous Figure 6(a), but in Figure 6(b), we noticed that the red color coded SSR tags increased another tag of "AATCTG@Coding" which was displayed in originally denoted as pink in Figure 6(a). Inversely, if we increased the *motif conserved ratio* to 100%, the result was shown in Figure 6(c) with no red color coded SSR tag in this cloud. Compared to Figure 6(a), the original red tag of "CCG@Coding" was changed into pink due to only 5 out of 6 species in the mammal group holding the tag of "CCG@Coding." In both Figures 6(b) and 6(c), we simply observed that color coded tags may switch their colors through different *motif conserved ratio* adjustments. The higher setting of *motif conserved ratio* reduces the amount of red, green, and blue color coded tags.

3.5. An Example of Genetic Disease of "Huntington's Disease (HD)". To demonstrate genetic diseases caused by abnormal distribution of SSR motifs, we have selected a well-known neurodegenerative genetic disease "Huntington's disease (HD)" as an example. HD was found as an irregular distribution of polyglutamine expansions (CAG repeats) located within the coding regions of ENSG00000197386 (*HTT*) gene at chromosome 4 [22]. It appears with involuntary movements caused by losing muscle coordination and leads to psychiatric problems. The nucleotide repeat length and the average age of symptom occurrence of Huntington's disease were in inverse relationship [23].

The verification results of SSR tag cloud were shown in Figure 7, and the parameter settings were defined as follows: *SSR quality* of 100% and 80%, *minimum SSR length* of 20

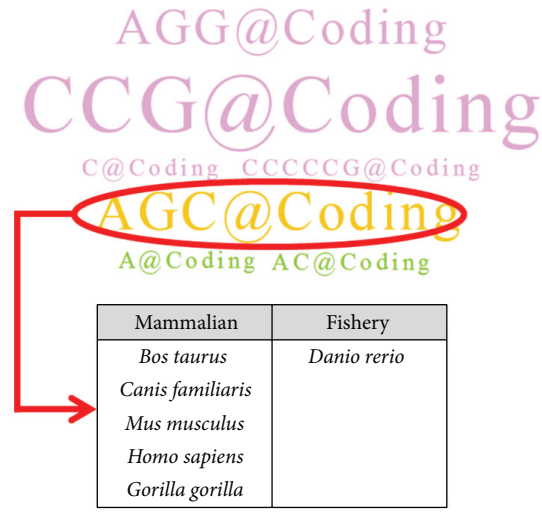


FIGURE 7: (a) SSR tag cloud for HTT gene with SSR quality of 80%, Motif Conserved Ratio of 80%, and 5 organisms holding the conserved SSR tag of "AGC@Coding"; (b) SSR quality of 100%, Motif Conserved Ratio of 80%, and only two species of human and cattle species holding the perfect SSR tag of "AGC@Coding".

nucleotides, *motif conserved ratio* of 80% (i.e., at least 5 species possessing identical SSR motifs in each species cluster), and with a selection of "show all SSRs". The first *species cluster* was assigned as mammal group while the second *species cluster* as fishery group. In Figure 7, we could observe the "AGC@Coding" in both two-tag clouds as an important biomarker. In fact, according to shifting transformation of SSR repeat pattern, the "AGC" repeat unit could be theoretically considered as the same pattern of "CAG" for efficient identification. However, SSRs located within coding regions would be further translated into their corresponding amino acid sequences according to precise loci verification on exon regions. Frame shifted SSRs in coding regions might result in different coded amino acids. For example, the coded amino acid of the trinucleotide pattern of "AGC" is serine(S) and "CAG" for glutamine (Q). Therefore, identified SSRs in coding regions should be carefully treated and translated into an appropriated protein sequence based on annotated genome database. In this example, we noticed that a significant SSR motif of "AGC@Coding" in HTT genes could be identified

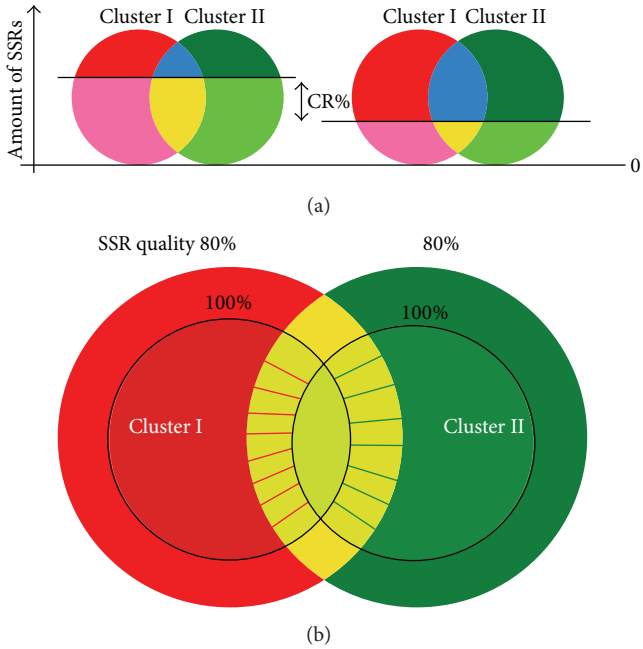


FIGURE 8: (a) Relationship between the parameter of *motif conserved ratio* and the amount of SSR tags in different colors; (b) relationship between the parameter of *SSR quality* and SSR tag colors.

with different sizes (occurrence rates) according to various *SSR quality* settings. This repeat motif in coding regions appears in most mammal species except macaque with a minimum length requirement of 20 nucleotides. Besides, only zebrafish possesses a similar repeat motif in coding region among all fishery species. When the parameter of *SSR quality* was increased to 100% (without any tolerance), the pattern of "AGC@Coding" (or equivalently to "CAG@Coding" in DNA sense strand) could be retrieved from both cattle and human in mammal species only. We could observe that the font size and color of each SSR tag were gradually changed according to different settings of tolerance rate. Accordingly, the tag of "AGC@Coding" appeared with the biggest icon in pink when compared to all other SSRs in coding regions, and it reflected the significance of exclusive features for mammal species compared to fishery species. These observations might also provide important information for biologists for animal species selection in future experimental studies regarding specific diseases.

4. Discussion

Two key parameters affect the color and size distribution within an SSR tag cloud. The first one is the *motif conserved ratio*. Different conserved ratio values change colors of SSR tags. When the *motif conserved ratio* increased, the amount of red, green, and blue tags might decrease. In Figure 8(a), Cluster I represents the first *species cluster* and Cluster II represents the second *species cluster*. The horizontal straight line in the figure represents a *motif conserved ratio* value. When the CR% threshold value is increased, the areas of red, blue, and dark green decreased. In contrast, when the CR% threshold value is decreased, the areas of red, blue, and dark green increased. The area is proportional to the amount of SSR tags.

The second important parameter for a tag cloud is the *SSR quality* threshold. As shown in Figure 8(b), different *SSR quality* values were not only changing the number of SSR tags but also transforming the colors. Increment of *SSR quality* value may reduce the amount of SSR tags, since the SSRs with higher qualities are always a subset of SSRs with lower qualities. When a quality threshold decreases to gain more SSR candidates, part of red and green tags might change their colors into yellow or blue tags, respectively. This is mainly caused by newly intersecting region after expanding SSR candidates.

Besides, a few common SSR tags originally coded in yellow might be transformed into either red or green through increasing the quality factors, which is mainly because the total number of species possessing certain SSR tag is decreased, and therefore the conserved SSR motifs between two species clusters might become representative SSR tags for one species cluster exclusively. In Table 2, a list of total amount of SSR motifs for each species is presented by setting a minimum *SSR length* of 20 nucleotides. The SSR quantities for mammal species are usually more than fishery species, and the increment of *SSR quality* value reduces the amount of SSR motifs in each species generally.

5. Conclusion

SSRs are nonrandomly distributed nucleotides in the genomes with repeating basic patterns of lengths from 1 to 6 nucleotides, and a large number of functional SSR motifs have been demonstrated as important biomarkers involved within various biological processes and gene regulations. Due to abundant number of SSRs in each species genomes, it is difficult to recognize significant SSR biomarkers or gene regulation related SSRs mainly based on repeat sequence length, genetic locations, and fundamental repeat pattern of an SSR motif. In this paper, we proposed the concept of identifying SSR biomarker candidate through cross-species cluster comparison on a specified set of target genes. The developed system provides an online tool with multiparameter selection functions, and the identified SSR motifs are displayed by a tag cloud visualization method. The exclusive and consensus SSR motifs between two species clusters are shown in different font colors and sizes in an efficient approach. The *in silico* comparison of SSR motifs across different species clusters may provide the clues and evidences for further understanding of evolutionary development and functional associations.

Conflict of Interests

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

Table 2: The number of SSR motifs of each species for various *SSR quality* settings.

Scientific name	Species name	*SSR quality* 80%	*SSR quality* 90%	*SSR quality* 100%
Danio rerio	Zebrafish	1,175,832	594,741	401,503
Gasterosteus aculeatus	Stickleback	160,413	87,343	51,779
Oryzias latipes	Medaka	122,505	37,730	15,460
Takifugu rubripes	Fugu	261,612	148,043	90,753
Tetraodon nigroviridis	Tetraodon	119,557	69,473	43,584
Gadus morhua	Cod	359,592	209,540	123,880
Homo sapiens	Human	3,023,284	1,406,186	644,338
Gorilla gorilla	Gorilla	757,571	344,973	152,403
Macaca mulatta	Macaque	1,075,737	526,515	225,403
Mus musculus	Mouse	2,463,222	1,301,019	812,873
Bos taurus	Cow	323,386	132,923	44,906
Canis familiaris	Dog	715,776	340,433	152,502
Caenorhabditis elegans	Roundworm	59,273	13,637	4,225
Drosophila melanogaster	Fruit fly	199,458	79,952	21,223

Acknowledgments

This work is supported by the Center of Excellence for the Oceans from National Taiwan Ocean University and National Science Council, Taiwan (NSC 102–2321-B-019-001 and NSC 101–2627-B-019-003 to T.-W. Pai), and Department of Health in Taiwan (DOH102-TD-B-111-004 to H.-T. Chang).

References

[1] B. Charlesworth, P. Sniegowski, and W. Stephan, "The evolutionary dynamics of repetitive DNA in eukaryotes," *Nature*, vol. 371, no. 6494, pp. 215–220, 1994.

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

[3] J. R. Brouwer, R. Willemsen, and B. A. Oostra, "Microsatellite repeat instability and neurological disease," *Bioessays*, vol. 31, no. 1, pp. 71–83, 2009.

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

[5] S. Mundlos, F. Otto, C. Mundlos et al., "Mutations involving the transcription factor CBFA1 cause cleidocranial dysplasia," *Cell*, vol. 89, no. 5, pp. 773–779, 1997.

[6] H. Y. Zoghbi and H. T. Orr, "Glutamine repeats and neurodegeneration," *Annual Review of Neuroscience*, vol. 23, pp. 217–247, 2000.

[7] C. L. Cheng, T. Q. Gao, Z. Wang, and D. D. Li, "Role of insulin/insulin-like growth factor 1 signaling pathway in longevity," *World Journal of Gastroenterology*, vol. 11, no. 13, pp. 1891–1895, 2005.

[8] K. A. Woods, C. Camacho-Hübner, D. Barter, A. J. L. Clark, and M. O. Savage, "Insulin-like growth factor I gene deletion causing intrauterine growth retardation and severe short stature," *Acta Paediatrica*, vol. 86, no. 423, pp. 39–45, 1997.

[9] N. B. Sutter, C. D. Bustamante, K. Chase et al., "A single IGF1 allele is a major determinant of small size in dogs," *Science*, vol. 316, no. 5821, pp. 112–115, 2007.

[10] 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.

[11] S. Lohmann, J. Ziegler, and L. Tetzlaff, "Comparison of tag cloud layouts: task-related performance and visual exploration," in *Human-Computer Interaction—INTERACT 2009*, vol. 5726 of *Lecture Notes in Computer Science*, pp. 392–404, 2009.

[12] S. Hennig, D. Groth, and H. Lehrach, "Automated gene ontology annotation for anonymous sequence data," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3712–3715, 2003.

[13] B. M. Good, E. A. Kawas, B. Kuo, and M. D. Wilkinson, "iHOPerator: User-scripting a personalized bioinformatics Web, starting with the iHOP website," *BMC Bioinformatics*, vol. 7, article 534, 2006.

[14] S. A. Samarajiwa, S. Forster, K. Auchettl, and P. J. Hertzog, "INTERFEROME: the database of interferon regulated genes," *Nucleic Acids Research*, vol. 37, no. 1, pp. D852–D857, 2009.

[15] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, "Revigo summarizes and visualizes long lists of gene ontology terms," *PLoS ONE*, vol. 6, no. 7, Article ID e21800, 2011.

[16] E. Birney, T. D. Andrews, P. Bevan et al., "An overview of Ensembl," *Genome Research*, vol. 14, pp. 925–928, 2004.

[17] C. M. Chen, C. C. Chen, T. H. Shih, T. W. Pai, C. H. Hu, and W. S. Tzou, "Efficient algorithms for identifying orthologous simple sequence repeats of disease genes," *Journal of Systems Science and Complexity*, vol. 23, pp. 906–916, 2010.

[18] É. Nascimento, R. Martinez, A. R. Lopes et al., "Detection and selection of microsatellites in the genome of Paracoccidioides brasiliensis as molecular markers for clinical and epidemiological studies," *Journal of Clinical Microbiology*, vol. 42, no. 11, pp. 5007–5014, 2004.

[19] A. Zimprich, A. Benet-Pagès, W. Struhal et al., "A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset parkinson disease," *The American Journal of Human Genetics*, vol. 89, no. 1, pp. 168–175, 2011.

[20] D. V. Luquetti, A. V. Hing, M. J. Rieder, D. A. Nickerson, E. H. Turner, J. Smith et al., "Mandibulofacial dysostosis with microcephaly caused by EFTUD2 mutations: expanding the phenotype," *The American Journal of Medical Genetics A*, vol. 161, pp. 108–113, 2013.

[21] M. A. Lines, L. Huang, J. Schwartzentruber et al., "Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly," *The American Journal of Human Genetics*, vol. 90, no. 2, pp. 369–377, 2012.

[22] J. W. Fondon III and H. R. Garner, "Molecular origins of rapid and continuous morphological evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 52, pp. 18058–18063, 2004.

[23] M. E. MacDonald, C. M. Ambrose, M. P. Duyao et al., "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes," *Cell*, vol. 72, no. 6, pp. 971–983, 1993.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 782490, 7 pages
http://dx.doi.org/10.1155/2014/782490

Research Article

Global Analysis of miRNA Gene Clusters and Gene Families Reveals Dynamic and Coordinated Expression

Li Guo,[1] Sheng Yang,[1] Yang Zhao,[1] Hui Zhang,[1] Qian Wu,[2] and Feng Chen[1]

[1] *Department of Epidemiology and Biostatistics and Ministry of Education Key Lab for Modern Toxicology, School of Public Health, Nanjing Medical University, Nanjing 211166, China*
[2] *State Key Laboratory of Reproductive Medicine, Department of Hygienic Analysis and Detection, School of Public Health, Nanjing Medical University, Nanjing 211166, China*

Correspondence should be addressed to Qian Wu; wuqian@njmu.edu.cn and Feng Chen; fengchen@njmu.edu.cn

Received 17 January 2014; Accepted 26 February 2014; Published 25 March 2014

Academic Editor: Xing-Ming Zhao

To further understand the potential expression relationships of miRNAs in miRNA gene clusters and gene families, a global analysis was performed in 4 paired tumor (breast cancer) and adjacent normal tissue samples using deep sequencing datasets. The compositions of miRNA gene clusters and families are not random, and clustered and homologous miRNAs may have close relationships with overlapped miRNA species. Members in the miRNA group always had various expression levels, and even some showed larger expression divergence. Despite the dynamic expression as well as individual difference, these miRNAs always indicated consistent or similar deregulation patterns. The consistent deregulation expression may contribute to dynamic and coordinated interaction between different miRNAs in regulatory network. Further, we found that those clustered or homologous miRNAs that were also identified as sense and antisense miRNAs showed larger expression divergence. miRNA gene clusters and families indicated important biological roles, and the specific distribution and expression further enrich and ensure the flexible and robust regulatory network.

1. Introduction

The small non-coding RNA regulatory molecules, microRNAs (miRNAs), play an important role in multiple biological processes through negatively regulating gene expression [1]. Abnormally expressed miRNAs may contribute to various human diseases, including cancer development, and some have been identified as potential oncomiRs or tumor suppressors [2, 3]. Some miRNAs are preferentially located at fragile sites and regions and are abnormally expressed in cancer samples [4]. Those deregulated miRNAs have been widely studied as potential biomarkers, especially for circulating miRNAs in human diseases [5–7].

miRNAs in gene cluster or family may have functional relationships via coregulating or coordinately regulating biological processes [8, 9], although they have various expression levels due to complex maturation and degradation mechanisms [10–12]. These clustered miRNAs are quite popular in metazoan genomes, and they may be involved in homologous miRNA genes via duplication evolutionary histories [13–15]. Simultaneously, the phenomenon of multicopy miRNA precursors (pre-miRNAs) further complicates the distributions of miRNA gene cluster and family and also implicates the dynamic evolutionary process in the miRNA world [15, 16]. The systematic analysis based on clustered and homologous miRNAs is quite necessary to unveil the potential functional correlation and contribution in tumorigenesis.

In the present study, to further understand the potential expression and functional correlations between miRNAs, we performed a global analysis of miRNA gene clusters and families in breast cancer using small RNA deep sequencing datasets. These related miRNAs may have higher sequence similarity (homologous miRNAs) or may be expressed in a single polycistronic transcript with close physical distance on chromosome (clustered miRNAs). They have been identified as cooperative regulatory molecules via contributing

to multiple biological processes. Simultaneously, they also have close phylogenetic relationships through complex evolutionary process. Based on their functional and evolutionary relationships, the expression analysis will provide information of indirect interaction between miRNAs and potential contribution in cancer development.

2. Materials and Methods

2.1. Source Data. High-throughput miRNA sequencing datasets of 4 paired tumor (breast cancer) and adjacent normal tissues (P1, P5, P6, and P7) were obtained from Guo et al. [17]. The information on miRNA gene clusters and families was obtained from the public miRBase database (Release 19.0, http://www.mirbase.org/). Abundantly expressed miRNA gene clusters and families were collected and further analyzed according to relative expression levels. To comprehensively track the expression profiles between clustered or homologous miRNAs, we collected and analyzed all the members of miRNA clusters and families if one member was abundantly expressed in a sample.

2.2. Expression Analysis. The expression patterns were estimated using the relative expression levels (percentage) in every miRNA gene cluster or family. Simultaneously, due to dynamic expression across different individuals, equally mixed datasets were also used to estimate the expression patterns. We analyzed the potential relationships between miRNA gene clusters and families, especially some miRNAs could be yielded by multicopy pre-miRNAs. According to abundantly expressed miRNAs, we attempted to discover the potential cross-distribution and expression patterns between clustered miRNAs and homologous miRNAs. Moreover, we also focused on those clustered miRNAs and homologous miRNAs that were identified as sense and antisense miRNAs in the specific genome locus. Further expression analysis was performed based on the 4 paired datasets and mixed datasets, respectively.

2.3. Gene Ontology Enrichment Analysis. Experimentally validated target mRNAs of deregulated miRNAs were obtained from the miRTarBase database [18]. For those miRNAs with less or no validated targets, target mRNAs were predicted based on "seed sequences" using the TargetScan program [19]. According to these target mRNAs of deregulated miRNA gene clusters and families, the functional enrichment analysis was performed using CapitalBio Molecule Annotation System V4.0 (MAS, http://bioinfo.capitalbio.com/mas3/).

3. Results

Abundantly expressed clustered and homologous miRNAs were selected to perform further analysis. Some abundantly and abnormally expressed miRNAs (such as miR-23a, miR-23b, miR-24, miR-222, and miR-29a) had been experimentally validated using real-time PCR in breast cancer samples [20]. Interestingly, we found that many miRNA gene clusters and families had close relationships or had overlapped members (Tables S1 and S2; see Supplementary Material available online at http://dx.doi.org/10.1155/2014/782490). Some miRNAs could be yielded by different pre-miRNAs, and the phenomenon of multicopy pre-miRNAs largely contributed to the complex relationships. Generally, these pre-miRNAs may be located on different chromosomes, different strands of the same chromosome (including sense and antisense strands), or different regions on the same strand. The various distributions complicated the compositions of miRNA gene clusters and families. For example, miR-221 and miR-222 were members of miR-221 gene family with higher sequence similarity, but they were also clustered on chromosome X and identified as miR-222 gene cluster. Homologous miRNA members could be located in different gene clusters through locating on different genomic regions or different chromosomes. For example, miR-23a and miR-27a were clustered on chromosome 19, while miR-23b and miR-27b were located in a cluster on chromosome 9. Simultaneously, sense and antisense miRNA genes were also involved in the gene cluster and family (Tables S1 and S2). miR-103a and miR-103b were homologous miRNA species (they were homologous members in miR-103 gene family), while their precursors were located on the sense and antisense strands of chromosomes 5 and 20, respectively (miR-103a-2 and miR-103a-1 gene clusters could be detected based on their multicopy pre-miRNAs).

Clustered and homologous miRNAs always showed consistent deregulation patterns in tumor samples (Figure 1(a)), although they had various expression levels (Figure 1(b)). They might show expression divergence as well as individual difference across different samples. The dynamic expression patterns in miRNA gene clusters and families were quite popular, even though they might be cotranscribed as a single polycistronic unit or had higher sequence similarity. For example, one member was abundantly expressed, while another clustered or homologous member had lower expression level (Figure 1(b)). The deregulation patterns were also influenced by the various expression levels, especially some were rarely expressed. The fold change ($\log_2$) showed larger divergence between different clustered or homologous miRNA species and between different individuals (Figure 1). Furthermore, we also performed the expression analysis based on the mixed datasets. Similar expression patterns could be detected (Figure 2). The divergence of fold change existed, but the difference had been largely reduced than the expression analysis based on each pair of samples (Figures 1 and 2).

For those miRNA gene clusters and families that were involved in sense and antisense miRNAs, we also analyzed their expression patterns. As expected, they always showed larger expression divergence (or both of them were rarely expressed): if one member had abundant expression level, another would be rarely detected (Figure 3). The sense and antisense miRNAs could be perfectly reverse complementarily binding to each other, although they may also be homologous miRNA genes with higher sequence similarity.

According to the predicted target mRNAs, the common targets could be detected between clustered or homologous

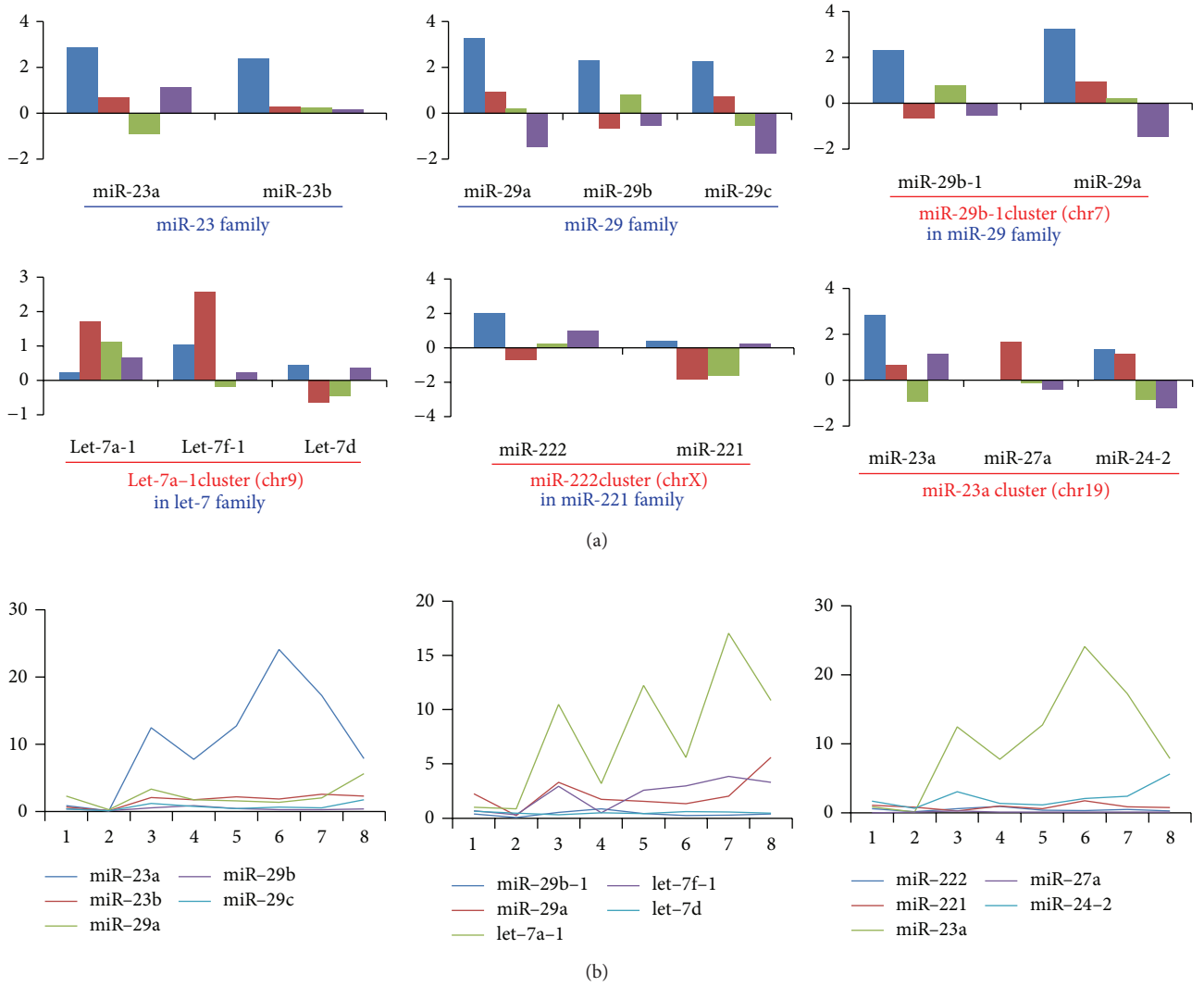


FIGURE 1: Examples of deregulation patterns of miRNA gene clusters and families (a) and their dynamic expression patterns (b). (a) Members in these miRNA gene clusters and families may be repeated. Some clusters are members of a specific gene family. The horizontal axis indicates the miRNA gene cluster or gene family; and the vertical axis indicates the fold change value ($\log_2$) based on each pair of tumor and adjacent normal samples. Bars in different colors (blue, red, green, and purple) indicate fold change value ($\log_2$) of the four pairs of tumor and adjacent normal samples, respectively; (b) dynamic expression across the 8 involved samples (P1-tumor, P1-normal, P5-tumor, P5-normal, P6-tumor, P6-normal, P7-tumor, and P7-normal). The horizontal axis indicates the 8 samples, and the vertical axis indicates the relative expression (percentage).

miRNAs (Table S3). Functional enrichment analysis of deregulated miRNA groups showed that they had versatile roles in multiple basic biological processes such as regulation of transcription and signal transduction (Table 1).

4. Discussion

miRNAs have been widely studied as crucial regulatory molecules, but the global expression patterns of miRNA gene clusters and families are little known. These clustered or homologous miRNAs have potential, functional, and evolutionary relationships, and they may coregulate or coordinately regulate multiple biological processes. The potential coordinated interaction complicates the coding-non-coding RNA regulatory network and enriches the miRNA-mRNA and miRNA-miRNA interactions [21, 22]. Sense and antisense miRNAs have been characterized as potential miRNA-miRNA interaction with larger expression divergence (Figure 3). Recent studies have shown that these endogenous complementary miRNAs can restrict the transcription or maturation process of one another [23–27]. The perfectly reverse binding suggests that miRNA-miRNA interaction may be a potential regulatory method in the miRNA world [21]. Further, the compositions of gene clusters and families are not random and independent, and the phenomenon of multicopy pre-miRNAs further

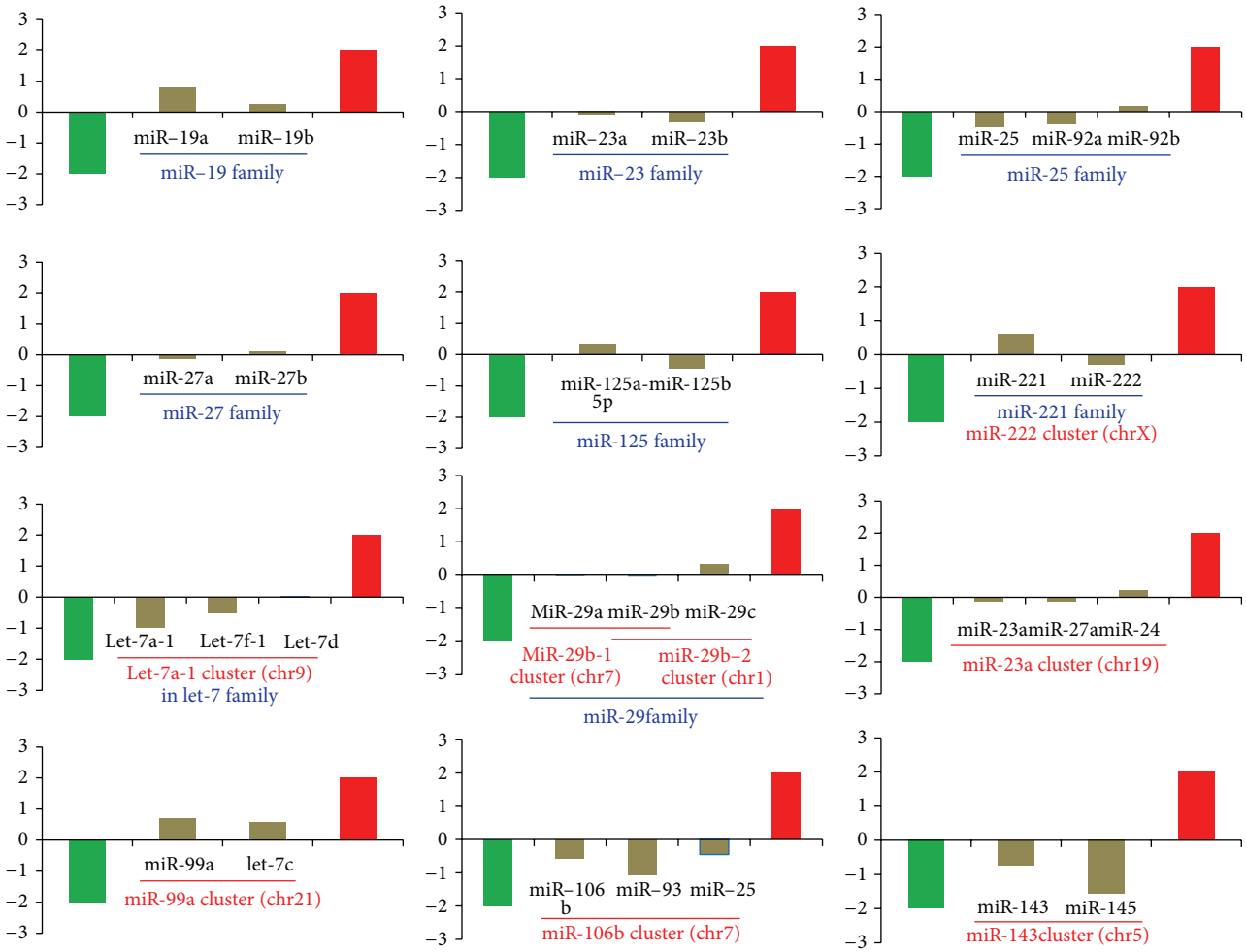


Figure 2: Examples of consistent or similar deregulation patterns in clustered and homologous miRNAs based on the equally mixed datasets. The horizontal axis indicates the miRNAs and the involved gene cluster and family, and the vertical axis indicates the fold change ($\log_2$) based on the equally mixed datasets of tumor and normal samples. The green and red bars indicate the threshold values (2 and −2).

Table 1: Enriched GO terms based on experimentally validated target mRNAs in Figure 1.

GO term	Count	*P* value
GO:0006355 regulation of transcription, DNA-dependent	24	$5.28E-26$
GO:0006350 transcription	18	$7.48E-17$
GO:0007165 signal transduction	18	$2.37E-14$
GO:0007275 development	15	$1.82E-13$
GO:0006508 proteolysis	14	$3.32E-17$
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	12	$1.05E-21$
GO:0007155 cell adhesion	10	$1.67E-12$
GO:0006915 apoptosis	10	$5.62E-12$
GO:0008285 negative regulation of cell proliferation	9	$7.05E-15$
GO:0006468 protein amino acid phosphorylation	9	$2.03E-11$
GO:0006917 induction of apoptosis	8	$9.59E-14$
GO:0042981 regulation of apoptosis	8	$2.75E-10$

Here, we only list important GO terms that involved at least 8 target mRNAs of differentially expressed miRNAs. Count indicates involved number of target mRNAs; *P* value indicates enrichment *P* value.

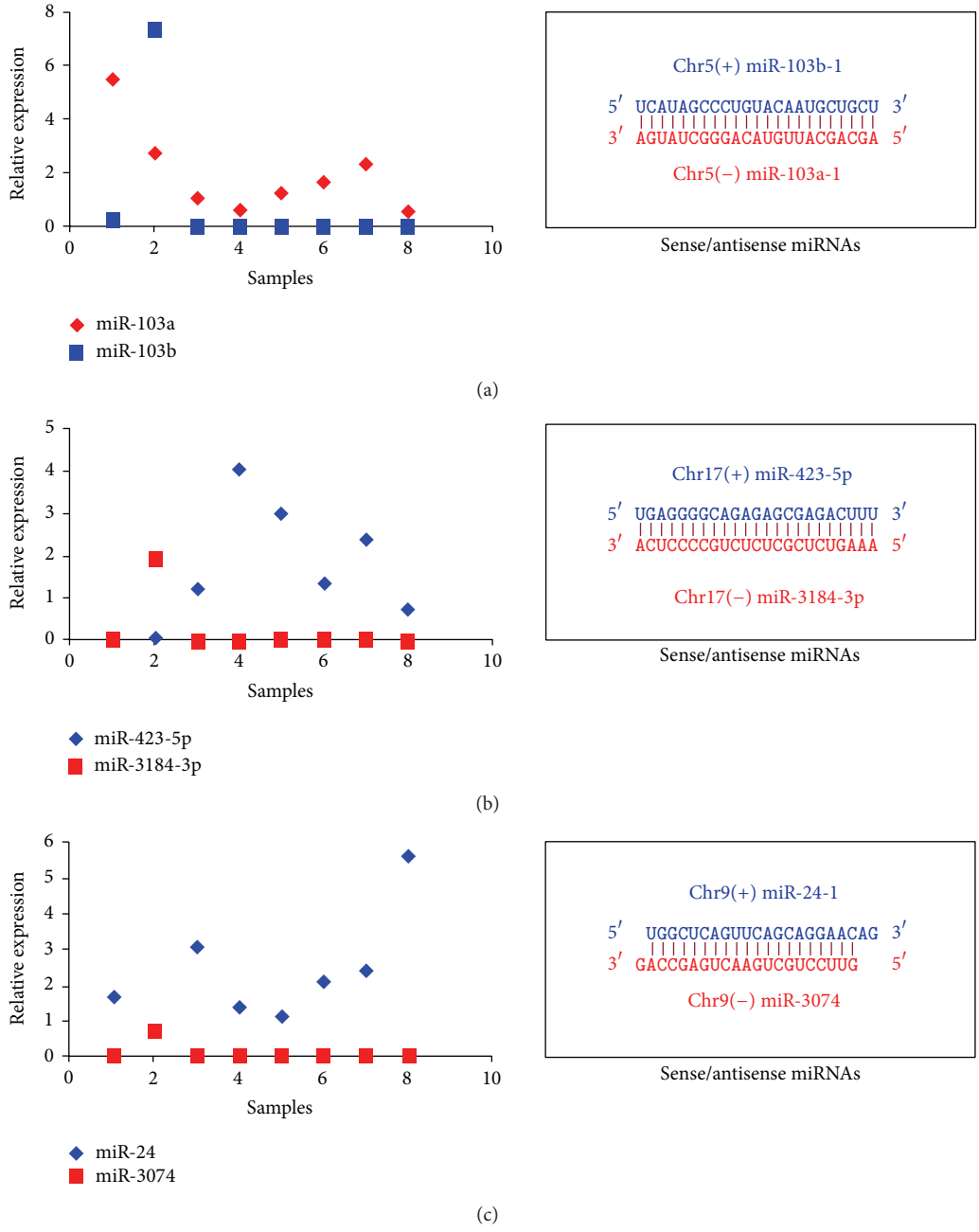


Figure 3: Examples of diverged expression of sense and antisense miRNAs in miRNA gene clusters and families. The detailed sense and antisense miRNA sequences are also presented on light, and they can perfectly reverse complementarily binding to each other. (a) miR-103a and miR-103b are homologous miRNAs and also clustered on chromosomes 5 and 20 as sense and antisense miRNAs (the figure only lists the sequences on chromosome 5); (b) miR-423 and miR-3184 are a pair of sense and antisense miRNAs on chromosome 17; (c) miR-24-1 and miR-3074 are sense and antisense miRNAs on chromosome 9 and also clustered in miR-23b gene cluster.

complicates the distributions of miRNAs [28]. Clustered and homologous miRNAs always have close relationships with overlapped members (Tables S1 and S2). The interesting distributions and relationships may be mainly derived from the complex duplication history that may adapt to the functional and evolutionary pressures [13–15, 29].

Although clustered and homologous miRNA members are involved in various and inconsistent enrichment levels via maturation and degradation mechanisms, they are prone to present consistent or similar deregulation patterns in tumor samples (Figures 1 and 3). Across different samples, miRNAs may show the larger expression divergence. The reason may be partly derived from the deep sequencing datasets with higher sensitivity and potential divergence during sequencing and sample preparation. On the other hand, the individual difference also leads to the expression

divergence, especially for these patients may be involved in different degrees or stages of breast cancer, although they are clinically characterized as primary breast cancer. Multiple factors may contribute to occurrence and development of breast cancer, and different samples may be prone to detect slightly inconsistent miRNA expression profiles. The dynamic expression patterns may contribute to the robust regulatory network and adapt to specific intracellular environment. Indeed, these miRNA gene clusters and families have important roles in multiple biological processes (Table 1). The consistent deregulation patterns contribute to their potential coordinated interaction, although they indicate various expression levels.

Furthermore, other factors also contribute to the expression divergence in miRNA gene clusters and families. Firstly, the phenomenon of cross-mapping or multiple mapping contributes to the relative expression levels [23, 30], especially between those homologous miRNAs. The same sequencing fragments can be mapped to different pre-miRNA sequences, and any arbitrary selection will influence the final expression analysis. Secondly, multiple pre-miRNAs have been identified that can yield the same miRNAs. However, it is hard to infer the genuine origin. These multiple pre-miRNAs are always located on different chromosomes or different strands on the same chromosomes. In the typical analysis, we always analyze the mature miRNAs and rarely consider their real origins. The default analysis would influence the expression patterns of members in miRNA gene clusters. Clustered miRNAs are characterized based on the location distributions of miRNA genes, but mature miRNAs are used to estimate the final expression levels. The arbitrary and default selection may lead to the imprecise expression analysis. Finally, an miRNA locus can yield many sequences with various $5'$ and/or $3'$ ends due to imprecise cleavage of Drosha and Dicer [31–33]. These multiple miRNA variants, also termed isomiRs, largely enrich the miRNA study and coding-non-coding RNA regulatory network as physical miRNA isoforms. These multiple isomiRs also influence the expression estimation, especially expression analysis based on the most abundant isomiR, the canonical miRNA, or sum of all isomiRs, respectively. Simultaneously, these various sequences also contribute to the phenomenon of cross-mapping between different miRNAs [23]. In the present study, the expression analysis at the miRNA level (based on the sum of all isomiRs) is not comprehensive. Collectively, expression divergence between miRNAs is more complexity *in vivo*, which may contribute to the dynamic regulatory network.

Taken together, although various expression levels can be detected, consistent or similar deregulation patterns are always found between clustered or homologous miRNAs. The expression patterns provide an opportunity to coregulate or coordinately regulate biological processes. Therefore, the dynamic and coordinated expression may have important biological roles, which should be derived from the functional and evolutionary pressures. As flexible regulatory molecules, multiple miRNAs can negatively regulate biological pathways based on potential coordinated interaction (e.g., based on miRNA gene clusters and families). Further study should be performed that clustered and/or homologous miRNAs would be potential biomarkers to study the mechanisms in tumorigenesis.

Conflict of Interests

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 61301251, 81072389, 81373102, and 81102182), the Research Fund for the Doctoral Program of Higher Education of China (nos. 211323411002 and 20133234120009), the China Postdoctoral Science Foundation funded project (no. 2012M521100), the Key Grant of Natural Science Foundation of the Jiangsu Higher Education Institutions of China (no. 10KJA33034), the National Natural Science Foundation of Jiangsu (no. BK20130885), the Natural Science Foundation of the Jiangsu Higher Education Institutions (nos. 12KJB310003 and 13KJB330003), the Jiangsu Planned Projects for Postdoctoral Research Funds (no. 1201022B), the Science and Technology Development Fund Key Project of Nanjing Medical University (no. 2012NJMU001), and the Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institutions.

References

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

[2] G. A. Calin, "A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia (vol 353, pg 1793, 2005)," *New England Journal of Medicine*, vol. 355, pp. 533–533, 2006.

[3] C. Caldas and J. D. Brenton, "Sizing up miRNAs as cancer genes," *Nature Medicine*, vol. 11, no. 7, pp. 712–714, 2005.

[4] 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.

[5] C. Swanton and C. Caldas, "Molecular classification of solid tumours: towards pathway-driven therapeutics," *British Journal of Cancer*, vol. 100, no. 10, pp. 1517–1522, 2009.

[6] D. Madhavan, M. Zucknick, M. Wallwiener et al., "Circulating miRNAs as surrogate markers for circulating tumor cells and prognostic markers in metastatic breast cancer," *Clinical Cancer Research*, vol. 18, pp. 5972–5982, 2012.

[7] M. Redova, J. Sana, and O. Slaby, "Circulating miRNAs as new blood-based biomarkers for solid cancers," *Future Oncology*, vol. 9, no. 3, pp. 387–402, 2013.

[8] L. P. Lim, M. E. Glasner, S. Yekta, C. B. Burge, and D. P. Bartel, "Vertebrate microRNA genes," *Science*, vol. 299, no. 5612, p. 1540, 2003.

[9] J. Z. Xu and C. W. Wong, "A computational screen for mouse signaling pathways targeted by microRNA clusters," *RNA*, vol. 14, no. 7, pp. 1276–1283, 2008.

[10] J. Yu, F. Wang, G. Yang et al., "Human microRNA clusters: genomic organization and expression profile in leukemia cell lines," *Biochemical and Biophysical Research Communications*, vol. 349, no. 1, pp. 59–68, 2006.

[11] L. Guo and Z. Lu, "Global expression analysis of miRNA gene cluster and family based on isomiRs from deep sequencing data," *Computational Biology and Chemistry*, vol. 34, no. 3, pp. 165–171, 2010.

[12] S. R. Viswanathan, C. H. Mermel, J. Lu, C. Lu, T. R. Golub, and G. Q. Daley, "MicroRNA expression during trophectoderm specification," *PLoS ONE*, vol. 4, no. 7, Article ID e6143, 2009.

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

[14] R. Zhang, Y. Peng, W. Wang, and B. Su, "Rapid evolution of an X-linked microRNA cluster in primates," *Genome Research*, vol. 17, no. 5, pp. 612–617, 2007.

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

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

[17] L. Guo, Y. Zhao, S. Yang, M. Cai, Q. Wu, and F. Chen, "Genome-wide screen for aberrantly expressed miRNAs reveals miRNA profile signature in breast cancer," *Molecular Biology Reports*, vol. 40, no. 3, pp. 2175–2186, 2013.

[18] S.-D. Hsu, F. Lin, W. Wu et al., "MiRTarBase: a database curates experimentally validated microRNA-target interactions," *Nucleic Acids Research*, vol. 39, no. 1, pp. D163–D169, 2011.

[19] B. P. Lewis, I.-H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge, "Prediction of mammalian microRNA targets," *Cell*, vol. 115, no. 7, pp. 787–798, 2003.

[20] Q. Wu, C. Wang, Z. Lu, L. Guo, and Q. Ge, "Analysis of serum genome-wide microRNAs for breast cancer detection," *Clinica Chimica Acta*, vol. 413, no. 13-14, pp. 1058–1065, 2012.

[21] L. Guo, B. Sun, Q. Wu, S. Yang, and F. Chen, "miRNA-miRNA interaction implicates for potential mutual regulatory pattern," *Gene*, vol. 511, no. 2, pp. 187–194, 2012.

[22] L. Guo, Y. Zhao, S. Yang, H. Zhang, and F. Chen, "Integrative analysis of miRNA-mRNA and miRNA-miRNA interactions," *BioMed Research International*, vol. 2014, Article ID 907420, 8 pages, 2014.

[23] L. Guo, T. Liang, W. Gu, Y. Xu, Y. Bai, and Z. Lu, "Cross-mapping events in miRNAs reveal potential miRNA-Mimics and evolutionary implications," *PLoS ONE*, vol. 6, no. 5, Article ID e20517, 2011.

[24] K. E. Shearwin, B. P. Callen, and J. B. Egan, "Transcriptional interference-a crash course," *Trends in Genetics*, vol. 21, no. 6, pp. 339–345, 2005.

[25] C. F. Hongay, P. L. Grisafi, T. Galitski, and G. R. Fink, "Antisense transcription controls cell fate in Saccharomyces cerevisiae," *Cell*, vol. 127, no. 4, pp. 735–745, 2006.

[26] A. Stark, N. Bushati, C. H. Jan et al., "A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands," *Genes and Development*, vol. 22, no. 1, pp. 8–13, 2008.

[27] E. C. Lai, C. Wiel, and G. M. Rubin, "Complementary miRNA pairs suggest a regulatory role for miRNA:miRNA duplexes," *RNA*, vol. 10, no. 2, pp. 171–175, 2004.

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

[29] A. M. Heimberg, L. F. Sempere, V. N. Moy, P. C. J. Donoghue, and K. J. Peterson, "MicroRNAs and the advent of vertebrate morphological complexity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 8, pp. 2946–2950, 2008.

[30] M. J. L. de Hoon, R. J. Taft, T. Hashimoto et al., "Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries," *Genome Research*, vol. 20, no. 2, pp. 257–264, 2010.

[31] R. D. Morin, M. D. O'Connor, M. Griffith et al., "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells," *Genome Research*, vol. 18, pp. 610–621, 2008.

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

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

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 127572, 12 pages
http://dx.doi.org/10.1155/2014/127572

Research Article

Identifying Potential Clinical Syndromes of Hepatocellular Carcinoma Using PSO-Based Hierarchical Feature Selection Algorithm

Zhiwei Ji[1] and Bing Wang[1,2,3]

[1] *School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China*
[2] *The Advanced Research Institute of Intelligent Sensing Network, Tongji University, Shanghai 201804, China*
[3] *The Key Laboratory of Embedded System and Service Computing, Tongji University, Ministry of Education, Shanghai 201804, China*

Correspondence should be addressed to Bing Wang; wangbing@ustc.edu

Received 17 December 2013; Revised 7 February 2014; Accepted 10 February 2014; Published 17 March 2014

Academic Editor: Jose C. Nacher

Hepatocellular carcinoma (HCC) is one of the most common malignant tumors. Clinical symptoms attributable to HCC are usually absent, thus often miss the best therapeutic opportunities. Traditional Chinese Medicine (TCM) plays an active role in diagnosis and treatment of HCC. In this paper, we proposed a particle swarm optimization-based hierarchical feature selection (PSOHFS) model to infer potential syndromes for diagnosis of HCC. Firstly, the hierarchical feature representation is developed by a three-layer tree. The clinical symptoms and positive score of patient are leaf nodes and root in the tree, respectively, while each syndrome feature on the middle layer is extracted from a group of symptoms. Secondly, an improved PSO-based algorithm is applied in a new reduced feature space to search an optimal syndrome subset. Based on the result of feature selection, the causal relationships of symptoms and syndromes are inferred via Bayesian networks. In our experiment, 147 symptoms were aggregated into 27 groups and 27 syndrome features were extracted. The proposed approach discovered 24 syndromes which obviously improved the diagnosis accuracy. Finally, the Bayesian approach was applied to represent the causal relationships both at symptom and syndrome levels. The results show that our computational model can facilitate the clinical diagnosis of HCC.

1. Introduction

Hepatocellular carcinoma (HCC) is the third most common cause of cancer-related death worldwide and the leading cause of death in patients with cirrhosis [1, 2]. In clinical practice, symptoms attributable to HCC are usually absent, so the majority of patients are diagnosed with advanced disease, often precluding potentially curative therapies. This has resulted, in part, in a 5-year overall survival rate of 12% and a median survival following diagnosis ranging from 6 to 20 months [3, 4]. Therefore, timely and accurate diagnosis is very important for treatment of HCC. Currently, the modalities employed in the diagnosis of HCC mainly include cross-sectional imaging, biopsy, and serum AFP, which depend on both the size of the lesion and underlying liver function, and some of them are controversial [5, 6].

Traditional Chinese Medicine (TCM) is one of the most popular complementary and alternative medicine modalities. It plays an active role in diagnosis and treatment of HCC in Chinese and East some Asian countries [7, 8]. Different from other diagnostic methods, it is possible to accurately diagnose HCC using inspection, auscultation and olfaction, inquiry, and pulse taking and palpation [8]. In this study, we will work on a TCM clinical dataset, which is observed from 120 HCC patients. Each patient is observed on 147 clinical symptoms and a positive score is evaluated to indicate total positive strength of symptoms. Based on this TCM dataset, we could achieve two aims: (1) screening the potential clinical syndromes for this cancer and (2) inferring the relationships among the potential clinical features via Bayesian network analysis. However, the computational cost will be exceedingly high if the dimensions of the raw dataset

are large. Furthermore, the causal relationships between all the features are difficult to infer because high dimensional data sharply increases the complexity of Bayesian network structure learning [9].

In this study, a particle swarm optimization-based hierarchical feature selection (PSOHFS) model was proposed to select potential clinical syndromes for HCC diagnoses. Firstly, all the 147 original symptoms were arranged into 27 groups according to the categories of clinical observations, and 27 new syndrome features were generated from these groups. Then, the hierarchical feature representation was built with a tree structure, in which different layers indicate different scales of clinical information (Figure 1). Secondly, an improved PSO algorithm was employed at the syndrome level to search an optimal syndrome subset for diagnoses. The experiment shows that 24 novel syndromes searched by PSOHFS could improve accuracy of diagnosis. In addition, Bayesian networks were further constructed at two levels: (1) a global network on the middle-layer features revealed the relationships among 24 potential syndromes; (2) the local networks were used to represent the connections of symptoms in the same groups.

The rest of the paper is organized as follows. Section 2 introduces the details about the experimental data and the feature selection approach. Sections 1 and 2 present the experiment design and results, respectively. Some important conclusions drawn are presented in Section 5.

2. Materials and Methods

2.1. Experimental Data. In this study, the raw data was observed from 120 HCC patients. The clinical dataset includes 300 samples and 147 clinical symptoms. The levels of positive of each symptom are quantified with nonnegative integers. The larger value indicates stronger positive symptom occurred. There are two types of data range for all the original symptoms: binary or integer. For example, the symptom "lip color is white" is binary (0 or 1); that means there are two possible states for this symptom: occurrence or nonoccurrence. Another example is "abdominal pain"; its data range is 0, 1, 2, and 3. The symptom is not positive if its value equals zero; otherwise, the larger the value is, the stronger positive symptom will be. In addition, each patient is marked with a score (nonnegative value) to represent the total evaluation of positive symptoms on this patient. It is obvious that if the HCC patients have larger positive scores than normal people, it is because some clinical symptoms appeared.

2.2. Feature Selection. Feature selection for classification or regression can be widely organized into three categories, depending on how they interact with the construction of model. Filter methods employ a criterion to evaluate each feature individually and is independent of the model [10]. Among them, feature ranking is a common method which involves ranking all the features based on a certain measurement and selecting a feature subset which contains high-ranked features [11]. Wrapper methods involve combinatorial searches through the feature space, guided by the predicting performance of a classification or regression model [12]. Embedded methods perform feature selection in the process of training a model [13].

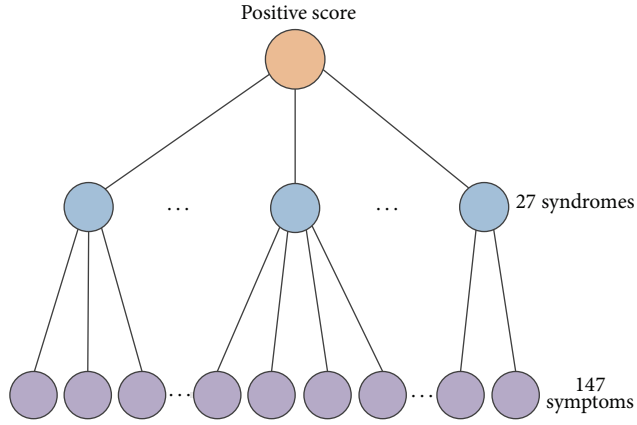


Figure 1: The hierarchical feature representation of TCM clinical dataset.

2.3. Hierarchical Feature Selection. When the raw dataset is high dimensional, the complexity of feature selection may be extremely high: (a) the computational cost will sharply increase, particularly for the wrapper and embedded methods; (b) the potential optimal feature subset may include many irrelevant or redundant features. Therefore, it is necessary to preliminarily reduce the dimension of original feature set before feature selection. As a common preselecting strategy, feature ranking-based approach could quickly reduce the feature space by picking up high-ranked features [14]. However, this type of approach always leads to inclusion of some redundant features. In addition, the optimal feature subset which covers high-ranked features may not provide the best performance in the classification (or regression) model. Ruvolo et al. proposed a novel hierarchical feature selection approach for the audio classification by converting the raw data to three-layer feature representation with a tree structure [15]. All the low-layer features are aggregated into several groups in a "bag of features" manner, and then a higher-layer feature is extracted based on the lower-layer features in the same group. Obviously, the high-layer feature set constitutes a reduced feature space with little redundancy and might provide lower computational cost for classification or regression model.

In this study, our raw TCM data is high dimensional and there are some redundant clinical symptom features included. For example, there are four redundant observed features to describe lip color of patients, such as "lip color is pale," "lip color is red," "lip color is pink," and "lip color is dark purple." Therefore, we aggregate several features into a group if they describe the same category of clinical symptoms or the same part of body and define a new syndrome feature for each symptom group. After extracting all the syndrome features, we build a tree structure to achieve the hierarchical feature representation (Figure 1). In this hierarchical structure, the

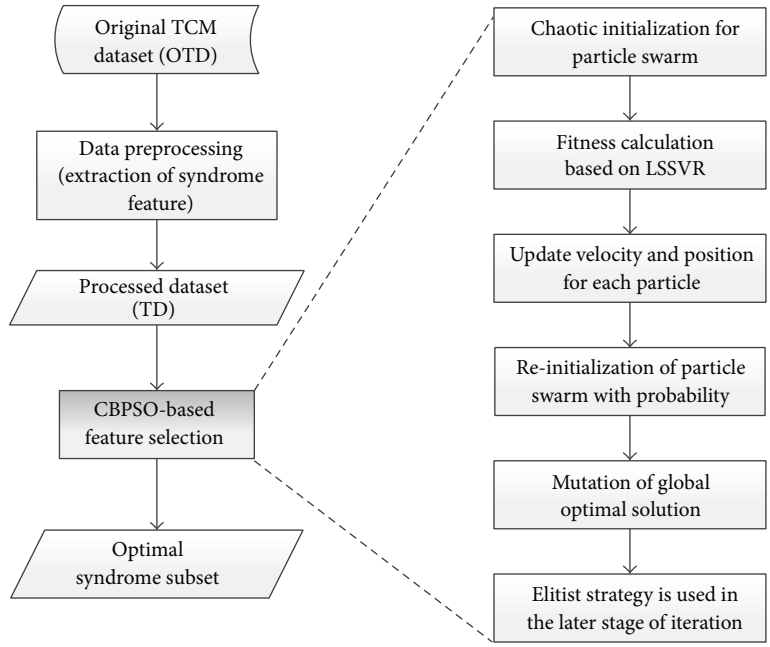


Figure 2: The flow chart of the proposed PSOHFS model for hierarchical feature selection.

bottom-layer nodes (leaf nodes) are the original clinical symptom features which are directly collected from the original TCM clinical dataset. And a middle-layer syndrome feature is defined on a group of symptoms which are related to the same part of the body. If the symptoms in the same group are not mutually exclusive (concurrent), the corresponding syndrome is defined as the sum of all these symptoms; otherwise, the level of positivity of the syndrome is based on the frequency of each symptom in all the patients (see Section 2). The top-layer node is the root of the tree, which denotes the positive score of a patient. It is obvious that each syndrome roughly represents the positive strength of one specific aspect or part of body, while symptom provides much more detailed information. Particularly, our study focuses on how to reasonably extract the syndrome features to generate a reduced feature set for feature selection and infer the causal relationships among these two-layer features.

2.4. Particle Swarm Optimization-Based Hierarchical Feature Selection (PSOHFS). Based on the hierarchical feature representation, the dimension of the processed TCM dataset is sharply reduced on the syndrome level. We designed a chaotic binary particle swarm optimization (CBPSO) algorithm to search potential syndromes for diagnosing efficiently. The flow chart of proposed CBPSO-based feature selection is shown in Figure 2.

Particle swarm optimization (PSO) is a population-based random optimization algorithm [16]. A swarm consists of N particles moving around in a D-dimensional search space. The position of the ith particle is represented as $X_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, and the velocity $V_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$, where $1 \le i \le N$. The positions and velocities of particles are confined within $[X_{\min}, X_{\max}]^D$ and $[V_{\min}, V_{\max}]^D$, respectively. Each particle coexists and evolves simultaneously based on knowledge shared with neighboring particles; it makes use of its own memory and knowledge gained by the swarm as a whole to find the best solution. The best previously encountered position of the ith particle is considered as its individual best position $p\text{best}_i = (p_{i1}, p_{i2}, \ldots, p_{iD})$. The best position of all the $p\text{best}_i$ is considered as the global best position $g\text{best} = (g_1, g_2, \ldots, g_D)$. The limitation of the standard PSO algorithm is applied to optimize the problems in continuous space. However, many optimization problems occur in a discrete feature space; thus binary PSO (BPSO) was proposed to combinatorial optimization [17]. In BPSO, each particle X_i is presented as a binary vector, thus, the overall velocity of particle may be described by the number of bits changed per iteration. Generally, each particle is updated as the following equations:

$$\begin{aligned} v_{id}^{\text{new}} &= w * v_{id}^{\text{old}} + c_1 * r_1 * \left(p\text{best}_{id} - x_{id}^{\text{old}}\right) \\ &\quad + c_2 * r_2 * \left(g\text{best}_d - x_{id}^{\text{old}}\right) \\ &\text{if } v_{id}^{\text{new}} \notin \left(V_{\min}, V_{\max}\right), \text{then} \\ v_{id}^{\text{new}} &= \max\left(\min\left(V_{\max}, v_{id}^{\text{new}}\right), V_{\min}\right) \\ S\left(v_{id}^{\text{new}}\right) &= \frac{1}{\left(1 + e^{-v_{id}^{\text{new}}}\right)} \\ &\text{if rand} < S\left(v_{id}^{\text{new}}\right), \text{then } x_{id}^{\text{new}} = 1; \text{ else } x_{id}^{\text{new}} = 0. \end{aligned} \tag{1}$$

Equation (1) will be used to update the velocities and positions of each particle in each generation. The inertia weight w controls the impact of the previous velocity of a particle on its current one. r_1 and r_2 are random numbers between

$[0, 1]$; c_1 and c_2 are acceleration constants that control how far a particle moves in a single generation. Velocities v_{id}^{new} and v_{id}^{old} denote the dth velocities of the ith particle in the current and the last generations, respectively. x_{id}^{new} and x_{id}^{old} indicate corresponding positions on the dth dimension, respectively. In our case, $V_{\max} = 6$, $V_{\min} = -6$.

Generally, the speed of convergence of BPSO is fast; however, it has high risk of converging to local optimum. Because chaos is a complex behavior of a nonlinear deterministic system which has ergodic and stochastic properties, we combine chaos theory with BPSO to design chaotic binary particle swarm optimization (CBPSO), which potentially promotes the convergence performance of BPSO [18].

CBPSO-based feature selection is introduced in the following steps (Figure 2).

(1) Chaotic Initialization of Particle Swarm. When CBPSO is used for feature selection, each particle indicates a candidate feature subset. Given an original feature set $F = \{f_1, f_2, \ldots, f_D\}$, each particle is denoted by $X_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, where D is the number of features. It is obvious that each particle represents a candidate feature subset. If x_{ij} equals 1 indicates the jth feature is selected; otherwise, is not selected. The performance of convergence about BPSO largely depends on initial particle swarm. The chaotic initialization via globally searching combined the ergodic and stochastic property of chaotic system is often has a better quality than random initialization.

The common chaotic model is logistic model; it can be shown as follows:

$$q_{k+1} = \mu q_k (1 - q_k), \quad k = 0, 1, 2, \ldots. \tag{2}$$

Equation (2) indicates a dynamical system, where μ is a control parameter. Given the value of μ, a time series $q_1, q_2, \ldots, q_k$ is generated from a random initial value q_0, which ranges from 0 to 1. When μ equals 4, there is no stable solution for the dynamic system. It appears as a complete chaotic state.

Now, an initial random vector $X_0 = \{x_{01}, x_{02}, \ldots, x_{0D}\}$ is generated. We substitute each element of X_0 into (2) orderly and iterate k times, respectively, and then obtain D chaotic variables $CX = [x_1, x_2, \ldots, x_D]$, which have different locus. When CX is substituted into (3), we get k binary vectors $[X_1; X_2; \ldots; X_k]$, where the binary vector $X_j = [g(x_{j1}), g(x_{j2}), \ldots, g(x_{jD})]$ represents a particle ($1 \le j \le k$):

$$g(x) = \begin{cases} 1, & x \ge 0.5 \\ 0, & x < 0.5. \end{cases} \tag{3}$$

At last, we select N top binary vectors to constitute initial particle swarm based on the fitness values. For fully traversal of chaotic variable, the iteration of chaotic series is always large (here, $k = 500$, $N < k$).

(2) Fitness Calculation Based on LSSVR. Support vector machine (SVM) has excellent capabilities in classification (SVC) or regression (SVR), even for small sample [19]. It minimizes an upper bound of the generalization error based on the principle of structure risk minimize. However, SVM training process will be time consuming if dataset is huge. Therefore, least squares support vector machine (LSSVM) is proposed to overcome the shortcoming of high computational cost [20]. Generally, LSSVM can be categorized into LSSVR which is used for regression and LSSVC for classification. Because the problem-solving process of the SVR is a QP problem, which will inevitably cause a high computational complexity especially for large-scale QP problem, LSSVR can overcome these shortcomings by a set of linear equations and squared loss function which lead to important reduction in computational complexity [21].

In this study, we use LSSVR as a regression model to evaluate the predicting performance of each candidate feature subset. We assume that an optimal feature subset not only has excellent performance of prediction but also contains more relevant features and less irrelevant features. The fitness function is defined in

$$\text{fitvalue}\ (X_i) = \text{pdterror}(X_i) + p * \text{mfr}(X_i). \tag{4}$$

X_i denotes a particle-coding binary vector which indicates a candidate feature subset. The function $\text{pdterror}(X_i)$ calculates the predicting error of LSSVR model based on the selected features in X_i. The parameter p is a weight between 0 and 1. Function $\text{mfr}(X_i)$ indicates the correlation measure between a feature subset and the target variable. In (5), the function $\text{fr}(f_{ij})$ measures the relevance between feature f_{ij} (included in X_i) and target value via a feature-ranking strategy. In our experiment, the more predictive features have smaller values of $\text{fr}(*)$ (see experiment in Section 3.2). Therefore, the smaller fitness value corresponds to the better candidate feature subset:

$$\text{mfr}(X_i) = \text{mean}(\text{fr}(f_{i1}), \text{fr}(f_{i2}), \ldots, \text{fr}(f_{iM})). \tag{5}$$

(3) Update the Velocity and Position for Each Particle. The velocity and position of each particle are updated according to (1). Considering the searching performance of CBPSO is affected largely by inertia weight (w), the value of w is dynamically updated in our CBPSO by using nonlinear decreasing strategy. Its calculation is as follows:

$$w = wl * \left(\frac{ws}{wl}\right)^{1/(1+c_3*(t/(t\max)))}. \tag{6}$$

In (6), $t\max$ is the number of iterations, t is the current iteration, and c_3 is a constant (set $c_3 = 10$). ws and wl, respectively, are the values of w on the initial and last generation ($ws > wl$). In our case, $ws = 1.2$, $wl = 0.4$. The performance of global search of CBPSO is increased using larger w at the beginning of iteration, and the local search will be enhanced using smaller w at the later stage.

(4) Reinitialization of Particle Swarm with Probability. The trajectory of particle is largely affected by $g$best and all the $p\text{best}_i$. At the beginning of iteration, the convergence rate of swarm is fast, but it is slow at the later stage which has high risk of converging to local optimum. For overcoming this shortcoming, each particle in each generation is reinitialized with small probability (Figure 3).

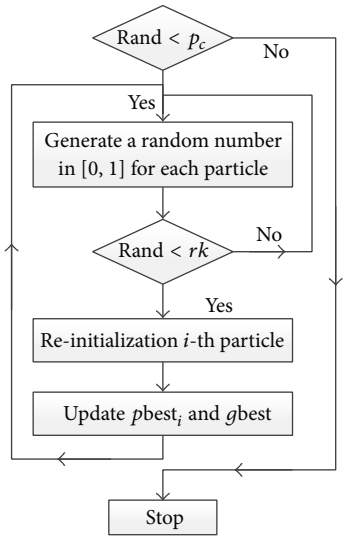


Figure 3: The flow chart of reinitialization of particle swarm.

In Figure 3, p_c is the probability of reinitialization for current particle swarm, with its calculation based on (7). At the early stage of iteration, there are many chances for particles to approximate the optimal solution, so that the probability of reinitialization for whole swarm is small. In the later stage, the probability of reinitialization is increased, it can largely avoid the particles fall into the local optimum. The parameter currun denotes the current iteration, and rk is a small random probability (in our case, $rk = 0.3$). When the better particle is found after reinitialization, update the current $g$best and $p$best$_i$:

$$p_c = 1 - \frac{1}{1 + \ln(\text{currun})}. \tag{7}$$

(5) Mutation of the Potential Global Optimal Solution. If the global optimal particle $g$best is not constantly improved for a long time, it is necessary to make variation for it to jump out from the local optimal point. In our case, when $g$best is invariant in 10 iterations, its binary coding vector will be mutated with a random probability. If a better particle is found, $g$best is updated again.

(6) Elitist Strategy Is Used in the Later Stage of Iteration. If step (4) could not obviously improve the $g$best further, a number of new particles are generated with a probability to instead some particles in current swarm so that the diversity of current swarm could be enhanced [22].

3. Experiment

3.1. Data Preprocessing. For hierarchical representation of clinical symptoms, our raw dataset should be preprocessed as in the following steps. Firstly, we manually divide all the 147 symptoms into 27 groups according to the categories of symptoms (Table SS in Supplementary Material available online at http://dx.doi.org/10.1155/2014/127572). Figure 4(a) shows an example of four clinical symptoms (pale, red, pink, and dark purple) being arranged to a group called "lip color." Hence, a syndrome feature "lip color" simply represents the states of lip color for a patient instead of four redundant symptom features. Secondly, we calculate each syndrome feature which is extracted from the corresponding clinical symptom group. Therefore, we obtain a new reduced feature space at syndrome level. Finally, combining the original symptom features, extracted syndrome features, and the positive score, we build a tree structure for hierarchical feature representation of the TCM clinical data. Two typical examples are given regarding how to extract the syndrome features from the group of symptoms.

Example 1. Figure 4(a) shows an example of several symptoms in the same group being mutually exclusive. That means if the lip color of a patient is red, the rest of the three colors will not appear with him/her. We name a new feature LC with five possible discrete values ($LC = 0, 1, 2, 3, 4$) to simplistically represent the combined meaning of four original symptoms. According to Figure 4(a), the states of lip color for a patient are presented with a binary vector (length is four) in original TCM data, while we can represent it with a single value LC, where $LC \in \{0, 1, 2, 3, 4\}$. If LC equals zero, that means all four symptoms are not positive. Otherwise, one of the symptoms appears positive. As for the mapping between four symptoms and four discrete values (1, 2, 3, and 4), we follow a simple rule to assign each candidate value to a possible level of this symptom: the larger discrete value of LC indicates that much more patients are positive on this clinical symptom. We count the statistic distributions of all the samples on these four symptoms, respectively, and map each discrete value to a symptom of lip color according to the mean value of positive scores on each symptom.

Example 2. The symptoms in the same group are not mutually exclusive. Figure 4(b) shows three clinical symptoms of emotion: irritability, depression, and sigh. These symptoms could be positive simultaneously on a patient. For example, the clinical symptoms of emotion for a patient are denoted by a vector $Es = [2, 0, 1]$ in original data, which means two emotion-related positive symptoms appeared with him/her. In this case, a new syndrome feature NEs is extracted from Es, where $NEs = \text{sum}(Es) = 3$. Therefore, if a patient has several positive symptoms which belong to the same syndrome, cumulative summation is a feasible strategy to get a total positive strength on this syndrome.

3.2. Experiment Design. First, we proposed a feature-ranking strategy for association analysis between individual syndrome and positive score (target value) with function fr(∗):

$$\text{fr}(f_i) = \frac{\text{mcc}(f_i, \text{ps}) + pcv(f_i)}{2} \tag{8}$$

$$\text{mcc}(f_i, \text{ps}) = 1 - \left|\text{corr}(f_i, \text{ps})\right|$$

$$pcv(f_i) = 1 - \frac{\text{pe}(f_i)}{\max\{\text{pe}(f_i), \text{pe}(f_2), \ldots, \text{pe}(f_D)\}}. \tag{9}$$

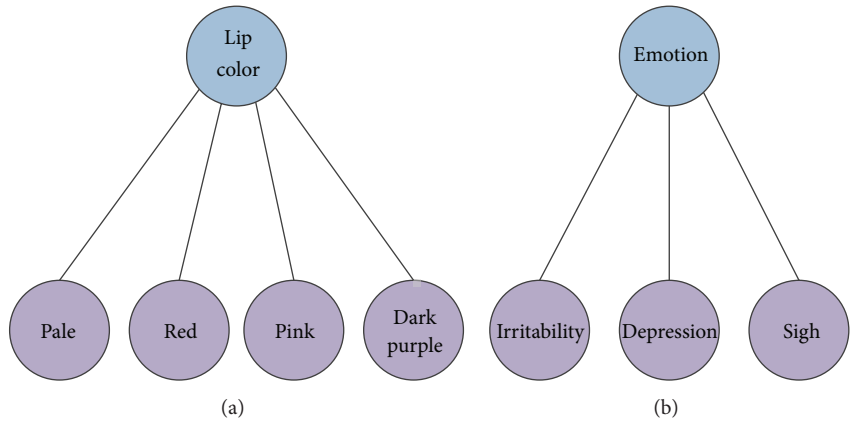


Figure 4: Two groups of symptoms are represented: "lip color" and "emotion." (a) The syndrome feature "lip color" defined on four clinical symptoms which describe four possible positive states of "lip colors." (b) The syndrome feature "emotion" defined on three clinical symptom features which describe three types of emotional states.

Combining ((4)-(5), (8)-(9)), we can determine the fitness function in the proposed PSOHFS model for feature subset optimizing. The function corr(f_i, ps) is the correlation coefficient between feature f_i and target value (ps). Function pe(f_i) denotes the predicting error of LSSVR model with all the features except f_i. If the predicting error is obviously increased after moving out f_i from the whole feature set, it indicates the feature f_i is high predictive. The smaller value of fr(f_i), the higher-ranked feature f_i will be. The result of feature ranking can provide a reference about the importance of each syndrome to positive score.

Secondly, our developed CBPSO algorithm was applied at the syndrome level for feature selection. Different swarm size and the number of iterations were chosen to test the searching performance of the proposed CBPSO. And then, the predicting performance of the optimal syndrome subset (OPS) by proposed model was further validated. On the one hand, we employed two well-established feature selection methods to compare them with our proposed PSOHFS model: (1) correlation-based filter method (CFM) [14, 23] and (2) PSO-based wrapper method (PWM) [14]. These standard approaches were applied on original symptom features. On the other hand, we further validated the performance of OPS by feature ranking on the syndrome feature level. Two types of syndrome subsets were selected to compare: (1) full collection with all the 27 syndromes (FCS) and (2) filter-based syndrome set by feature ranking via (8). Here, we set threshold 0.8 and 0.9 to get two potential syndrome subsets: FRS1 and FRS2.

Finally, based on the optimal potential syndrome subset inferred by our PSOHFS model, Bayesian networks were constructed, respectively, at the symptom and syndrome levels. On the one hand, the global Bayesian network on potential syndromes was inferred using GES algorithm [24]. Such coarser-grained network can roughly reveal the causal relationships among these potential syndromes of this cancer. Before structure learning of global network, the processed TCM dataset (TD) in Section 3.1 should be firstly discretized according to

$$DTD(:, j) = \begin{cases} TD(:, j), & \text{if length}(\text{unique}(TD(:, j))) \le 4 \\ \dfrac{TD(:, j)}{\max(TD(:, j))/\text{itvnum}(TD(:, j))}, & \text{else} \end{cases} \quad (10)$$

$$\text{itvnum}(TD(:, j)) = \lfloor \log 2(\text{length}(\text{unique}(TD(:, j)))) \rfloor + 1.$$

TD(:, j) denotes all the calculated values of jth syndrome. Function itvnum (TD(:, j)) is used to estimate the optimal intervals of discretization for the sample of jth syndrome. If the number of positive levels for a syndrome is larger than four, the discretization is necessary on this syndrome. On the other hand, we chose three syndromes as examples to construct local networks using GES algorithm (Table 4). When a network structure is learned, Maximum Likelihood Estimation (MLE) is utilized to compute all the conditional probability tables. Then, the probability inference could be achieved using inference algorithm, such as junction tree method [25, 26].

3.3. Experimental Parameters. The simulating experiments were implemented under the environment of MATLAB2011a with Intel Core i5-2410 CPU @ 2.3GHZ, 4 GB RAM. In the LSSVR regression model, Gaussian RBF kernel is employed, and the kernel parameters σ^2 and γ should be determined firstly. Currently, many approaches have been applied in parameter optimization of LSSVR, such as grid search [27], cross-validation [28, 29], genetic algorithm (GA) [30], and simulated annealing algorithm [31]. In our study, grid search was selected to determine the parameters in the range

Table 1: The result of feature ranking for all the syndromes.

Syndrome (f_i)	Name of syndrome	Abbreviation	Size	mcc(f_i, ps)	pe(f_i)	fr(f_i)	Rank
1	Lip color	LC	4	0.9257	0.1688	0.9480	24
2	Tongue color	Tc	4	0.9293	0.1813	0.9487	25
3	Appearance of tongue-1	At1	3	0.8123	0.5808	0.8550	16
4	Appearance of tongue-2	At2	5	0.9712	0.2998	0.9592	27
5	Coated tongue color	Ctc	3	0.8589	0.1914	0.9126	21
6	Texture of coated tongue	Tct	7	0.9039	0.2518	0.9298	23
7	Position of coated tongue	Pct	5	0.9629	0.2685	0.9578	26
8	The color of complexion	Coc	8	0.6396	2.7350	0.5790	6
9	Whole body condition	Wbc	8	0.9326	1.0378	0.8749	19
10	Odor	Od	1	0.6948	0.6055	0.7941	13
11	Chilly	Ch	1	0.6011	0.4767	0.7586	10
12	Hectic fever	Hf	1	0.7890	0.4248	0.8571	17
13	Fever	Fe	1	0.7304	0.2969	0.8391	15
14	Sweating	St	2	0.6270	0.4875	0.7706	11
15	Facial features	Ff	13	0.2177	5.6792	0.1088	1
16	Cardiothoracic condition	Ca	4	0.4923	1.2036	0.6402	8
17	Sternocostal and abdominal pain	Sap	16	0.4010	1.6943	0.5513	5
18	Diet	Diet	7	0.2937	1.7266	0.4948	3
19	Defecate and urine	Du	10	0.4016	1.7268	0.5488	4
20	Sleep	Slp	2	0.4382	0.9854	0.6324	7
21	Emotion	NEs	3	0.5141	1.1600	0.6549	9
22	Skin of the limbs	Sl	10	0.2091	2.4312	0.3905	2
23	Bump in ribs	Bir	1	0.6543	0.5541	0.7784	12
24	Ascites	Ass	1	0.7279	0.4304	0.8260	14
25	Pleural effusion	Pe	1	0.7894	0.2301	0.8745	18
26	Pulse condition in left	Pcle	13	0.8630	0.3003	0.9051	20
27	Pulse condition in right	Pcrt	13	0.8716	0.2556	0.9133	22

of [0.1, 100000] for σ^2 and [0.1, 10000] for γ. For a pairwise (σ^2, γ), we used 10-fold cross-validation to evaluate the performance of LSSVR model.

To evaluate the accuracy of prediction, three statistical metrics are widely employed: (1) mean square error (MSE), (2) root mean square error (RMSE), and (3) mean relative percentage error (MRPE). In (11), where y_i and y_i' are the observed value and predicted value, the smaller MSE, RMSE, and MRPE are, the better the LSSVR model will be:

$$\begin{aligned} \text{MSE} &= \frac{1}{n}\sum_{i=1}^{n}\left[y_i - y_i'\right]^2 \\ \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[y_i - y_i'\right]^2} \\ \text{MRPE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i' - y_i}{y_i}\right| \times 100\%. \end{aligned} \tag{11}$$

In our experiment, we used MSE to calculate the values of function pdterror$(*)$ and pe$(*)$.

Moreover, the Matlab Bayes Net Toolbox FullBNT-1.0.7 [32] and BNT Structure Learning Package BNT_SLP_1.5 were, respectively, used in the Bayesian network structure learning, parameters learning, and probability inference. The probability distribution between nodes in a Bayesian network could be computed according to the inferred network structure and conditional probability tables.

4. Results and Discussion

Table 1 shows the results of association analysis between individual syndromes and positive score. mcc(f_i, ps) reflects the predicting performance of feature f_i to ps (positive score). The smaller the value of mcc is, the more important the feature f_i will be. The value of pe(f_i) indicates predicting error of LSSVR model based on all the features except f_i; it is measured by MSE. Here, it is obvious that the higher-ranked features have lower values of fr(f_i). We clearly see some important syndromes are high predictive, such as "facial features," "skin of the limbs," "diet," "sternocostal and abdominal pain," and so forth.

Our developed CBPSO algorithm was applied to search the optimal syndrome subset on the processed TCM dataset. Assigning different swarm size and the number of iterations, this CBPSO algorithm shows excellent convergence performance (Figure 5). Different assignments of parameters for CBPSO finally got the same optimal solution:

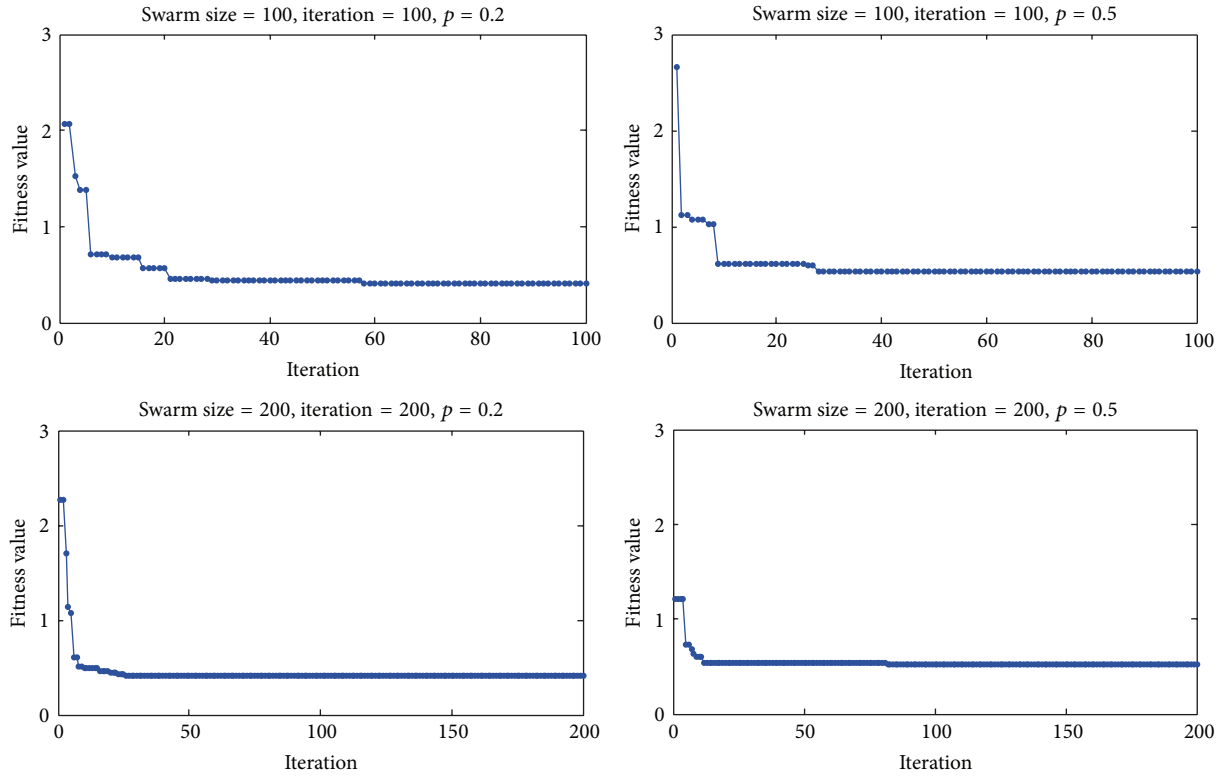


Figure 5: The results of CBPSO-based feature selection under different parameters. Four subfigures show the CBPSO algorithm rapidly approximate the optimal solution in the reduced feature space.

Table 2: The optimal solutions of our CBPSO using different parameters.

Swarm size	Iteration	P	The optimal solution of CBPSO	Fitness value
100	100	0.2	00110111111111111111111111111	0.40911
100	100	0.5	00110111111111111111111111111	0.53205
200	200	0.2	00110111111111111111111111111	0.41062
200	200	0.5	00110111111111111111111111111	0.52183

00110111111111111111111111111. It means the potential syndrome subset containing 24 syndromes is a steady solution for this NP-hard problem (Table 2). These 24 syndromes reflect many cancer-related parts of body or aspects of observation, which are helpful to clinically diagnose HCC.

Now, two well-established feature selection methods were introduced to be compared with our proposed PSOHFS model: (1) correlation-based filter method (CFM) [14, 23] and (2) PSO-based wrapper method (PWM) [14]. The first one uses correlation-based feature ranking as the principle criteria for feature selection by ordering. The second one uses standard BPSO algorithm to search an optimal feature subset. These two methods were all applied on the original symptom features. For CFM, we used 15% and 30% top-ranked features to validate its performance, while, for PWM, we set population size equal to 100 and iterations equal to 100 and 200. Table 3 shows the error of prediction of the LSSVR model based on candidate optimal feature subsets. Five candidate feature subsets were searched by the above two methods and PSOHFS model, respectively. In Table 3, the values of MSE, MRSE, and MRPE were calculated based on LSSVR by 5-fold cross-validation.

Comparing the values of MSE, RMSE, and MRPE in Table 3, we can see that the optimal syndrome set (OPS) searched by our PSOHFS model has the obvious superiority in the predicting performance. The dimension of the PSOHFS-based optimal syndrome subset equals 24, which is significantly smaller relatively to the dimension of the original symptoms (147). Because CFM and PWM work directly on the original high dimensional feature space, it is hard for them to achieve an optimized prediction performance and the dimension of potential feature subset, simultaneously. PWM searches for the optimal solution depending on the evaluation of regression model, so the optimal feature subset from PWM is more predictive than CFM's. However, standard wrapper-based methods do not

Table 3: The predicting performance of the optimal feature subsets obtained from different feature selection methods.

Approaches	Dimension of the optimal feature subset	MSE	RMSE	MRPE (%)	Time (second)
PSOHFS	**24 (syndromes)**	**0.1622**	**0.4027**	**1.0700**	**3.0108**
CFM (top 15%)	22 (symptoms)	14.4575	3.8023	11.8907	2.8510
CFM (top 30%)	45 (symptoms)	6.2611	2.5022	7.8632	4.8010
PWM (100 iterations)	92 (symptoms)	3.2268	1.7963	5.5645	8.8760
PWM (200 iterations)	89 (symptoms)	2.7516	1.6588	5.2351	8.7390

Table 4: Comparisons of the PSOHFS-based optimal syndrome set with other potential syndrome subsets.

Feature set	Dimension	MSE	RMSE	MRPE (%)	Time (second)
OPS	**24**	**0.1622**	**0.4027**	**1.0700**	**3.0108**
FCS	27	0.1834	0.4283	1.9572	3.2604
FRS1	13	3.3735	1.8367	6.2871	2.4024
FRS2	19	1.7084	1.3071	4.5202	2.9640

optimize the size of optimal feature subset. CFM got the worst result is reasonable because the correlation measurement can only detect linear dependencies between variable and target.

Next, we further validate the performance of OPS on the syndrome level. Two types of syndrome subsets were selected to compare: (1) full collection with all the 27 syndromes (FCS) and (2) filter-based syndrome subset by feature ranking via (8). Here, we chose threshold 0.8 and 0.9 to get two potential syndrome subsets: FRS1 and FRS2 (Table 1). In Table 4, we obviously find OPS can get good balance between the dimension and predicting performance. The verification on FRS1 and FRS2 proves the fact that, although feature-ranking methods run quickly, they still easily lead to worse results because feature-ranking filter ignores the possible interactions and dependences among the features [29]. The difference between Tables 3 and 4 indicates the feature selection on a reduced feature space of original dataset potentially obtains a better solution. 24 potential syndrome features could quickly diagnose the positive level of HCC patients with high accuracy. Our result suggested that "lip color," "tongue color," and "coated tongue color" could be ignored during the process of prediction because they are weak predictive features for discriminating these HCC samples.

Finally, based on the hierarchical feature representation and the result of feature selection on syndromes, Bayesian network on two layers was constructed and the conditional probability tables were inferred. Here, we picked up three cases to explain what we can obtain from the Bayesian network analysis in the symptom and syndrome feature space (Table 5). Figure 6(a) shows the Bayesian network structure of "emotion" syndrome. We can clearly see that there is a causal relationship between "depression" and "sigh." When a patient is depressive, sigh is a usual symptom with him/her. While "irritability" seems to reflect inversely comparing to "depression"; therefore it is an independent node in this inferred network structure. The conditional probability tables

Table 5: The details of three syndromes.

Syndrome	Symptoms	The number of level of positive symptom
Emotion	Irritability	4
	Depression	3
	Sigh	3
Cardiothoracic condition	Tightness in the chest	4
	Shortness of breath	3
	Palpitations	3
	Pain in the chest	3
Diet	Anorexia	4
	Tired of greasy	4
	Nausea	3
	Hiccups	3
	Acid reflux	3
	Water reflux	3
	Gastric discomfort	2

of "emotion" are shown as in Supplementary Table S1A-S1C. For example, P("irritability" = 0, "depression" = 1, "sigh" = 1) = 0.027 suggests the probability of the clinical symptoms "depression" and "sigh" is positive on a patient. Figure 6(b) shows the network structure of "cardiothoracic condition" syndrome. From Figure 6(b), "tightness in the chest" might lead to three other clinical symptoms: "shortness of breath," "palpitations," and "pain in chest." The conditional probability tables of "cardiothoracic condition" are shown in Supplementary Table S2A-S2D. For example, P("tightness in the chest" = 1, "shortness of breath" = 1, "palpitations" = 1, "pain in chest" = 0) = 0.01143. Similarly, Figure 6(c) shows the network structure of "diet" syndrome. The conditional probability tables of "diet" are shown in Supplementary Table S3A-S3G. At last, Figure 7 represents the global network on 24 potential syndromes. There are three subnetwork modules and six independent nodes in Figure 7. All the relationships among these syndromes were represented. Their conditional probability tables were listed in Supplementary Table SS1-SS24. Based on the hierarchical feature representation, the Bayesian networks potentially provided us with useful knowledge with multi-granularity. From Table 6, we can clearly see that the computational cost of network structure learning is sharply increased when the number of nodes in the network is increasing. It further proves that if we construct Bayesian network on 147 original clinical symptoms directly, it will

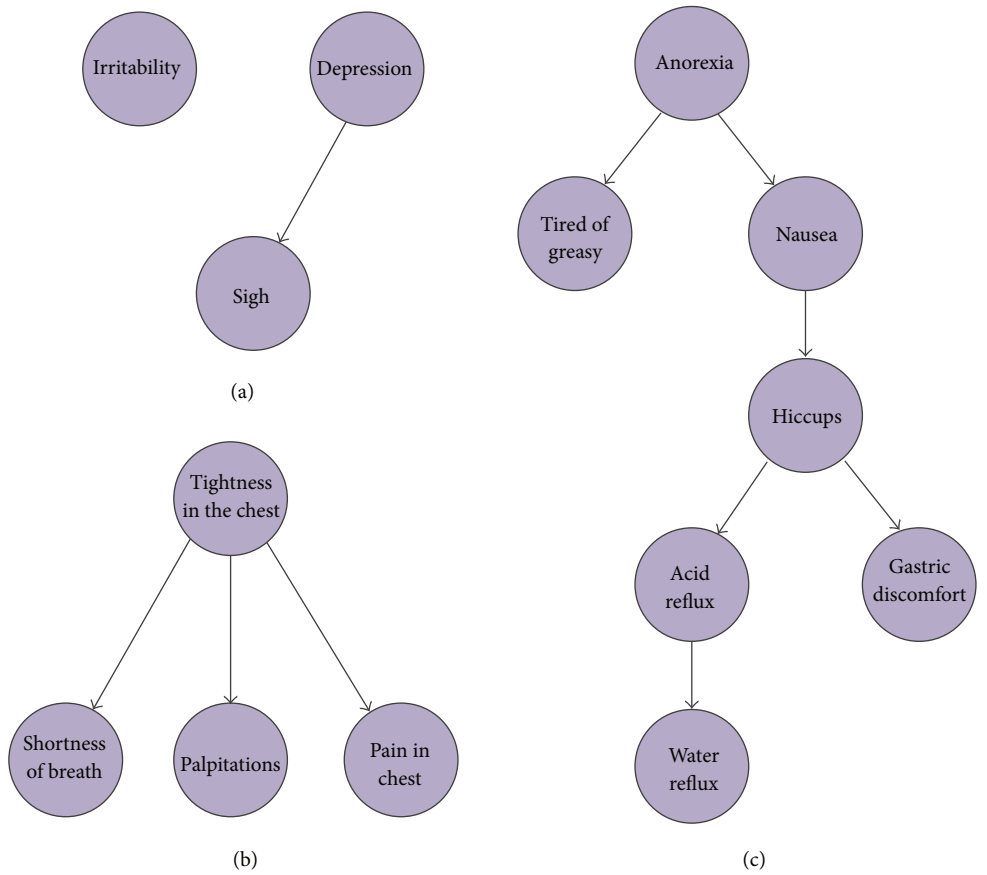


Figure 6: Three inferred Bayesian networks based on symptom features. (a) The casual relationships among three clinical symptoms of "emotion" group. "Depression" might cause "sigh," while "irritability" is an isolated node. (b) The casual relationships among four clinical symptoms of "cardiothoracic condition" group. (c) The casual relationships among seven clinical symptoms of "diet."

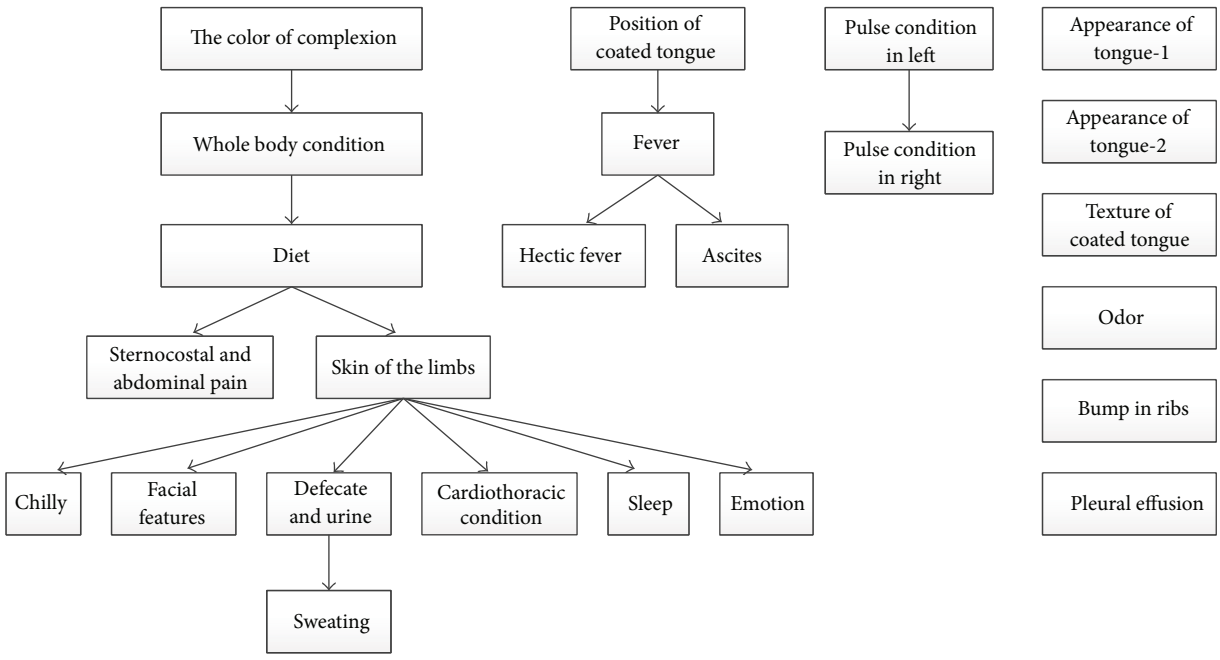


Figure 7: The global Bayesian network based on 24 potential syndromes.

Table 6: The computational cost of structure learning for some Bayesian networks.

Bayesian networks	The number of nodes	Computational time of structure learning (second)
Emotion	3	0.06
Cardiothoracic condition	4	0.41
Diet	7	4.23
Potential syndrome set (OPS)	24	606.88

meet unimaginable computational complex; therefore, our method proposed in this paper provided a good solution.

5. Conclusions

In this paper, a particle swarm optimization-based hierarchical feature selection (PSOHFS) model was proposed to infer potential clinical features of HCC on a Traditional Chinese Medicine dataset which was collected from 120 patients. The PSOHFS model firstly arranged all the 147 original symptoms into 27 groups according to the categories of clinical symptoms and extracted a new syndrome feature from each group. The raw TCM clinical dataset was represented in a reduced feature space so that we can build a hierarchical feature representation pattern with a tree structure. Based on such hierarchical feature graph, we reached two aims: (1) based on a significant reduced feature space, the feature selection can be easily realized, and the optimal feature subset could diagnose patient samples efficiently; (2) we constructed Bayesian network on symptom and syndrome levels. A global Bayesian network for all the potential syndromes roughly described the relationships among the main important aspects of HCC. While each local network was constructed for the symptom features in the same group, the causal relationships among them could be inferred.

In our simulating experiment, our CBPSO algorithm in PSOHFS model discovered an optimal syndrome subset of HCC, which included 24 syndromes. With a LSSVR regression model built by these 24 potential syndromes, the diagnosis accuracy of HCC is high and computational cost is sharply reduced. The significance of the proposed model is as follows: (1) feature selection is implemented on a reduced feature space, so that the dimension of optimal feature subset is smaller; (2) the fitness function in CBPSO algorithm optimizes the predicting performance and the correlation between features and target variable. Based on the results of feature selection, we further achieved the Bayesian network construction at both syndrome and symptom levels to explain the relationships among all the nodes and the probability inference could be computed based on learned network structure and conditional probability tables.

However, our model also has some shortcomings: (1) most of syndrome groups were aggregated from the clinical symptoms observed from the same parts of body, while much more evidence proved that there are significant relationships between symptoms which describe different parts (aspects) of body; (2) we did not study the relationships of clinical symptom features which belong to different groups. In the future, we will collect more clinical samples of HCC to deeply analyze the correlation between any clinical features. Also, some high-predictive clinical features inferred in this study need to be validated further in other TCM clinical datasets. If we can discover and validate some high-predictive clinical features in the next step of research, that might be the significant phenotypes of this cancer.

Conflict of Interests

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

Acknowledgments

This work was supported by the National Science Foundation of China (nos. 61272269 and 61133010). The data in this work was collected by the Changhai Hospital in Shanghai, China. The authors give special thanks to Professor X. Q. Yue for his work in data preprocessing.

References

[1] F. X. Bosch, J. Ribes, R. Cleries, and M. Diaz, "Epidemiology of hepatocellular carcinoma," *Clinics in Liver Disease*, vol. 9, no. 2, pp. 191–211, 2005.

[2] 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.

[3] H. B. El-Serag, "Hepatocellular carcinoma," *The New England Journal of Medicine*, vol. 365, no. 12, pp. 1118–1127, 2011.

[4] C. Gallo, "A new prognostic system for hepatocellular carcinoma: a retrospective study of 435 patients: the Cancer of the Liver Italian Program (CLIP) investigators," *Hepatology*, vol. 28, no. 3, pp. 751–755, 1998.

[5] G. Miller, L. H. Schwartz, and M. D'Angelica, "The use of imaging in the diagnosis and staging of hepatobiliary malignancies," *Surgical Oncology Clinics of North America*, vol. 16, no. 2, pp. 343–368, 2007.

[6] A. Forner, R. Vilana, C. Ayuso et al., "Diagnosis of hepatic nodules 20 mm or smaller in cirrhosis: prospective validation of the noninvasive diagnostic criteria for hepatocellular carcinoma," *Hepatology*, vol. 47, no. 1, pp. 97–104, 2008.

[7] Y. H. Liao, C. C. Lin, T. C. Li, and J. G. Lin, "Utilization pattern of traditional Chinese medicine for liver cancer patients in Taiwan," *BMC Complementary & Alternative Medicine*, vol. 12, article 146, 2012.

[8] R. Mourad, C. Sinoquet, and P. Leray, "Probabilistic graphical models for genetic association studies," *Briefings in Bioinformatics*, vol. 13, no. 1, pp. 20–33, 2012.

[9] X.-W. Chen, G. Anantha, and X. T. Lin, "Improving Bayesian network structure learning with mutual information-based node ordering in the K2 algorithm," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 5, pp. 628–640, 2008.

[10] A. Sharma, S. Imoto, and S. Miyano, "A filter based feature selection algorithm using null space of covariance matrix for

DNA microarray gene expression data," *Current Bioinformatics*, vol. 7, no. 3, pp. 289–294, 2012.

[11] F. Bellal, H. Elghazel, and A. Aussem, "A semi-supervised feature ranking method with ensemble learning," *Pattern Recognition Letters*, vol. 33, no. 10, pp. 1426–1433, 2012.

[12] H. W. Chang, Y. H. Chiu, H. Y. Kao, C. H. Yang, and W. H. Ho, "Comparison of classification algorithms with wrapper-based feature selection for predicting osteoporosis outcome based on genetic factors in a taiwanese women population," *International Journal of Endocrinology*, vol. 2013, Article ID 850735, 8 pages, 2013.

[13] M. B. Imani, M. R. Keyvanpour, and R. Azmi, "A novel embedded feature selection method: a comparative study in the application of text categorization," *Applied Artificial Intelligence*, vol. 27, no. 5, pp. 408–427, 2013.

[14] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[15] P. Ruvolo, I. Fasel, and J. R. Movellan, "A learning approach to hierarchical feature selection and aggregation for audio classification," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1535–1542, 2010.

[16] A. R. Jordehi and J. Jasni, "Parameter selection in particle swarm optimisation: a survey," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 25, no. 4, pp. 527–542, 2013.

[17] A. H. El-Maleh, A. T. Sheikh, and S. M. Sait, "Binary particle swarm optimization (BPSO) based state assignment for area minimization of sequential circuits," *Applied Soft Computing*, vol. 13, no. 12, pp. 4832–4840, 2013.

[18] Q. Zhao and S. Z. Yan, "Collision-free path planning for mobile robots using chaotic particle swarm optimization," in *Proceedings of the 1st International Conference on Natural Computation (ICNC '05)*, vol. 3612, part 3 of *Lecture Notes in Computer Science*, pp. 632–635, Changsha, China, August 2005.

[19] W. Guan and A. Gray, "Sparse high-dimensional fractional-norm support vector machine via DC programming," *Computational Statistics & Data Analysis*, vol. 67, pp. 136–148, 2013.

[20] A. Mellit, A. M. Pavan, and M. Benghanem, "Least squares support vector machine for short-term prediction of meteorological time series," *Theoretical and Applied Climatology*, vol. 111, no. 1-2, pp. 297–307, 2013.

[21] G. Xie, S. Y. Wang, Y. X. Zhao, and K. K. Lai, "Hybrid approaches based on LSSVR model for container throughput forecasting: a comparative study," *Applied Soft Computing*, vol. 13, no. 5, pp. 2232–2241, 2013.

[22] I. J. Leno, S. S. Sankar, M. V. Raj, and S. G. Ponnambalam, "An elitist strategy genetic algorithm for integrated layout design," *The International Journal of Advanced Manufacturing Technology*, vol. 66, no. 9 –12, pp. 1573–1589, 2013.

[23] J. Z. Wang, L. S. Wu, J. Kong, Y. X. Li, and B. X. Zhang, "Maximum weight and minimum redundancy: a novel framework for feature subset selection," *Pattern Recognition*, vol. 46, no. 6, pp. 1616–1627, 2013.

[24] D. M. Chickering, "Learning equivalence classes of Bayesian-network structures," *Journal of Machine Learning Research*, vol. 2, no. 3, pp. 445–498, 2002.

[25] R. G. Cowell, "Local propagation in conditional Gaussian Bayesian networks," *Journal of Machine Learning Research*, vol. 6, pp. 1517–1550, 2005.

[26] M. M. Zhu, S. Y. Liu, Y. L. Yang, and K. Liu, "Using junction trees for structural learning of Bayesian networks," *Journal of Systems Engineering and Electronics*, vol. 23, no. 2, pp. 286–292, 2012.

[27] L. F. Bo, L. Wang, and L. C. Jiao, "Multiple parameter selection for LS-SVM using smooth leave-one-out error," in *Proceedings of the 2nd International Symposium on Neural Networks: Advances in Neural Networks (ISNN '05)*, vol. 3496, part 1 of *Lecture Notes in Computer Science*, pp. 851–856, Chongqing, China, June 2005.

[28] S. J. An, W. Q. Liu, and S. Venkatesh, "Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression," *Pattern Recognition*, vol. 40, no. 8, pp. 2154–2162, 2007.

[29] G. Rubio, H. Pomares, I. Rojas, and L. J. Herrera, "A heuristic method for parameter selection in LS-SVM: application to time series prediction," *International Journal of Forecasting*, vol. 27, no. 3, pp. 725–739, 2011.

[30] Z. Yang, X. S. Gu, X. Y. Liang, and L. C. Ling, "Genetic algorithm-least squares support vector regression based predicting and optimizing model on carbon fiber composite integrated conductivity," *Materials & Design*, vol. 31, no. 3, pp. 1042–1049, 2010.

[31] Y. L. Liu, L. Tao, J. J. Lu, S. Xu, Q. Ma, and Q. Duan, "A novel force field parameter optimization method based on LSSVR for ECEPP," *FEBS Letters*, vol. 585, no. 6, pp. 888–892, 2011.

[32] Y. H. Zhang, W. S. Zhang, and Y. Xie, "Improved heuristic equivalent search algorithm based on maximal information coefficient for Bayesian network structure learning," *Neurocomputing*, vol. 117, pp. 186–195, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 362738, 8 pages
http://dx.doi.org/10.1155/2014/362738

Review Article

A Survey on Evolutionary Algorithm Based Hybrid Intelligence in Bioinformatics

Shan Li,[1] Liying Kang,[1] and Xing-Ming Zhao[2]

[1] *Department of Mathematics, Shanghai University, Shanghai 200444, China*
[2] *Department of Computer Science, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China*

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

Received 3 December 2013; Revised 29 January 2014; Accepted 29 January 2014; Published 6 March 2014

Academic Editor: Jean X. Gao

With the rapid advance in genomics, proteomics, metabolomics, and other types of omics technologies during the past decades, a tremendous amount of data related to molecular biology has been produced. It is becoming a big challenge for the bioinformatists to analyze and interpret these data with conventional intelligent techniques, for example, support vector machines. Recently, the hybrid intelligent methods, which integrate several standard intelligent approaches, are becoming more and more popular due to their robustness and efficiency. Specifically, the hybrid intelligent approaches based on evolutionary algorithms (EAs) are widely used in various fields due to the efficiency and robustness of EAs. In this review, we give an introduction about the applications of hybrid intelligent methods, in particular those based on evolutionary algorithm, in bioinformatics. In particular, we focus on their applications to three common problems that arise in bioinformatics, that is, feature selection, parameter estimation, and reconstruction of biological networks.

1. Introduction

During the past decade, large amounts of biological data have been generated thanks to the development of high-throughput technologies. For example, 1,010,482 samples were profiled and deposited in Gene Expression Omnibus (GEO) database [1] by the writing of this paper, where around thousands of genes on average were measured for each sample. The recently released pilot data from the 1000 genomes project indicate that there are 38 million SNPs (single-nucleotide polymorphism) and 1.4 million biallelic indels within the 14 populations investigated [2]. Beyond that, other large-scale omics data, for example, RNA sequencing and proteomics data, can be found in public databases and are being generated everyday around the world. Despite the invaluable knowledge hidden in the data, unfortunately, the analysis and interpretation of these data lag far behind data generation.

It has been a long history that intelligent methods from artificial intelligence were widely used in bioinformatics, where these approaches were utilized to analyze and interpret the big datasets that cannot be handled by biologists. For example, in their pioneering work, Golub et al. utilized self-organizing maps (SOMs) to discriminate acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL) based only on gene expression profiles without any prior knowledge [3]. Later, support vector machine was employed to classify 14 tumor types based on microarray gene expression data [4]. Except for diagnosis, intelligent methods have been exploited to identify biomarkers [5], annotate gene functions [6], predict drug targets [7, 8], and reverse engineering signaling pathways [9], among others.

Despite the success achieved by standard intelligent methods, it is becoming evident that it is intractable to analyze the large-scale omics data with only single standard intelligent approaches. For example, when diagnosing cancers based on gene expression profiles, low accuracy is expected if a traditional classifier, for example, linear discriminant analysis (LDA), is employed to classify the samples based on all the genes measured. This phenomenon is caused due to the "large p small n" paradigm which arises in microarray data, where there are generally around 20 thousand of genes or variables that were measured for each sample while only tens or at most hundreds of samples were

considered in each experiment. In other words, there are very few samples while a much larger number of variables are to be learned by the intelligent methods, that is, the curse of dimensionality problem. Therefore, it is necessary to employ other intelligent techniques to select a small number of informative features first, based on which a classifier can be constructed to achieve the desired prediction accuracy. Such hybrid intelligent methods, that is, the combination of several traditional intelligent approaches, are being proved useful in analyzing the big complex biological data and are therefore becoming more and more popular.

In this paper, we survey the applications of hybrid intelligent methods in bioinformatics, which can help the researchers from both fields to understand each other and boost their future collaborations. In particular, we focus on the hybrid methods based on evolutionary algorithm due to its popularity in bioinformatics. We introduce the applications of hybrid intelligent methods to three common problems that arise in bioinformatics, that is, feature selection, parameter estimation, and molecular network/pathway reconstruction.

2. Evolutionary Algorithm

In this section, we first briefly introduced evolutionary algorithm, which is actually a family of algorithms inspired by the evolutionary principles in nature. In the evolutionary algorithm family, there are various variants, such as genetic algorithm (GA) [10, 11], genetic programming (GP) [12], evolutionary strategies (ES) [13], evolutionary programming (EP) [14], and differential evolution (DE) [15]. However, the principle underlying all these algorithms is the same that tries to find the optimal solutions by the operations of reproduction, mutation, recombination, and natural selection on a population of candidate solutions. In the following parts, we will take genetic algorithm (GA) as an example to introduce the evolutionary algorithm.

Figure 1 presents a schematic flowchart of genetic algorithm. In genetic algorithm, each candidate solution should be represented in an appropriate way that can be handled by the algorithm. For example, given a pool of candidate solutions X of size M, $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}^T$, a candidate solution $\mathbf{x}_i$, that is, an individual, can be represented as a binary string $\mathbf{x}_i = [0, 0, 1, 0, \ldots, 1]$. Take feature selection as an example; each individual represents a set of features to be selected, where element 1 in the individual means that the corresponding feature is selected and vice versa. After the representation of individuals is determined, a pool of initial solutions is generally randomly generated first.

To evaluate each individual in the candidate solution pool, a fitness function or evaluation function F is defined in the algorithm. The fitness function is generally defined by taking into account the domain knowledge and the optimal objective function to be solved. For instance, the prediction accuracy or classification error can be used as fitness function. If an individual leads to better fitness, it is a better solution and vice versa.

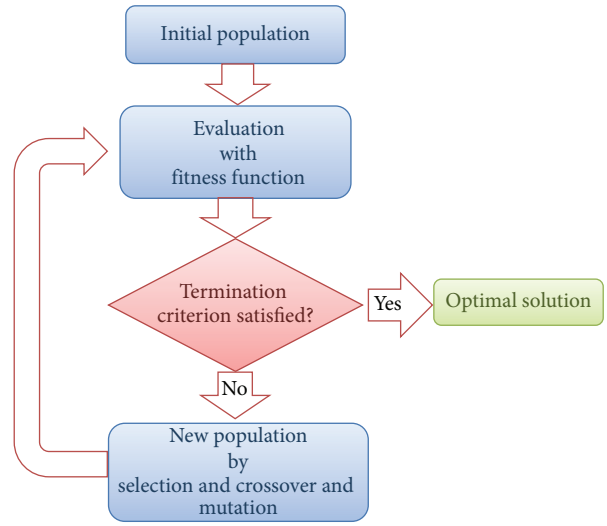


Figure 1: The schematic flowchart of genetic algorithm.

Once the fitness function is determined, the current population will go through two steps: selection and crossover and mutation. In selection step, a subset of individual solutions will be selected generally based on certain probability, and the selected solutions will be used as parents to breed next generation. In the next step, a pair of parent solutions will be picked from the selected parents to generate a new solution with crossover operation; meanwhile, mutation(s) can be optionally applied to certain element(s) within a parent individual to generate a new one. The procedure of crossover and/or mutation continues until a new population of solutions of similar size is generated.

The genetic algorithm repeats the above procedure until certain criterion is met; that is, the preset optimal fitness is found or a fixed number of generations are reached. Despite the common principles underlying the evolutionary algorithm family, other variants of the algorithm may have implementation procedures that are different from the genetic algorithm. For example, in differential evolution, the individuals are selected based on greedy criterion to make sure that all individuals in the new generation are better than or at least as good as the corresponding ones in current population. Another alternative of the traditional genetic algorithm, namely, memetic algorithm (MA), utilizes a local search technique to improve the fitness of each individual and reduce the risk of premature convergence.

Since the evolutionary algorithm starts with a set of random candidate solutions and evaluates multiple individuals at the same time, the risk of getting stuck in a local optimum is reduced. Furthermore, the evolutionary algorithm can generally find optimal solutions within reasonable time, thereby becoming a popular technique in various fields.

3. Feature Selection in Bioinformatics

In bioinformatics, various problems are equivalent to feature selection problem. For example, in bioinformatics, biomarker

discovery is one important and popular topic that tries to identify certain markers, for example, genes or mutations, which can be used for disease diagnosis. It is obvious that biomarker identification is equivalent to feature selection if we consider genes or mutations of interest as variables, where the informative genes or mutations are generally picked to discriminate disease samples from normal ones. However, it is not an easy task to select a few informative variables (generally <20) from thousands or even tens of thousands of features. Under the circumstances, the evolutionary algorithm has been widely adopted for identifying biomarkers along with other intelligent methods. Figure 2 depicts the procedure of feature selection with GA, where GA generally works together with a classifier as a wrapper method and the classifier is used to evaluate the selected features in each iteration. For example, Li et al. [16] utilized genetic algorithm and k-nearest neighbor (KNN) classifier to find discriminative genes that can separate tumors from normal samples based on gene expression data, and robust results were obtained by the hybrid GA/KNN method. Later, Jirapech-Umpai and Aitken [17] applied the GA/KNN approach to leukemia and NCI60 datasets, where the prediction results by the hybrid method are found to be consistent with clinical knowledge, indicating the effectiveness of the hybrid method. Since the simple genetic algorithm (SGA) often converges to a point in the search space, Goldberg and Holland adopted the speciated genetic algorithm, which controls the selection step by handling its fitness with the niching pressure, for gene selection along with artificial neural network (SGANN) [18]. Benchmark results show that SGANN reduces much more features than SGA and performs pretty well [19]. Recently, the hybrid approaches that, respectively, combined Pearson's correlation coefficient (CC) and Relief-F measures with GA were proposed by Chang et al. [20] to select the key features in oral cancer prognosis. These hybrid approaches outperform other popular techniques, such as adaptive neurofuzzy inference system (ANFIS), artificial neural network (ANN), and support vector machine (SVM). In addition to gene selection, the hybrid methods involving evolutionary algorithm have been successfully used to identify SNPs associated with diseases [21, 22] and peptides related to diseases from proteomic profiles [23–25].

Beyond biomarker identification, the evolutionary algorithm based hybrid intelligent methods have also been successfully applied to other feature selection problems in bioinformatics. For example, Zhao et al. [26] proposed a novel hybrid method based on GA and support vector machine (SVM) to select informative features from motif content and protein composition for protein classification, where the principal component analysis (PCA) was further used to reduce the dimensionality while GA was utilized to select a subset of features as well as optimize the regularization parameters of SVM at the same time. Results on benchmark datasets show that the hybrid method is really effective and robust. The hybrid method that integrates SVM and GA was also successfully used to select SNPs [27] and genes [28] associated with certain phenotypes and predict protein subnuclear localizations based on physicochemical composition features [29]. Recently, the hybrid SVM/GA approach was also utilized for selecting the optimum combinations of specific histone epigenetic marks to predict enhancers [30]. Saeys et al. predicted splice sites from nucleotide acid sequence by utilizing the hybrid method combining SVM and estimation of distribution algorithms (EDA) that is similar to GA [31]. Nemati et al. further combined GA and ant colony optimization (ACO) together for feature selection, and the hybrid method was found to outperform either GA or ACO alone when predicting protein functions [32]. In addition, Kamath et al. [33] proposed a feature generation with an evolutionary algorithm (FG-EA) approach, which employs a standard GP algorithm to explore the space of potentially useful features of sequence data. The features obtained from FG-EA enable the SVM classifier to get higher precision.

Feature selection is an important topic in bioinformatics and is involved in the analysis of various kinds of data. The hybrid methods that utilize the evolutionary algorithm have been proven useful for feature selection when handling the complex biological data due to their efficiency and robustness.

4. Parameter Estimation in Modeling Biological Systems

In bioinformatics, one biological system can be modeled as a set of ordinary differential equations (ODEs) so that the dynamics of the systems can be investigated and simulated. For example, Zhan and Yeung modeled a molecular pathway with the following ODEs [34]:

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t), \theta), \\ x(t_0) &= x_0, \\ y(t) &= g(x(t)) + \eta(t), \end{aligned} \tag{1}$$

where $x \in R^n$ is the state vector of the system, $\theta \in R^k$ is a parameter vector, $u(t) \in R^p$ is the system's input, $y \in R^m$ is the measured data, $\eta(t) \sim N(0, \sigma^2)$ is the Gaussian white noise, and x_0 denotes the initial state. f is designed as a set of nonlinear transition functions to represent the dynamical properties of the biological system and g is a measurement function. It can be seen that, to make the model work, it is necessary to estimate the parameters in the model, which can be transformed into an optimization problem as follows:

$$P : \min_{\widehat{\theta}, \widehat{x}_0} \sum_{j=0}^{N-1} \sum_{i=1}^{n} w_{ij} \left\| y_i(t_j) - \widehat{y}_i(t_j \mid \widehat{\theta}) \right\|_l, \tag{2}$$

where $\widehat{y}(t_j) = g(\widehat{x}(t_j \mid \widehat{\theta}))$, $\|\cdot\|_l$ denotes the l-norm, $\widehat{x}(t_j \mid \widehat{\theta})$ is the variable at time t_j with parameter $\widehat{\theta}$, w_{ij} denotes the weight, and $\widehat{y}$ means the estimated value. The problem P could be solved easily by employing the evolutionary algorithms [35–37]. For example, Katsuragi et al. [38] employed GA to estimate the parameters required by the simulation of dynamics of the metabolite concentrations, and Ueda et al. [39] applied the real-coded genetic algorithm to find the optimal values of the parameters. Recently, in order

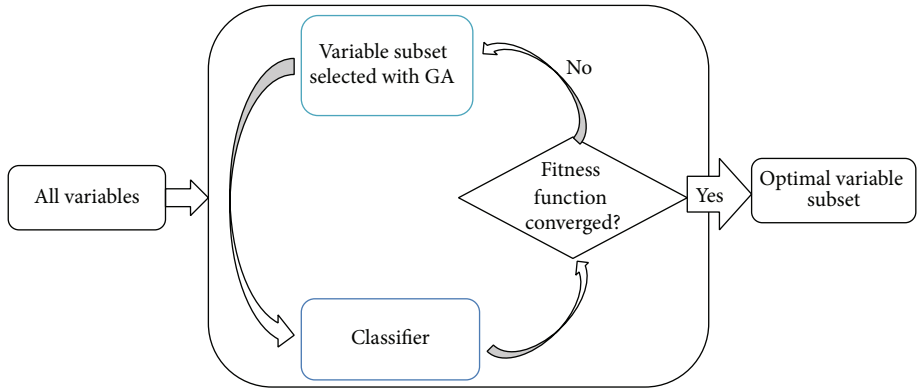


Figure 2: The flowchart of feature selection based on GA and classifier.

to improve the accuracy of parameter estimation, Abdullah et al. [40] proposed a novel approach that combines differential evolution (DE) with the firefly algorithm (FA), which outperformed other well-known approaches, such as particle swarm optimization (PSO) and Nelder-Mead algorithm.

In biological experiments, most data observed are measured at discrete time points while the traditional ODE model is a set of continuous equations, which makes it difficult to estimate the parameters in an accurate way. Therefore, the S-system, which is a type of power-law formalism and a particular type of ODE model, was widely used instead. For example, Savageau and Rosen [41] modeled the genetic network with the following S-system model:

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n+m} X_j^{h_{ij}}, \tag{3}$$

where X_i denotes the variable or reactant, n and m, respectively, denote the number of dependent and independent variables, α_i and β_i are nonnegative rate constants, and g_{ij} and h_{ij} are kinetic orders. Here, the parameters α_i, β_i, g_{ij}, and h_{ij} must be estimated. To optimize the parameters, Tominaga and Okamoto [42] utilized GA to approach the optimization problem with the following evaluation function E:

$$E = \sum_{i=1}^{n+m} \sum_{t=1}^{T} \left(\frac{X_i'(t) - X_i(t)}{X_i(t)} \right)^2, \tag{4}$$

where T is the number of sampling points and $X_i(t)$ and $X_i'(t)$, respectively, denote experimentally observed and estimated value at time t for X_i. Later, Kikuchi et al. [43] found that it is difficult to estimate all the parameters from limited time-course data of metabolite concentrations. Hence, they changed the evaluation function E as follows:

$$E = \sum_{i=1}^{n+m} \sum_{t=1}^{T} \left(\frac{X_i'(t) - X_i(t)}{X_i(t)} \right)^2 + c(n+m)T \left\{ \sum_{i,j} \left| g_{ij} \right| + \sum_{i,j,i \neq j} \left| h_{ij} \right| \right\}, \tag{5}$$

where c is a penalty constant that balances the two evaluation terms. Moreover, they adopted the simplex operations [44] instead of the random ones to accelerate the searching in GA. Considering only a few genes affecting both the synthesis and degradation processes of specific genes, Noman and Iba [45] further simplified the evaluation function as follows:

$$E_i = \sum_{t=1}^{T} \left(\frac{X_i'(t) - X_i(t)}{X_i(t)} \right)^2 + c \sum_{j=1}^{n+m-1} \left(\left| K_{i,j} \right| \right), \tag{6}$$

where $K_{i,j}$ is the kinetic order of gene i. With this objective function, they adopted a novel hybrid evolutionary algorithm, namely, memetic algorithm (MA) [46], that combines global optimization and local search together to find the optimal solutions. Considering that the traditional S-system can only describe instantaneous interactions, Chowdhury et al. [47] introduced the time-delay parameters to represent the system dynamics and refined the evaluation function as follows:

$$E = \sum_{t=1}^{T} \left(\frac{X_i^{\text{cal}}(t) - X_i^{\text{exp}}(t)}{X_i^{\text{exp}}(t)} \right)^2 + B_i \times C_i \frac{2N}{2N - r_i}, \tag{7}$$

where r_i is the number of all actual regulators, B_i is a balancing factor between the two terms, and C_i is the penalty factor for gene i. The trigonometric differential evolution (TDE) technique was adopted to estimate the set of parameters because of its better performance than other traditional evolutionary algorithms.

Parameter estimation is a key step in mathematical modeling of biological systems, which is however a nontrivial task considering the possible huge search space. Due to its excellent searching capability, the evolutionary algorithm is able to help determine the model parameters along with other intelligent approaches.

5. Molecular Network/Pathway Reconstruction

Recently, the network biology that represents a biological system as a molecular network or graph is attracting more and more attention. In the molecular network, the nodes denote the molecules, for example, proteins and metabolites, while

edges denote the interactions/regulations or other functional links between nodes. Although it is easy to observe the activity of thousands of molecules at the same time with high-throughput screening, it is not possible to detect the potential interactions/regulations between molecules right now.

Under the circumstances, a lot of intelligent methods have been presented to reconstruct the molecular networks, such as Boolean network and Bayesian network. When reconstructing the molecular networks, one critical step is to determine the topology of the network to be modeled, based on which the interactions/regulations between molecules can be investigated. The topology determination problem can be treated as an optimization problem that is ready to be solved with the help of the evolutionary algorithm.

Take a gene regulatory network as an example; Figure 3 shows the flowchart of reconstructing the regulatory network based on gene expression data by utilizing Boolean network and evolutionary algorithm. In the example, we want to reconstruct the regulatory circuit that controls the gene expression of five genes. Since at least one edge exists while at most 10 edges exist in the network, the number of possible network structures will be $M = \sum_{i=1}^{10} C_{10}^{i} = 2^{10} - 1 \approx 2^{10}$. It is impossible to validate all network topologies by biologists in lab. With appropriate fitness function, the evolutionary algorithm is able to identify the optimal network structure that fits best the gene expression data, where the consistence between network topology and gene expression data is evaluated with Boolean network based on certain rules.

Repsilber et al. [48] modeled the gene regulatory network with a Boolean model as a directed acyclic graph $G = (V, F)$, where $V = \{x_1, x_2, \ldots, x_n\}$ denotes the set of genes in the regulatory network and $F = \{f_1, f_2, \ldots, f_n\}$ denotes the Boolean rules that describe the regulations between nodes (or genes). To determine the topology of the regulatory network that better fits the observed data, they employed GA with the following fitness function f:

$$f = \frac{1}{1 + (1/D) \sum_{ijk} \delta_{ijk}^2}, \tag{8}$$

where $\delta_{ijk} = (\text{sim\_data}_{ijk} - \text{network\_output}_{ijk})$ is the difference between the observed data and those estimated from the generated network. In this way, they successfully reconstructed the gene regulatory network that generates the expression profiles consistent with experiments.

Later, Mendoza and Bazzan [49] presented inconsistency ratio (IR) to evaluate each individual node in the network, where the IR is defined as follows:

$$\text{IR}_i = w^{-1} \sum_{k=1}^{2^K} \min\left(w_k(0), w_k(1)\right). \tag{9}$$

Here, $k = 1, 2, 3, \ldots, 2^K$ is the number of possible input combinations for a node, $w_k(0)$ denotes the weight of measurements with output of 0 while $w_k(1)$ denotes those with output of 1, and w is the sum of all weights. With the IR defined above, an evaluation function defined below was used to investigate the inconsistency between the network generated and the experimental data:

$$\phi = \frac{1}{1 + \left(\sum_{i=1}^{N} \text{IR}_i / (N \times 0.5)\right) + \left(\text{NP}/N^2\right)}, \tag{10}$$

where $N \times 0.5$ denotes the maximum inconsistency to be generated by the network while (NP/N^2) is a penalty factor. With this evaluation function, the differential evolution (DE) approach was used to identify the optimal network structure [50].

Recently, to understand the signaling in distinct physiological situations, Terfve et al. [51] proposed a CellNOptR approach, which derives a Boolean logic model from a "prior knowledge network" and uses GA to search the optimal network structure that is consistent with the perturbation data. Later, Crespo et al. [52] employed Boolean logic model and genetic algorithm to predict missing gene expression values from experimental data and obtained promising results.

Although the Boolean network is simple and capable of handling large networks, it fails to provide quantitative information about regulations between molecules, which is however the key to understand the regulation process. In this case, the Bayesian network is widely adopted. Considering the expensive computation time required by Bayesian network, the evolutionary algorithm is widely used to determine the structures of the molecular networks modeled. In the Bayesian network, the molecular network is regarded as a directed acyclic graph described as follows:

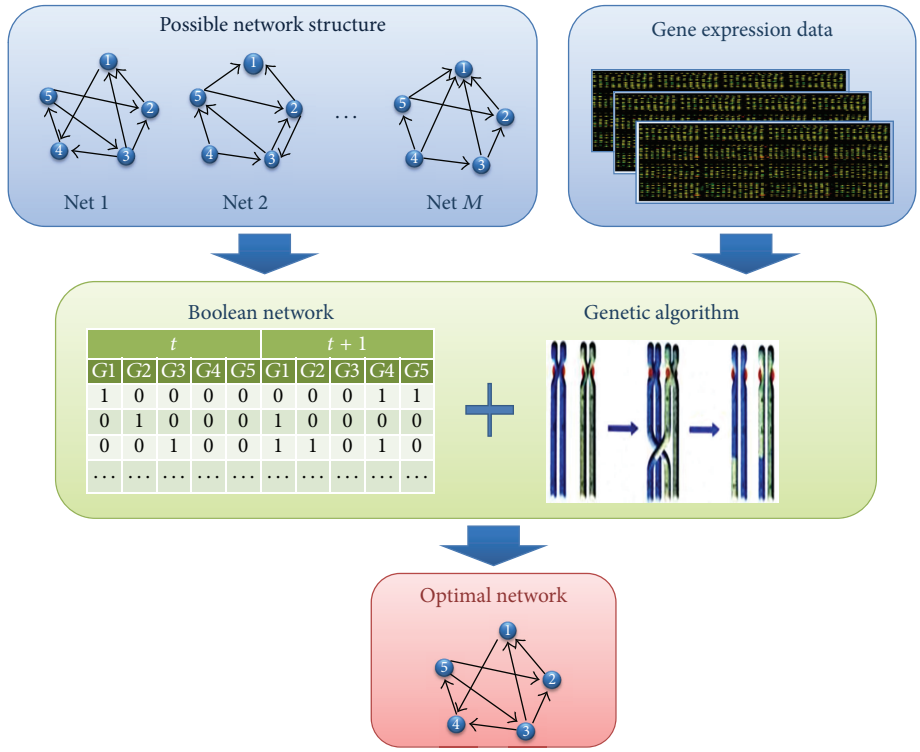
$$P\left(x_1, x_2, \ldots, x_n\right) = \prod_{i=1}^{n} P\left(x_i \mid \pi_i\right), \tag{11}$$

where x_i denotes node i in the set of variables, that is, the molecules considered, and π_i denotes the parent node of x_i. For example, Yu et al. [53] utilized GA to determine the optimal network structure consistent with experimental data along with the dynamic Bayesian network by defining an evaluation function based on Bayesian dirichlet equivalence (BDe) score and Bayesian information criterion (BIC) score. Later, Xing and Wu [54] employed the maximum likelihood (ML) score and the minimal description length (MDL) score as fitness values and determined the topology of gene regulatory networks with GA, where the regulatory network is modeled with Bayesian network. Recently, Li and Ngom [55] proposed a new high-order dynamic Bayesian network (HO-DBN) learning approach to identify genetic regulatory networks from gene expression time-series data and obtained the optimal structure of the networks with GA. In their method, the optimal structure $\widehat{S}$ was estimated by the maximum likelihood as follows:

$$\widehat{S} = \int_{\theta_s} P\left(X \mid \theta_s\right) P\left(\theta_s \mid S\right) d\theta_s, \tag{12}$$

where $X = \{x_1, x_2, \ldots, x_n\}$ and $\theta_s = \{\theta_1, \theta_2, \ldots, \theta_n\}$ is the parameter set.

In addition to Boolean and Bayesian networks, the Petri net [56] is also widely employed to reconstruct biological



t					$t+1$				
G1	G2	G3	G4	G5	G1	G2	G3	G4	G5
1	0	0	0	0	0	0	0	1	1
0	1	0	0	0	1	0	0	0	0
0	0	1	0	0	1	1	0	1	0
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯

Figure 3: The reconstruction of gene regulatory network based on gene expression with the hybrid method consisting of Boolean network and evolutionary algorithm.

networks. For example, in the Petri net model of metabolic networks, the nodes named places denote metabolites or products while transitions representing reactions are edges, where the values accompanying transitions denote rate constants. The input places for a transition denote the reaction's reactants while the output places denote its products, and the value of a place can be represented by its corresponding amount of substance. If a transition is deleted, a reaction happens, in which reactants are consumed and products are yielded. To find the optimal solutions, Nummela and Juistrom [57] defined a fitness function F as follows:

$$F = \sum \frac{|c_{mi} - c_{mi0}|}{n_m n_p} + 0.1 \times n_r, \tag{13}$$

where c_{mi} means the computed concentration of the mth metabolite at time i, c_{mi0} is the corresponding target concentration, n_m means the number of metabolites, n_p is the number of time steps, and n_r is the number of reactions. With the hybrid method combining the Petri net and GA, they successfully identified a network that is consistent with the simulated data. Later, Koh et al. [58] have also successfully employed this hybrid method to model the AKt and MAPK signaling pathways.

The molecular networks enable one to investigate the biological systems from a systematic perspective, whereas the network topology is the key to construct and understand the network. Accumulating evidence demonstrates that the hybrid heuristic methods involving evolutionary algorithm are able to help determine the network topology consistent with experimental data in an accurate way due to its significant efficiency.

6. Conclusions

In this paper, we surveyed the applications of hybrid intelligent methods, which combine several traditional intelligent approaches together, in bioinformatics. Especially, we introduced the hybrid methods involving evolutionary algorithm and their applications in three common problems in bioinformatics, that is, feature selection, parameter estimation, and reconstruction of biological networks. The evolutionary algorithm was selected here due to its capability of finding global optimal solutions and its robustness. The hybrid intelligent approaches that combine evolutionary algorithm together with other standard intelligent approaches have been proved extremely useful in the above three topics. We hope this review can help the researchers from both bioinformatics and informatics to understand each other and boost their future collaborations. We believe that, with more effective hybrid intelligent methods introduced in the future, it will become relatively easier to analyze the ever-growing complex biological data.

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 National Natural Science Foundation of China (91130032 and 61103075), the Innovation Program of Shanghai Municipal Education Commission (13ZZ072), and Shanghai Pujiang Program (13PJD032).

References

[1] T. Barrett and R. Edgar, "Gene expression omnibus: microarray data storage, submission, retrieval, and analysis," *Methods in Enzymology*, vol. 411, pp. 352–369, 2006.

[2] G. R. Abecasis, A. Auton, L. D. Brooks et al., "An integrated map of genetic variation from 1, 092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[3] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[4] S. Ramaswamy, P. Tamayo, R. Rifkin et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 15149–15154, 2001.

[5] K. Q. Liu, Z. P. Liu, J. K. Hao et al., "Identifying dysregulated pathways in cancers from pathway interaction networks," *BMC Bioinformatics*, vol. 13, article 126, 2012.

[6] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble, "Kernel-based data fusion and its application to protein function prediction in yeast," *Pacific Symposium on Biocomputing*, pp. 300–311, 2004.

[7] D. Barh, K. Gupta, N. Jain et al., "Conserved host-pathogen PPIs. Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in C. pseudotuberculosis, C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis, and E. coli targeted by Piper betel compounds," *Integrative Biology*, vol. 5, no. 3, pp. 495–509, 2013.

[8] 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.

[9] X.-M. Zhao, R.-S. Wang, L. Chen, and K. Aihara, "Uncovering signal transduction networks from high-throughput data by integer linear programming," *Nucleic Acids Research*, vol. 36, no. 9, article e48, 2008.

[10] A. S. Fraser, "Simulation of genetic systems by automatic digital computers. I. Introduction," *Australian Journal of Biological Sciences*, vol. 10, pp. 484–491, 1957.

[11] J. H. Holland, *Adaptation in Natural and Artificial Systems: an Introductory analysis with Applications to Biology, Control, and Artificial Intelligence*, MIT Press, 1992.

[12] N. A. Barricell, "Numerical testing of evolution theories," *Journal of Statistical Computation and Simulation*, vol. 1, no. 2, pp. 97–127, 1972.

[13] I. Rechenberg, *Evolutionsstrategie: Optimierung Technischer Systeme Nach Prinzipien Der Biologischen Evolution*, Technical University of Berlin, 1971.

[14] L. J. Fogel, A. J. Owens, and M. J. Walsh, *Artificial Intelligence Through Simulated Evolution*, John Wiley & Sons, 1966.

[15] R. Storn and K. Price, "Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[16] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, no. 12, pp. 1131–1142, 2002.

[17] T. Jirapech-Umpai and S. Aitken, "Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes," *BMC Bioinformatics*, vol. 6, article 148, 2005.

[18] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," *Machine Learning*, vol. 3, no. 2, pp. 95–99, 1988.

[19] J.-H. Hong and S.-B. Cho, "Efficient huge-scale feature selection with speciated genetic algorithm," *Pattern Recognition Letters*, vol. 27, no. 2, pp. 143–150, 2006.

[20] S. W. Chang, S. Abdul-Kareem, A. F. Merican et al., "Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods," *BMC Bioinformatics*, vol. 14, article 170, 2013.

[21] G. Mahdevar, J. Zahiri, M. Sadeghi, A. Nowzari-Dalini, and H. Ahrabian, "Tag SNP selection via a genetic algorithm," *Journal of Biomedical Informatics*, vol. 43, no. 5, pp. 800–804, 2010.

[22] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson, "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium," *American Journal of Human Genetics*, vol. 74, no. 1, pp. 106–120, 2004.

[23] K. A. Baggerly, J. S. Morris, J. Wang, D. Gold, L.-C. Xiao, and K. R. Coombes, "A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples," *Proteomics*, vol. 3, no. 9, pp. 1667–1672, 2003.

[24] E. F. Petricoin III, A. M. Ardekani, and B. A. Hitt, "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, vol. 359, pp. 572–577, 2002.

[25] L. Li, H. Tang, Z. Wu et al., "Data mining techniques for cancer detection using serum proteomic profiling," *Artificial Intelligence in Medicine*, vol. 32, no. 2, pp. 71–83, 2004.

[26] X.-M. Zhao, Y.-M. Cheung, and D.-S. Huang, "A novel approach to extracting features from motif content and protein composition for protein sequence classification," *Neural Networks*, vol. 18, no. 8, pp. 1019–1028, 2005.

[27] B. Gong, Z. Guo, J. Li et al., "Application of a genetic algorithm -Support vector machine hybrid for prediction of clinical phenotypes based on genome-wide SNP profiles of sib pairs," in *Proceedings of the 2nd International Confernce on Fuzzy Systems and Knowledge Discovery (FSKD '05)*, pp. 830–835, August 2005.

[28] L. Li, W. Jiang, X. Li et al., "A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset," *Genomics*, vol. 85, no. 1, pp. 16–23, 2005.

[29] W.-L. Huang, C.-W. Tung, H.-L. Huang, S.-F. Hwang, and S.-Y. Ho, "ProLoc: prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features," *BioSystems*, vol. 90, no. 2, pp. 573–581, 2007.

[30] M. Fernández and D. Miranda-Saavedra, "Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines," *Nucleic Acids Research*, vol. 40, no. 10, p. e77, 2012.

[31] Y. Saeys, S. Degroeve, D. Aeyels, P. Rouzé, and Y. Van de Peer, "Feature selection for splice site prediction: a new method using EDA-based feature ranking," *BMC Bioinformatics*, vol. 5, article 64, 2004.

[32] S. Nemati, M. E. Basiri, N. Ghasem-Aghaee, and M. H. Aghdam, "A novel ACO-GA hybrid algorithm for feature selection in protein function prediction," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12086–12094, 2009.

[33] U. Kamath, J. Compton, R. Islamaj-Dogan, K. A. De Jong, and A. Shehu, "An evolutionary algorithm approach for feature generation from sequence data and its application to DNA splice site prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 5, pp. 1387–1398, 2012.

[34] C. Zhan and L. F. Yeung, "Parameter estimation in systems biology models using spline approximation," *BMC Systems Biology*, vol. 5, article 14, 2011.

[35] M. Ashyraliyev, Y. Fomekong-Nanfack, J. A. Kaandorp, and J. G. Blom, "Systems biology: parameter estimation for biochemical models," *FEBS Journal*, vol. 276, no. 4, pp. 886–902, 2009.

[36] J. R. Banga and E. Balsa-Canto, "Parameter estimation and optimal experimental design," *Essays in Biochemistry*, vol. 45, pp. 195–209, 2008.

[37] C. G. Moles, P. Mendes, and J. R. Banga, "Parameter estimation in biochemical pathways: a comparison of global optimization methods," *Genome Research*, vol. 13, no. 11, pp. 2467–2474, 2003.

[38] T. Katsuragi, N. Ono, K. Yasumoto et al., "SS-mPMG and SS-GA: tools for finding pathways and dynamic simulation of metabolic networks," *Plant Cell Physiology*, vol. 54, no. 5, pp. 728–739, 2013.

[39] T. Ueda, D. Tominaga, N. Araki et al., "Estimate hidden dynamic profiles of siRNA effect on apoptosis," *BMC Bioinformatics*, vol. 14, article 97, 2013.

[40] A. Abdullah, S. Deris, S. Anwar, and S. N. Arjunan, "An evolutionary firefly algorithm for the estimation of nonlinear biological model parameters," *PloS One*, vol. 8, no. 3, Article ID e56310, 2013.

[41] M. A. Savageau and R. Rosen, *Biochemical Systems Analysis: A Study of Function and Design in molecular Biology (*, Addison-Wesley, 1976.

[42] D. Tominaga and M. Okamoto, "Design of canonical model describing complex nonlinear dynamics," in *Proceedings of the 7th International Conference on Computer Applications in Biotechnology*, pp. 85–90, 1998.

[43] S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi, and M. Tomita, "Dynamics modeling of genetic networks using genetic algorithm and S-system," *Bioinformatics*, vol. 19, no. 5, pp. 643–650, 2003.

[44] S. Tsutsui, M. Yamamura, and T. Higuchi, "Multi-parent recombination with simplex crossover in real coded genetic algorithms," in *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 657–664, 1999.

[45] N. Noman and H. Iba, "Reverse engineering genetic networks using evolutionary computation," *Genome Informatics*, vol. 16, no. 2, pp. 205–214, 2005.

[46] P. Moscato, "On evolution, search, optimization, genetic algorithms and martial arts: towards memetic algorithms," Caltech Concurrent Computation Program 826, 1989.

[47] A. R. Chowdhury, M. Chetty, and N. X. Vinh, "Incorporating time-delays in S-System model for reverse engineering genetic networks," *BMC Bioinformatics*, vol. 14, article 196, 2013.

[48] D. Repsilber, H. Liljenström, and S. G. E. Andersson, "Reverse engineering of regulatory networks: simulation studies on a genetic algorithm approach for ranking hypotheses," *BioSystems*, vol. 66, no. 1-2, pp. 31–41, 2002.

[49] M. R. Mendoza and A. L. C. Bazzan, "Evolving random boolean networks with genetic algorithms for regulatory networks reconstruction," in *Proceedings of the 13th Annual Genetic and Evolutionary Computation Conference (GECCO '11)*, pp. 291–298, July 2011.

[50] A. Esmaeili and C. Jacob, "A multi-objective differential evolutionary approach toward more stable gene regulatory networks," *BioSystems*, vol. 98, no. 3, pp. 127–136, 2009.

[51] C. Terfve, T. Cokelaer, D. Henriques et al., "CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms," *BMC Systems Biology*, vol. 6, article 133, 2012.

[52] I. Crespo, A. Krishna, A. Le Bechec, and A. del Sol, "Predicting missing expression values in gene regulatory networks using a discrete logic modeling optimization guided by network stable states," *Nucleic Acids Research*, vol. 41, no. 1, article e8, 2013.

[53] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis, "Advances to Bayesian network inference for generating causal networks from observational biological data," *Bioinformatics*, vol. 20, no. 18, pp. 3594–3603, 2004.

[54] Z. Xing and D. Wu, "Modeling multiple time units delayed gene regulatory network using dynamic Bayesian network," in *Proceedings of the 6th IEEE International Conference on Data Mining—Workshops (ICDM '06)*, pp. 190–195, December 2006.

[55] Y. Li and A. Ngom, "The max-min high-order dynamic Bayesian network learning for identifying gene regulatory networks from time-series microarray data," in *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB '13)*, pp. 83–90, 2013.

[56] G. Rozenberg and E. Engelfriet, "Elementary net systems," in *Lectures on Petri Nets I: Basic Models*, vol. 1497, pp. 12–121, 1998.

[57] J. Nummela and B. A. Juistrom, "Evolving, petri nets to represent metabolic pathways," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '05)*, pp. 2133–2139, June 2005.

[58] G. Koh, H. F. C. Teong, M.-V. Clément, D. Hsu, and P. S. Thiagarajan, "A decompositional approach to parameter estimation in pathway modeling: a case study of the Akt and MAPK pathways and their crosstalk," *Bioinformatics*, vol. 22, no. 14, pp. e271–e280, 2006.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 420856, 7 pages
http://dx.doi.org/10.1155/2014/420856

Research Article

Sparse Representation for Tumor Classification Based on Feature Extraction Using Latent Low-Rank Representation

Bin Gan,[1] Chun-Hou Zheng,[1,2] Jun Zhang,[2] and Hong-Qiang Wang[3]

[1] *College of Information and Communication Technology, Qufu Normal University, Rizhao 276800, China*
[2] *College of Electrical Engineering and Automation, Anhui University, Hefei 230000, China*
[3] *Intelligent Computing Lab, Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230000, China*

Correspondence should be addressed to Chun-Hou Zheng; zhengch99@126.com and Hong-Qiang Wang; hqwang@ustc.edu

Received 13 November 2013; Revised 27 December 2013; Accepted 27 December 2013; Published 11 February 2014

Academic Editor: Xing-Ming Zhao

Accurate tumor classification is crucial to the proper treatment of cancer. To now, sparse representation (SR) has shown its great performance for tumor classification. This paper conceives a new SR-based method for tumor classification by using gene expression data. In the proposed method, we firstly use latent low-rank representation for extracting salient features and removing noise from the original samples data. Then we use sparse representation classifier (SRC) to build tumor classification model. The experimental results on several real-world data sets show that our method is more efficient and more effective than the previous classification methods including SVM, SRC, and LASSO.

1. Introduction

Tumor is a solid lesion caused by the abnormal growth of cells. A timely accurate treatment is very important clinically. The premise of an accurate treatment is an exact diagnosis due to the heterogeneity of cancer. That is, we need to classify them accurately before treating tumors. Current methods for classifying cancer malignancies mostly rely on a variety of morphological, clinical, or molecular variables. Despite recent progresses, there are still many uncertainties in diagnosis. The advent of DNA microarray and RNA_seq [1] makes it possible to analyze tumor samples and classify them based on gene expression profiles. Moreover, we can get the expression data of tens of thousands of genes through DNA microarray or RNA-seq simultaneously.

Many methods for molecular data classification or clustering based on gene expression data have appeared in this area [2–14]. Huang and Zheng used independent component analysis [5] to extract features; Gao and Church introduced sparse nonnegative matrix factorization for feature extraction [4]; Zheng et al. proposed metasample-based sparse representation [7], and Furey et al. used support vector machines [8] to classify the gene expression data. All these methods have achieved impressive classification performances.

Recently published sparse representation classification (SRC) is also a powerful tool for processing gene expression data. SRC method was inspired by many theories such as Basis pursuing [15], compressive sensing for signal reconstruction [16], and least absolute shrinkage. It has already been widely used in face recognition [17] and texture classification [18]. In SRC method, test samples can be only represented as a sparse linear combination of the training samples from the same class. Furthermore, an imposed l_1-regularized least square optimization is used to calculate an SR coefficient vector with only a few significant coefficients. In theory, a test sample can be well represented by only using the training samples from the same class. However, there is too much noise in gene expression data, which causes that the discriminative features are not obvious and the test samples can also be represented by some training samples from different classes. This will decrease the classification accuracy. To reduce noise [19–21] and get salient features [20] for tumor classification, in this paper, we introduce latent

low-rank representation to preprocess gene expression data. By combining it with SRC algorithm, we propose a new method for tumor classification.

Latent low-rank representation (LatLRR) is a kind of theory which can be used to extract principal and salient features from original data. LatLRR is the improved version of LRR. The two methods can be resolved by the inexact augmented Lagrange multiplier (ALM) optimization. In [19–22], LRR has been successfully used for the recovery of subspace structure, subspace segmentation, feature extraction, outlier detection, and so forth. In [23], the author introduced LRR theory for face recognition in order to remove noise and achieved an impressive result. Based on these successful applications, in this paper, we introduce LatLRR into sparse representation classifier for tumor classification. Firstly, we use LatLRR to remove noise from original data and extract salient features. Then based on the new extracted salient features, we design sparse representation classifier to classify new test samples. We referred to the proposed method as SRC-based latent low-rank representation (SRC- LatLRR).

The rest of the paper is organized as follows. Section 2 describes our proposed SRC-LatLRR method in detail. We firstly review SRC and latent low-rank representation methods in Sections 2.1 and 2.2, respectively. Then we present our method in detail in Section 2.3. Section 2.4 specifies our experimental setting. In Section 3, we evaluate our method using several publicly available gene expression data sets. Section 4 concludes the paper and outlines our future work.

The abbreviations used in this paper are summarized in the Abbreviations section.

2. Methods

2.1. Sparse Representation Classification. Sparse representation classification is a supervised classification. Let $W \in R^{m\times n}$ denote a training sample matrix with n samples and m genes. As we know, each DNA microarray chip usually contains thousands of genes; the number of genes is much larger than tumor samples; that is, $m \gg n$.

Let c_l be the lth sample of W and the n samples are divided into k object classes. Assuming that there are n_i samples belonging to ith class and making up $W_i = [c_{i,1}, c_{i,2}, \ldots, c_{i,n_i}]$, the whole data set can be reexpressed as $W = [W_1, W_2, \ldots, W_k]$. Suppose that a new testing sample $y \in R^m$ belongs to ith class. Based on the theory of sparse representation, y would lie in the linear span of the training samples W_i; that is,

$$y = \alpha_{i,1}c_{i,1} + \alpha_{i,2}c_{i,2} + \cdots + \alpha_{i,n_i}c_{i,n_i}, \quad (1)$$

where $\alpha_{i,j} \in R$ is a scalar and $j = 1, 2, \ldots, n_i$.

Supposing a linear representation coefficient vector $x_0 \in R^n$, y can be also rewritten as

$$y = Wx_0. \quad (2)$$

Ideally, if the training samples are sufficient and the training samples sets that belong to different class are disjoint each other, then we have

$$x_0 = \left[0, \ldots, 0, \alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,n_i}, 0, \ldots, 0\right] \in R^n; \quad (3)$$

that is, in x_0, only the entries corresponding to the same class as y are nonzero.

From the above analysis, it can be seen that we can classify the test sample y according to x_0. So the key problem is how to calculate x_0 in (2). As in [7], x_0 would be sparse if the number of object classes k is large; this is what sparse representation implies. According to the theory of compressive sensing [16, 24–26] and SR, x_0 can be achieved by solving the following l_1-minimization problem:

$$\hat{x}_1 = \arg\min \|x\|_1 \quad \text{s.t. } Wx = y. \quad (4)$$

This problem can be solved by standard linear programming methods [15]. But (4) has no exact solutions since $m \gg n$. Then a generalized version of (4) can be conceived:

$$J(x, \lambda) = \min_x \left\{ \|Wx - y\|_2 + \lambda \|x\|_1 \right\}, \quad (5)$$

where λ is a scalar regularization. This function can balance the degree of noise by using λ. In this study, we solve this function by the truncated Newton interior-point method [27].

2.2. Latent Low-Rank Representation. Latent low-rank representation is an extension of low-rank representation. Consider an observed data matrix $X = [x_1, x_2, \ldots, x_n] \in R^{D\times n}$, where each column vector x_i is a sample, and a dictionary $A = [a_1, a_2, \ldots, a_m] \in R^{D\times m}$, where a_i is also a sample. X can be linearly represented by the dictionary. That is,

$$X = AZ, \quad (6)$$

where $Z = [z_1, z_2, \ldots, z_n] \in R^{m\times n}$ is a coefficient matrix and each z_i is the representation of x_i. Equation (6) means that each column vector of X can be represented by a linear combination of the bases in A. In (6), the dictionary A should be overcomplete enough to represent any observed data matrix X. But meanwhile, this causes multiple feasible solutions of Z to (6). To achieve the optimal solution, low rankness criterion is introduced to (6):

$$\min_Z \operatorname{rank}(Z), \quad \text{s.t. } X = AZ. \quad (7)$$

Here, the optimal solution Z^* is the so-called lowest-rank representation of data X with respect to the dictionary A. Unfortunately, function (7) can not be easy to solve because of the discrete nature of the rank function. By matrix completion method [28–30], we replace solving low-rank problem with dealing with nuclear norm [31]; then problem (7) can be rerepresented as

$$\min_Z \|Z\|_*, \quad \text{s.t. } X = AZ, \quad (8)$$

where $\|Z\|_*$ means the nuclear norm of matrix Z, that is, the sum of the singular values of matrix Z.

Strictly speaking, the dictionary A should be overcomplete and noiseless. But this kind of dictionary is difficult to get. In practice, we usually use observed data matrix X itself

as the dictionary [19, 21, 32]. Finally we have the following convex optimization problem:

$$\min_{Z} \|Z\|_*, \quad \text{s.t. } X = XZ. \tag{9}$$

To solve this equation, two conditions need to be met. Firstly, the data sampling X should be sufficient. Secondly, the data sampling X should also contain sufficient noiseless data to achieve robust capability. In fact, the first one can be easily met but the second one not. Because gene expression data are usually noisy, in reality, function (9) may be invalid and not robust.

To solve the problem in (9), we introduce the following LRR problem [20]:

$$\min_{Z} \|Z\|_*, \quad \text{s.t. } X_O = [X_O, X_H] Z, \tag{10}$$

where X_O is the observed data matrix and the X_H is the unobserved data, that is, the hidden data. We use the concatenation of X_O and X_H as a dictionary. The optimal result of (10) is $Z^*_{O,H} = [Z^*_{O|H}; Z^*_{H|O}]$, where $Z^*_{O|H}$ and $Z^*_{H|O}$ correspond to X_O and X_H, respectively.

By solving (10), the two problems above can be solved well. Then our next mission is to recover the affinity matrix $Z^*_{O|H}$ by using only X_O in the absence of the hidden data X_H. The method is called latent low-rank representation (LatLRR), which is an improvement of LRR.

Supposing we have two matrices X_O and X_H, then by solving (10) we have the following equations:

$$Z^*_{O|H} = V_O V_O^T, \qquad Z^*_{H|O} = V_H V_O^T, \tag{11}$$

where V_H and V_O can be obtained through computing the skinny singular value decomposition of $[X_O, X_H] = U \sum V^T$, and $V = [V_O; V_H]$. Namely, $X_O = U \sum V_O^T$ and $X_H = U \sum V_H^T$.

Depending on function (11), we have

$$\begin{aligned} X_O &= [X_O, X_H] Z^*_{O,H} \\ &= X_O Z^*_{O|H} + X_H Z^*_{H|O} \\ &= X_O Z^*_{O|H} + X_H V_H V_O^T \\ &= X_O Z^*_{O|H} + U \sum V_H^T V_H V_O^T \\ &= X_O Z^*_{O|H} + U \sum V_H^T V_H \sum^{-1} U^T X_O. \end{aligned} \tag{12}$$

Let $L^*_{H|O} = U \sum V_H^T V_H \sum^{-1} U^T$; then we have the following simple function:

$$X_O = X_O Z^*_{O|H} + L^*_{H|O} X_O. \tag{13}$$

If X_O and X_H come from the same collection of low-rank subspaces, then both $Z^*_{O|H}$ and $L^*_{H|O}$ should be of low-rank, so we can achieve

$$\begin{aligned} &\min_{Z_{O|H}, L_{H|O}} \quad \operatorname{rank}(Z_{O|H}) + \operatorname{rank}(L_{H|O}) \\ &\text{s.t.} \quad X_O = X_O Z_{O|H} + L_{H|O} X_O. \end{aligned} \tag{14}$$

Just as in [28–30], we also change the above rank minimization problem to the nuclear norm. Then we have the following convex optimization problem:

$$\min_{Z,L} \|Z\|_* + \|L\|_* \quad \text{s.t. } X = XZ + LX. \tag{15}$$

Here, we replace X_O, $Z_{O|H}$, and $L_{H|O}$ with X, Z, and L, respectively, for ease of representation. In (15), X is the noiseless observed data. By considering there may exist corrupted data or noise in X, we also need to introduce a denoising model about (15); then we have

$$\min_{Z,L} \|Z\|_* + \|L\|_* + \lambda \|E\|_1 \quad \text{s.t. } X = XZ + LX + E, \tag{16}$$

where $\lambda > 0$ is a scalar and $\|E\|_1$ is the l_1-norm of sparse noise matrix E. If $\lambda \to +\infty$, the problem (16) will be equivalent to (15), that is, no noise in the observed data X. In (16), the optimal solutions XZ^*, L^*X, and E^* represent the principal features, salient features, and noise, respectively.

To solve the LatLRR problem listed in (16), we introduce the augmented Lagrange multiplier (ALM) [33] method and revise (16) as follows to meet the requirement of ALM algorithm:

$$\begin{aligned} \min_{Z,L,J,S,E} \|Z\|_* + \|L\|_* + \lambda \|E\|_1 \quad &\text{s.t. } X = XZ + LX + E, \\ &Z = J, \ L = S. \end{aligned} \tag{17}$$

This problem can be solved by ALM method which minimizes the following augmented Lagrange function:

$$\begin{aligned} &\|J\|_* + \|S\|_* + \lambda \|E\|_1 + \operatorname{tr}\left(Y_1^T (X - XZ - LX - E)\right) \\ &\quad + \operatorname{tr}\left(Y_2^T (Z - J)\right) + \operatorname{tr}\left(Y_3^T (L - S)\right) \\ &\quad + \frac{\mu}{2}\left(\|X - XZ - LX - E\|_F^2 + \|Z - J\|_F^2 + \|L - S\|_F^2\right), \end{aligned} \tag{18}$$

where $\operatorname{tr}(\cdot)$ and $\|\cdot\|_F$ denote the trace and Frobenius norm of a matrix, respectively. $\mu > 0$ is a penalty parameter. More details about (18) can be found in [33].

2.3. Sparse Representation Classification Based on LatLRR. Since LatLRR can extract the salient features and remove noise from original data sets, in this study, before using observed data for classification, we firstly use LatLRR to suppress noise and get the salient features. Then we use the denoised data for tumor classification; that is, we factorize the observed data X into

$$X = XZ + LX + E. \tag{19}$$

Here, we only use $D = LX$ for data classification. For a test sample y, we can calculate its SR by the following function:

$$J(x, \lambda) = \min_{x} \{\|Dx - Ly\|_2 + \lambda \|x\|_1\}, \tag{20}$$

where the parameter $\lambda > 0$ can be determined experimentally and x is a coefficient vector. Assuming the test sample y

belongs to one of target classes, the training data set is sufficient. When classifying y, we introduce Ly, where L is a square matrix obtained through LatLRR method when extracting the salient features.

Ideally, Ly can be linearly represented by the samples from the same class in D. Namely, the representation vector x should be sparse and the nonzero entries are associated with the columns of D from the same class. This will lead us to classify the test samples. However, noise and modeling errors will also introduce some nonzero entries to x which correspond to the columns of D from the multiple classes [17]. To solve this problem, we classify Ly based on how well it can be reconstructed by using the coefficients from each class as in [17].

Using the result of (20), we construct $\delta_i(x)$ as the characteristic function which selects the coefficients associated with the ith class in the coefficient vector x. By only using ith class coefficients to reconstruct the test sample Ly as $\widehat{y}_i = D\delta_i(x)$, we can classify Ly into the minimum residual class between Ly and $\widehat{y}$; that is,

$$\min_i r_i\ (y) = \left\| Ly - D\delta_i\ (x) \right\|_2. \tag{21}$$

Our classification algorithm can be summarized as follows.

Input. Observed data $X \in R^{m\times n}$ for k classes; test sample y.

Step 1. Normalize the columns of X.

Step 2. Extract the salient features of X and remove to some extent noise to get data D defined in (19).

Step 3. Solve the optimization problem defined in (20).

Step 4. Compute the residuals $r_i(y) = \|Ly - D\delta_i(x)\|_2$.

Output. Identity(y) = arg $\min_i r_i(y)$.

Our method can be seen as the combination of SRC [17] and latent low-rank representation for feature extraction [20], so we named it as SRC-LatLRR. In SRC, the test sample is represented as a sparse linear combination of the training samples from the same class. In LatLRR, noise is removed to some extent and salient features are simultaneously extracted from the training samples. So the introduction of LatLRR can improve the classification accuracy of SRC in a way.

2.4. Evaluation of the Performance. To evaluate our proposed method, we compare our method with SRC [17, 34], LASSO [35], and SVM [8, 36, 37]. SVM has been proved to be one of the best classifiers for classifying data in the area of "high dimensionality and small sample size" [36, 37]. We do binary classification and multiclass classification experiments in Sections 3.1 and 3.2, respectively. During the experiment, the best results of SRC, LASSO, and SVM are also used to compare with those of our method, which were achieved by choosing appropriate parameters experimentally. As the number of tumor sample is too small, we use stratified 10-fold cross validation in all our experiments. In the multiclass classification experiments, we do not use LASSO method because it is designed only for binary class classification problems [35]. As we know, dimensionality reduction can improve the classification performance and computing speed, so we reduce data dimensionality using between-category to within-category sums of squares methods in our experiments.

Table 1: Three binary data sets used in the experiments.

Datasets	Samples		Genes
	Class 1	Class 2	
Colon cancer	40	22	2000
Prostate cancer	77	59	12600
DLBCL	58	19	5469

Table 2: Classification accuracies by different methods for the three binary data sets.

Datasets	SVM	LASSO	SRC	SRC-LatLRR
Colon cancer	85.48	85.48	85.48	**90.32**
Prostate cancer	91.18	91.91	**94.85**	94.12
DLBCL	96.10	96.10	**97.40**	**97.40**

3. Experimental Results

3.1. Two-Class Classification Problem. In this subsection, three two-class microarray data sets are used to evaluate our method: colon cancer [38], prostate cancer [39], and diffuse large B-cell lymphoma [40].

The colon data set contains 62 samples consisting of 40 tumor and 22 normal. The prostate data set contains prostate tumors and normal prostate samples, each consisting of the expression levels of 12600 genes. For the DLBCL data set, the gene expression values were measured by high-density oligonucleotide microarrays. An overview of the three data sets is given in Table 1.

The classification results by using SVM, LASSO, SRC, and the proposed SRC-LatLRR are listed in Table 2. From Table 2, we can see that our method SRC-LatLRR performs well on all the three data sets. Even the performance of SRC-LatLRR is not better than SRC on the prostate cancer data set, but it is better than SVM and LASSO. In summary, SRC has an advantage for the prostate cancer and DLBCL data sets, but SRC-LatLRR is the best classifier for the colon cancer and DLBCL data sets.

To further evaluate our method, in this experiment, we also introduced BW feature selection in our method to classify these three data sets. The results are listed in Table 3, and the number of genes selected is given in the parenthesis behind data set. From Table 3, we can see that after feature selection, our proposed classification method outperforms the other three classification methods, and it can even achieve an accuracy of 100% for the DLBCL data set.

3.2. Multiclass Classification Problem. In this subsection, we use four multiclass data sets to further check the classification

Table 3: Classification accuracies by different methods with gene selection for the three binary data sets.

Datasets	SVM	LASSO	SRC	SRC-LatLRR
Colon cancer (1000)	87.1	87.1	87.1	**91.94**
Prostate cancer (1500)	94.85	91.18	95.59	**96.32**
DLBCL (800)	97.40	93.51	97.40	**100**

Table 4: Descriptions of the four multiclass data sets used in DNA classification experiments.

Dataset	Class counts	Samples	Genes
Lung cancer	5	203	12600
Leukemia	3	72	11225
11_tumors	11	174	12533
9_tumors	9	60	5726

Table 5: Classification accuracies by different methods for the multiclass data sets.

Dataset	SVM	SRC	SRC-LatLRR
Lung cancer	**96.05**	95.07	95.07
Leukemia	96.60	95.83	**98.61**
11_tumors	94.68	**94.83**	**94.83**
9_tumors	65.10	**66.67**	**66.67**

performance of SRC-LatLRR. The four data sets are lung cancer [41], leukemia [42], 11_tumors [43], and 9_tumors [44].

In lung cancer data set, there are four classes of lung cancer and normal class. This data set contains 203 samples. For leukemia data set, all the samples are classified into acute myelogenous leukemia, acute lymphoblastic leukemia, or mixed-lineage leukemia. The data set includes 72 samples with 11225 genes. For 11_tumors, there are 11 classes of samples, which are ovary, bladder/ureter, breast, colorectal, gastroesophagus, kidney, liver, prostate, pancreas, adeno lung, and squamous lung. This data set includes 174 samples. For the 9_tumors data set, there are 60 samples with 5726 genes. These 9 types of tumors are non-small-cell lung, colon, breast, ovarian, leukemia, renal, melanoma, prostate, and central nervous system. The detailed descriptions about these four data sets are listed in Table 4. All the four data sets were produced by oligonucleotide microarrays and the analysis tool Affymetrix GENECHIP [36].

The experimental results are listed in Table 5. From these results, we can see that the proposed method SRC-LatLRR does not have a clear advantage over SVM and SRC. The reason may be that in these data sets, the training samples of each class are very few so that the sample space is not complete.

We then introduced BW feature selection before applying our method. The obtained results are listed in Table 6. From the results we can see that the proposed method classified leukemia well. For the other data sets, it has no clear advantage. But it performed better than SRC for all the four data sets.

Table 6: Classification accuracies by different methods with gene selection for the multiclass data sets.

Dataset	SVM	SRC	SRC-LatLRR
Lung cancer (2000)	**96.62**	95.07	95.57
Leukemia (3000)	96.90	95.83	**98.61**
11_tumors (1000)	**96.07**	95.40	95.40
9_tumors (2000)	**85.84**	71.67	80.00

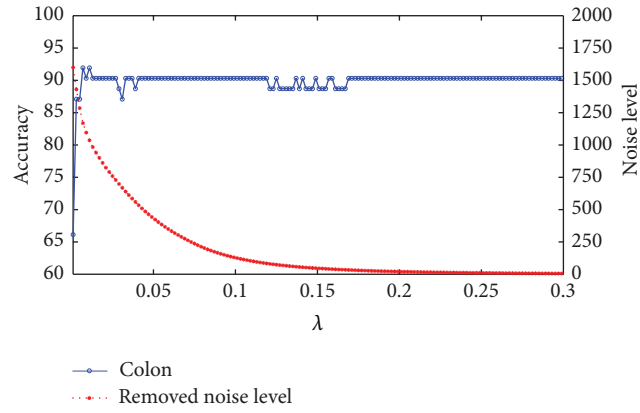


Figure 1: The changing curves of classification accuracy and removed noise level with λ on the colon data set.

3.3. The Choice of the Balanced Parameter. In this section, we use the data sets described in Section 3.1 to check how λ in (16) affect the classification performance. We show the accuracies and the removed noise level by our method at different values of λ in Figures 1, 2, and 3 for the colon, prostate, and DLBCL data sets, respectively. From (16), we know that the lower the λ is, the bigger the noise level is removed. For these three figures we use $\|E\|_1$ to represent the level of the removed noise. From these three figures we can see that the noise that we remove from the original data can not be too much, or it will reduce the accuracy. The reason is that if λ is set to be too small, useful information may be also removed besides noise. On the contrary, if λ is too big, the noise that was removed is too little, and we still can not get a good classification result. The experiment suggests that for colon data sets, $\lambda = 0.011$ is the best choice and $\lambda = 0.096$ and $\lambda = 0.1$ for the prostate and DLBCL data sets, respectively.

4. Conclusions

For gene expression data, cancer diagnosis is one of the most important clinical applications. In this paper, we have proposed a new SR-based method for tumor classification which uses the noiseless salient features extracted from the original samples to classify a test sample. We compared our method with several state-of-the-art methods including SVM, LASSO, and SRC on seven data sets. The results of experiments show that the proposed method is better than

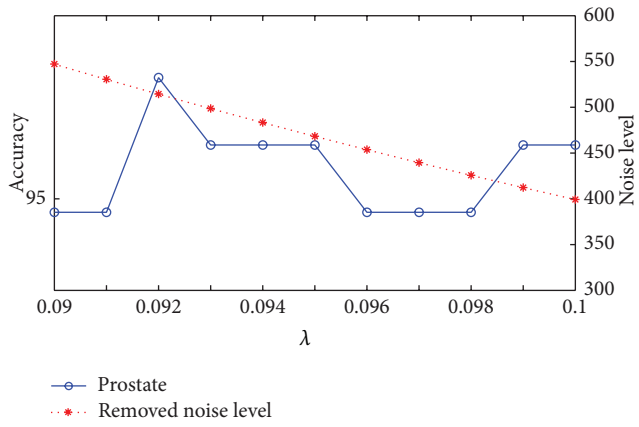


Figure 2: The changing curves of classification accuracy and removed noise level with λ on the prostate data set.

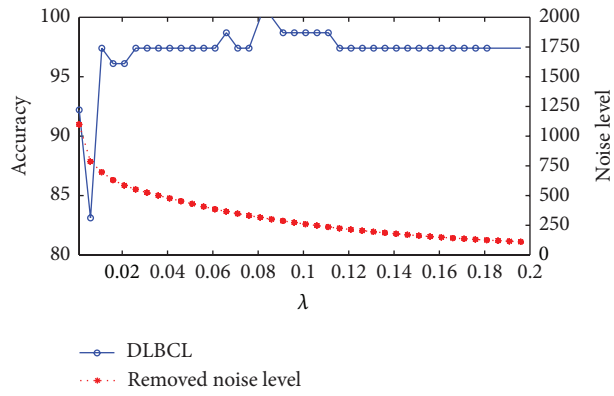


Figure 3: The changing curves of classification accuracy and removed noise level with λ on the DLBCL data set.

SVM, LASSO, and SRC in a way. These demonstrate that SRC-LatLRR is effective and efficient for tumor classification. We also introduced gene selection into our method. The results show that gene selection can improve the classification accuracy to some extent.

During the study we also found that, for the optimal result of LatLRR on the observed samples, Z^* represents the affinity matrix of samples [21]. In theory, the affinity matrix can be used to cluster samples. In future, we will extend it to investigate the property of sample clusters.

Abbreviations

SR: Sparse representation
SRC: Sparse representation classification
LRR: Low-rank representation
LatLRR: Latent low-rank representation
ALM: Augmented Lagrange multiplier
SVM: Support vector machines
LASSO: Least absolute shrinkage and selection operator
BW: Between-categories to within-category.

Conflict of Interests

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

Acknowledgments

This work was supported by the National Science Foundation of China under Grant nos. 61272339 and 61271098, and 61374181, the Natural Science Foundation of Anhui Province under Grant no. 1308085MF85, and the Key Project of Anhui Educational Committee, under Grant no. KJ2012A005.

References

[1] A. Cánovas, G. Rincon, A. Islas-Trejo, S. Wickramasinghe, and J. F. Medrano, "SNP discovery in the bovine milk transcriptome using RNA-Seq technology," *Mammalian Genome*, vol. 21, no. 11-12, pp. 592–598, 2010.

[2] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[3] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 12, pp. 4164–4169, 2004.

[4] Y. Gao and G. Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," *Bioinformatics*, vol. 21, no. 21, pp. 3970–3975, 2005.

[5] D. S. Huang and C. H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," *Bioinformatics*, vol. 22, no. 15, pp. 1855–1862, 2006.

[6] T. K. Paul and H. Iba, "Prediction of cancer Class with majority voting genetic programming classifier using gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 2, pp. 353–367, 2009.

[7] C. H. Zheng, L. Zhang, T. Y. Ng, C. K. Shiu, and D. S. Huang, "Metasample-based sparse representation for tumor classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 5, pp. 1273–1282, 2011.

[8] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[9] Y. S. Lee, A. Krishnan, Q. Zhu, and O. G. Troyanskaya, "Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies," *Bioinformatics*, vol. 29, no. 23, pp. 3036–3044, 2013.

[10] M. Tanic, E. Andress, S. M. Rodriguez-Pinilla et al., "MicroRNA-based molecular classification of non-BRCA1/2 hereditary breast tumours," *British Journal of Cancer*, vol. 109, no. 10, pp. 2724–2734, 2013.

[11] J. H. Huang, H. L. Xie, J. Yan, H. M. Lu, Q. S. Xu, and Y. Z. Liang, "Rsing random forest to classify T-cell epitopes based on amino acid properties and molecular features," *Analytica Chimica Acta*, vol. 804, pp. 70–75, 2013.

[12] L. Nanni, S. Brahnam, S. Ghidoni, E. Menegatti, and T. Barrier, "Acomparison of methods for extracting information from

the co-occurrence matrix for subcellular classification," *Expert Systems with Applications*, vol. 40, no. 18, pp. 7457–7467, 2013.

[13] G. R. Lioyd, L. M. Almond, N. Stone et al., "Utilising non-consensus pathology measurements to improve the diagnosis of oesophageal cancer using a raman spectroscopic probe," *The Analyst*, vol. 139, no. 2, pp. 381–388, 2014.

[14] G. Braz, S. V. da Rocha, M. Gattass, A. C. Silva, and A. C. de Paiva, "A mass classification using spatial diversity approaches in mammography images for false positive reduction," *Expert Systems with Applications*, vol. 40, no. 18, pp. 7534–7543, 2013.

[15] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[16] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[17] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[18] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS '09)*, 2009.

[19] G. C. Liu, Z. C. Lin, S. C. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[20] G. Liu and S. C. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 1615–1622, Barcelona, Spain, November 2011.

[21] G. C. Liu, Z. C. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proceedings of the 27th International Conference on Machine Learning (ICML '10)*, pp. 663–670, 2010.

[22] G. C. Liu, H. Xu, and S. C. Yan, "Exact subspace segmentation and outlier detection by low-rank representation," *JMLR: Workshop and Conference Proceedings*, vol. 22, pp. 703–711, 2012.

[23] L. Ma, C. H. Wang, B. H. Xiao, and W. Zhou, "Sparse representation for face recognition based on discriminative low-rank dictionary learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2586–2593, June 2012.

[24] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[25] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[26] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society B*, vol. 73, no. 3, pp. 273–282, 2011.

[27] S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l_1-regularized least squares," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2007.

[28] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[29] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[30] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *Journal of Machine Learning Research*, vol. 11, pp. 2057–2078, 2010.

[31] M. Fazel, *Matrix rank minimization with applications [Ph.D. thesis]*, Stanford University, Stanford, Calif, USA, 2002.

[32] Y. Ni, J. Sun, X. Yuan, S. Yan, and L. F. Cheong, "Robust low-rank subspace segmentation with semidefinite guarantees," in *Proceedings of the 10th IEEE International Conference on Data Mining Workshops (ICDMW '10)*, pp. 1179–1188, Sydney, Australia, December 2010.

[33] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," Tech. Rep. UILU-ENG-09-2215, 2009.

[34] X. Hang and F. X. Wu, "Sparse representation for classification of tumors using gene expression data," *Journal of Biomedicine and Biotechnology*, vol. 2009, Article ID 403689, 6 pages, 2009.

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

[36] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.

[37] N. Pochet, F. de Smet, J. A. K. Suykens, and B. L. R. de Moor, "Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction," *Bioinformatics*, vol. 20, no. 17, pp. 3185–3195, 2004.

[38] U. Alon, N. Barka, D. A. Notterman et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6745–6750, 1999.

[39] D. Singh, P. G. Febbo, K. Ross et al., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[40] M. A. Shipp, K. N. Ross, P. Tamayo et al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[41] A. Bhattacharjee, W. G. Richards, J. Staunton et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13790–13795, 2001.

[42] S. A. Armstrong, J. E. Staunton, L. B. Silverman et al., "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, vol. 30, no. 1, pp. 41–47, 2002.

[43] A. I. Su, J. B. Welsh, L. M. Sapinoso et al., "Molecular classification of human carcinomas by use of gene expression signatures," *Cancer Research*, vol. 61, no. 20, pp. 7388–7393, 2001.

[44] J. E. Staunton, D. K. Slonim, H. A. Coller et al., "Chemosensitivity prediction by transcriptional profiling," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 19, pp. 10787–10792, 2001.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 732650, 12 pages
http://dx.doi.org/10.1155/2013/732650

Research Article

Walking on a Tissue-Specific Disease-Protein-Complex Heterogeneous Network for the Discovery of Disease-Related Protein Complexes

Thibault Jacquemin and Rui Jiang

MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, FIT 1-107, Beijing 100084, China

Correspondence should be addressed to Rui Jiang; ruijiang@tsinghua.edu.cn

Received 11 September 2013; Accepted 7 October 2013

Academic Editor: Xing-Ming Zhao

Besides the pinpointing of individual disease-related genes, associating protein complexes to human inherited diseases is also of great importance, because a biological function usually arises from the cooperative behaviour of multiple proteins in a protein complex. Moreover, knowledge about disease-related protein complexes could also enhance the inference of disease genes and pathogenic genetic variants. Here, we have designed a computational systems biology approach to systematically analyse potential relationships between diseases and protein complexes. First, we construct a heterogeneous network which is composed of a disease-disease similarity layer, a tissue-specific protein-protein interaction layer, and a protein complex membership layer. Then, we propose a random walk model on this disease-protein-complex network for identifying protein complexes that are related to a query disease. With a series of leave-one-out cross-validation experiments, we show that our method not only possesses high performance but also demonstrates robustness regarding the parameters and the network structure. We further predict a landscape of associations between human diseases and protein complexes. This landscape can be used to facilitate the inference of disease genes, thereby benefiting studies on pathology of diseases.

1. Introduction

With a vast amount of genetic variants detected by such techniques as traditional genome-wide association studies [1, 2] and recent exome sequencing studies [3, 4], connecting functional implications of these genetic variants to human inherited diseases has now become a standard task [5]. For genetic variants occurring in protein coding regions, a typical approach to this task is to screen out a set of candidate genes around the genomic positions where the genetic variants occur and then prioritize the candidates to identify genes that are most likely to be associated with a disease of interest [6, 7].

To achieve this goal, quite a few approaches have been proposed from the perspective of computational systems biology. For example, Endeavour resorted to the guilt-by-association principle [8] to rank candidate genes according to their functional similarities to a set of predefined seed genes [9]. Cipher integrated a phenotype similarity profile and a protein-protein interaction (PPI) network to make a global inference of disease genes [10]. The idea of relying on phenotype similarities between diseases instead of between predefined seed genes to make inferences has then been extended by a series of methods, including RWRH [11], PRINCE [12], AlignPI [13], MAXIF [14], and many others [15–17]. In these studies, PPI networks have also been dominantly used to provide a simplified yet systematic measure of functional similarities between gene products [7], and recent studies have shown the advantage of using tissue-specific PPI networks over using generic ones [18–20].

However, a biological function usually arises from the cooperation of multiple proteins. These proteins link to each other by noncovalent interactions, forming a protein complex. Hence, genetic variants occurring at different loci might affect the structure of a member protein of a complex, alter the

function of the entire complex, and cause a disease. For example, it has been reported that seven pathogenic genes responsible for a heterogeneous syndrome called Fanconi anemia (FA) form a protein complex with functions related to DNA repair [21]. Therefore, besides the prioritization of candidate genes for a disease of interest, it is also of great importance to identify protein complexes underlying a query disease, thereby shedding light on biological processes and functional mechanisms of the occurrence and development of the disease under investigation.

Some methods for identifying disease genes have paid attention to linking protein complexes to diseases and then made use of such information to facilitate the prediction of disease genes. For example, Lage et al. proposed to identify the aggregates of proteins connected to a candidate protein in a PPI network as a protein complex by a virtual pull-down procedure and infer the association between the candidate protein and a query disease based on members of the protein complex [15]. Vanunu et al. proposed to analyze the PPI network and to establish a prioritization procedure in order to identify densely connected subnetworks that contain high scoring proteins as disease-related protein complexes [12]. Yang et al. proposed to infer disease genes from relationship between protein complexes and diseases [22]. These studies demonstrate that association relationships between protein complexes and a query disease could enhance the inference of disease genes. However, so far it still lacks a computational approach to systematically analyze potential relationships between known protein complexes and human diseases.

With the above understandings, we propose in this paper a computational systems biology approach for the identification of protein complexes that are related to a query disease via a random walk model on a heterogeneous network that is composed of a disease-disease similarity layer, a tissue-specific protein-protein interaction layer, and a protein complex membership layer. Starting from the query disease at the disease layer, our method simulates the process in which a random walker travels in the three-layered disease-protein-complex network, scores a protein complex using the probability that the walker stays in the protein complex at the steady state, and then ranks candidate protein complexes according to their scores. With a series of large-scale leave-one-out cross-validation experiments, we systematically show that our method not only possesses high performance but also demonstrates robustness to parameters involved and the network structure. As an application of our approach, we predict a landscape of associations between human diseases and known protein complexes and provide free downloads of the prediction results at http://bioinfo.au.tsinghua.edu.cn/jianglab/complex.

2. Methods

2.1. Overview of the Proposed Method. We model the problem of identifying protein complexes associated with a query disease as a prioritization problem and propose to solve this problem with a three-step approach. As illustrated in Figure 1, given a query disease and a set of predefined protein complexes as inputs, we first identify the tissue to which the disease is most likely related. Then, we construct a tissue-specific disease-protein-complex heterogeneous network, which is composed of three layers: a disease-disease similarity layer on the top, a protein-protein interaction layer in the middle, and a protein complex membership layer at the bottom. In this procedure, we use a PPI network that is specific to the tissue identified in the first step as the middle layer. Finally, we apply a random walk with restart algorithm to the three-layer network to calculate a score for each candidate complex and further rank the candidates to obtain a ranking list as the output.

2.2. Construction of the Disease-Protein-Complex Network. The disease-protein-complex network is composed of three layers. The top layer is a disease-disease similarity network derived from a phenotype similarity profile [23]. The middle layer is a tissue-specific PPI network derived using generic PPI information [24] and tissue-specific gene expression data [25]. The bottom layer reflects relationships between proteins and complexes that are extracted from the database [26].

At the top layer, given a disease phenotype similarity profile (a real-valued matrix) that quantifies pairwise overlaps of diseases in their clinic traits, we construct the disease-disease similarity network by using two strategies. First, with a k-nearest neighbour (k-NN) strategy (used as the default in our study), we link each disease to its k nearest neighbours, which correspond to the k highest phenotype similarity scores. Second, with a δ-threshold strategy, we set up a cut-off value δ and then connect two diseases by an undirected edge if and only if their similarity is greater than or equal to the cut-off. In both strategies, we further consider two variations for edges: weighting edges by the original similarity values or treating edges as unweighted.

At the middle layer, given generic PPI network and tissue-specific gene expression data, we get a tissue-specific PPI network from the literature [18]. These networks have been constructed by using one of the two following strategies. The first one is a naïve node removal (NR) strategy: a tissue-specific network is constructed by removing proteins that are not expressed in the given tissue from the generic PPI network. The second one is an edge reweight (ERW) strategy (used as the default in our study): each edge in the tissue-specific network is assigned a weight (controlled by a parameter $0 \leq \mathrm{rw} \leq 1$ with default value 0.1 [18]), reflecting the possibility that both endpoints of the edge are expressed in the given tissue. We further connect the top layer and the middle layer by undirected edges that correspond to known associations between diseases and proteins, and we weight these edges by a positive real-valued parameter α.

At the bottom layer, given a collection of protein complexes, we connect each of them to all of its member proteins in the PPI network at the middle layer by undirected edges, while leaving protein complexes unconnected. We weight the introduced edges by a positive real-valued parameter β.

Formally, we describe the disease-disease similarity network by a weight matrix $\mathbf{D} = (d_{ij})_{l\times l}$, where l is the number of diseases and d_{ij} is the weight of the edge between the ith and

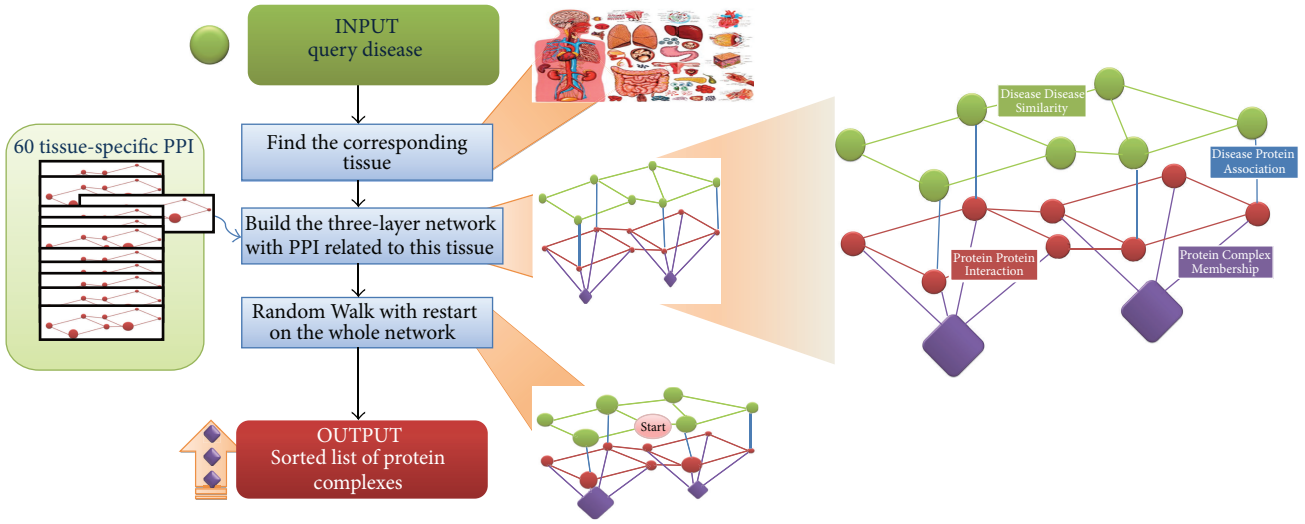


FIGURE 1: Illustration of the proposed method. Our method takes as inputs a query disease and a set of candidate protein complexes and gives a ranking list of the candidates as the output. For this purpose, we construct a tissue-specific disease-protein-complex heterogeneous network, apply a random walk with restart algorithm to the network to obtain scores for candidate protein complexes, and rank the candidates according to their scores.

jth diseases or 0 if the edge is absent. We describe the tissue-specific PPI network by a weight matrix $\mathbf{P} = (p_{ij})_{m\times m}$, where m is the number of proteins and p_{ij} is the weight of the edge between the ith and jth proteins or 0 if the edge is absent. We describe connections between the diseases and proteins by a weight matrix $\mathbf{A} = (a_{ij})_{l\times m}$, where $a_{ij} = \alpha$ is the weight of the edge between the ith disease and the jth proteins or 0 if the edge is absent. We describe connections between proteins and complexes by a weight matrix $\mathbf{B} = (b_{ij})_{m\times n}$, where $b_{ij} = \beta$ is the weight of the edge between the ith protein and the jth complex or 0 if the edge is absent. Put together, the disease-protein-complex network can be represented using a block matrix, as

$$\mathbf{H} = \begin{pmatrix} \mathbf{D} & \mathbf{A} & \mathbf{0} \\ \mathbf{A}^T & \mathbf{P} & \mathbf{B} \\ \mathbf{0} & \mathbf{B}^T & \mathbf{0} \end{pmatrix}, \quad (1)$$

where $\mathbf{0}$ stands for a zero matrix and the superscript T stands for the transposition of a matrix.

2.3. Random Walking on the Disease-Protein-Complex Network. We achieve the goal of identifying protein complexes related to a specific query disease by calculating a score for each candidate complex and then rank the candidates to obtain a ranking list. The higher the rank, the more likely to be related to the query disease. For this purpose, we adapt the random walk with restart model [11, 27] to the constructed disease-protein-complex network.

At a quick glance, our model simulates the process that a random walker wanders on the three-layered disease-protein-complex network. When starting on, the walker chooses the query disease of interest as the starting point. In each step of the walking process, the walker may start on a new journey with probability γ or move on with probability $1-\gamma$. When moving on, the walker may move at random to one of its direct neighbours in the same layer, jump from the disease layer to the protein layer or vice versa, or jump from the protein layer to the complex layer or vice versa.

Formally, as illustrated in Algorithm 1, we use a vector $\mathbf{q}^{(0)} = (q_i^{(0)})_{(l+m+n)\times 1}$ to represent initial probabilities when a random walker starts a journey, with $q_i^{(0)}$ $(i = 1, \ldots, l+m+n)$ being the probability that the walker initially starts from the ith node. In this vector, the element corresponding to the query disease is set to 1, and all of the other elements are set to 0. We normalize each row of the weight matrix $\mathbf{H}$ for the disease-protein-complex network to obtain a transition matrix $\mathbf{T} = (t_{ij})_{(l+m+n)\times(l+m+n)}$, in which $t_{ij} = h_{ij} / \sum_{j=1}^{l+m+n} h_{ij}$ represents the probability that a random walker moves from the ith node to the jth node, with each node being a disease, a protein, or a complex. We use a vector $\mathbf{q}^{(t)} = (q_i^{(t)})_{(l+m+n)\times 1}$ to represent probabilities that the random walker stays on nodes at step t, with $q_i^{(t)}$ $(i = 1, \ldots, l+m+n)$ being the probability that the walker stays on the ith node. We then have the iterative updating formula as

$$\mathbf{q}^{(t+1)} = (1-\gamma)\,\mathbf{T}^T\mathbf{q}^{(t)} + \gamma\mathbf{q}^{(0)}. \quad (2)$$

After a number of updates, the probabilities that the random walker staying on nodes will reach a steady state, which can be determined by checking whether the difference between $\mathbf{q}^{(t)}$ and $\mathbf{q}^{(t+1)}$ is sufficiently small. In our implementation, we check whether the L_2 norm of $\Delta\mathbf{q} = \mathbf{q}^{(t+1)} - \mathbf{q}^{(t)}$ is less than or equal to a small positive number ϵ (with the default value 10^{-5}). With the steady-state probability (denoted by $\mathbf{q}^{(\infty)}$)

Require: A query disease i, the transition matrix $\mathbf{T}$ of the disease-protein-complex network.
Ensure: A score for each protein complex.

(1) $\mathbf{q}^{(0)} \leftarrow \mathbf{0}; q_i^{(0)} \leftarrow 1;$
(2) $\Delta\mathbf{q} \leftarrow +\infty; t \leftarrow 0;$
(3) **WHILE** $\Delta\mathbf{q} \geq \epsilon$
  (a) $\mathbf{q}^{(t+1)} = (1-\gamma)\,\mathbf{T}^T\mathbf{q}^{(t)} + \gamma\mathbf{q}^{(0)};$
  (b) $\Delta\mathbf{q} \leftarrow \left\|\mathbf{q}^{(t+1)} - \mathbf{q}^{(t)}\right\|;$
  (c) $t \leftarrow t + 1;$
(4) **END**
(5) $\mathbf{q}^{\infty} \leftarrow \mathbf{q}^{(t)};$
(6) **FOR** j **FROM** 1 **TO** n
  (a) $s_j = \mathbf{q}_{l+m+j}^{(\infty)} / \sum_{j=1}^{n} \mathbf{q}_{l+m+j}^{(\infty)};$
(7) **END**

Algorithm 1: The random walk algorithm on the disease-protein-complex heterogeneous network.

obtained, we further calculate a normalized score s_i for the ith complex as

$$s_i = \frac{q_{l+m+i}^{(\infty)}}{\sum_{i=1}^{n} q_{l+m+i}^{(\infty)}} \tag{3}$$

and use this score to quantify the strength of association between the complex and the query disease. With such scores calculated for candidate complexes, we further rank the candidates in nonincreasing order according to their scores to obtain the final ranking list.

In this paper, we set the default values for the parameters as disease-protein weight $\alpha = 1$, protein-complex weight $\beta = 1$, and restart probability $\gamma = 0.5$. By simulation studies, we find that our model is not sensitive to these parameters (see results for details).

2.4. Validation Method. We adopt a leave-one-out cross-validation experiment to assess the capability of our method to identify protein complexes that are associated with human diseases. For this reason, we define a protein complex as associated with a disease if at least one member protein of the complex has been annotated as associated with the disease, and we collect a set of test protein complexes as those associated with at least one disease. Then, in each validation run, we take a test protein complex, identify a query disease as the one with which the complex is associated, pretend that all annotated associations between the query disease and proteins (or corresponding genes) are unknown, and then rank the test protein complex against a collection of control protein complexes.

In the context of the disease-protein-complex network, the above validation procedure is equivalent to remove all edges connecting the query disease and proteins and see whether protein complexes containing these proteins could receive high ranks. In the context of genetics, this validation procedure is equivalent to hide all known genetic bases of the query disease and see whether some of them could be recovered at the protein complex level.

2.5. Evaluation Criteria. We adopt three classes of criteria to quantify the performance of our method. First, let us suppose that we have performed a total of N validation runs and collected the same number of ranking lists. We calculate a criterion named TOP which is the number of test protein complexes ranked first in their corresponding list. We also divide this number by N to obtain the fraction of first ranked test protein complexes and call this fraction precision (PRE). Second, we calculate the average rank of all test protein complexes as the second criterion called mean rank (MR). Alternatively, we normalize ranks of test protein complexes by the lengths of ranking lists to obtain relative ranks, and we calculate the average relative rank of all test protein complexes to obtain mean relative rank (MRR). Third, given a threshold of the relative rank, we calculate the sensitivity (true positive rate) as the fraction of test protein complexes ranked above the threshold and the specificity (true negative rate) as the fraction of control protein complexes ranked below the threshold. Varying the threshold value from 0.0 to 1.0, we draw a rank receiver operating characteristic (ROC) curve and further calculate the area under this curve (AUC). Obviously, larger TOP (PRE)/AUC and smaller MR/MRR indicate higher performance.

3. Results

3.1. Data Sources. We obtained disease-tissue associations from the literature [28]. Briefly, Lage et al. studied co-occurrence patterns of disease-tissue pairs in PubMed abstracts and quantified the strength of association between a disease and a tissue by a normalized Ochiai's coefficient [29], resulting in a matrix that contains association scores between 926 diseases and 60 tissues. Following the literature [18], we associated a disease with the tissue of the highest score among all tissues, obtaining a total of 926 disease-tissue associations.

We obtained disease-disease similarity scores from the literature [23]. Briefly, van Driel et al. used terms in the anatomy and disease sections of the medical subject headings vocabulary (MeSH) [30] as a standard vocabulary to analyse

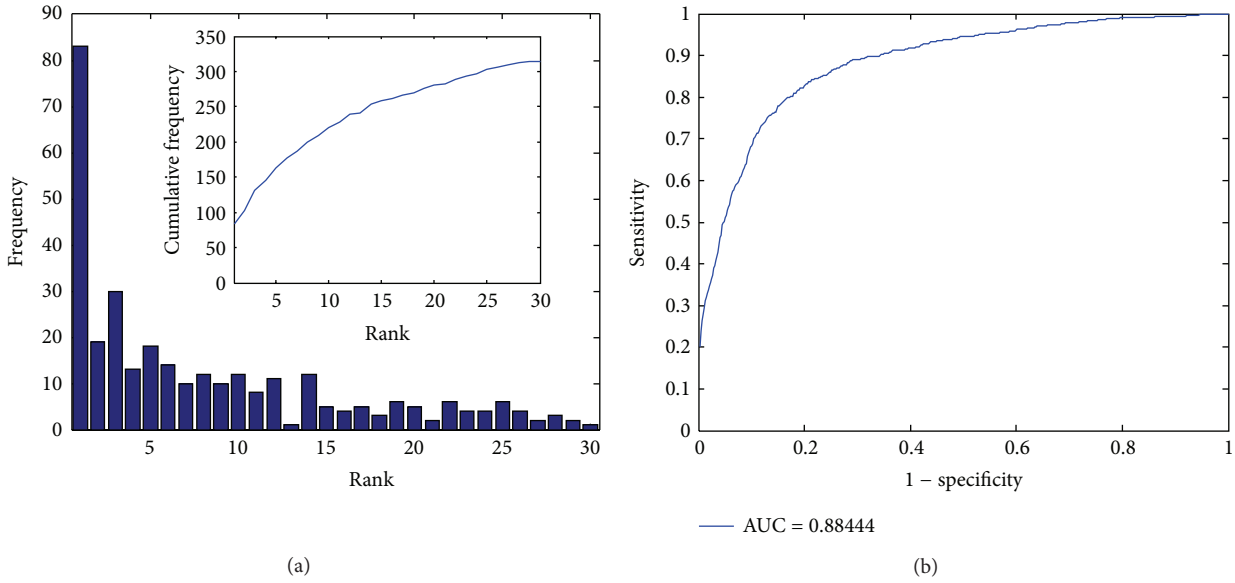


Figure 2: Performance of the proposed method. (a) Histogram of the ranks for the test protein complexes in the validation experiment. (b) The rank receiver operating characteristic (ROC) curve.

the full-text and clinical synopsis fields of OMIM records. By characterizing a disease using a vector composed of weighted phenotypic terms, they quantified the similarity between two diseases as the cosine of the angle of their vectors and obtained a matrix that contains pairwise similarity scores for 5,080 diseases [23].

We obtained tissue-specific PPI networks from the literature [18]. Given a specific tissue and a generic PPI network (9,998 proteins as nodes and 41,049 interactions as edges) extracted from the Human Protein Reference Database (HPRD) [24], Magger et al. derived two tissue-specific PPI networks for each of the 60 tissues by using both the edge reweight strategy and the node removal strategy [18].

We extracted disease-protein associations from the Ensembl database using the tool Biomart [31], obtaining a total of 5,164 associations between 3,504 diseases and 3,066 proteins (on February 26, 2013). Focusing on diseases with similarity scores and proteins that can be mapped back to the HPRD database, we obtain 1,962 associations between 1,548 diseases and 1,244 proteins.

We extracted 1,343 human protein complexes from the core set of the CORUM database (release in February 2013) [26], each of which contains at least one protein that can be mapped back to the HPRD database. By considering a protein complex as associated with a disease if at least one of its member protein has been annotated as associated with the disease, we collected a set of 939 disease-related protein complexes as test cases.

3.2. Performance of the Proposed Method. With the collected data and the default parameter setting ($k = 15$, $\alpha = 1$, $\beta = 1$, $\gamma = 0.5$), we constructed a disease-protein-complex network that was composed of 5,080 diseases, 9,998 proteins, and 1,343 protein complexes. There were a total of 107,661 edges in the network, among which 58,448 are between diseases, 41,049 are between proteins, 1,962 are connecting diseases and proteins, and 6,202 are connecting proteins and protein complexes.

We then performed the leave-one-out cross-validation experiment using this network and showed the results in Figure 2. By counting the number of test protein complexes with different ranking position, we observed that 83 (8.84%) test cases were ranked first, 163 (17.36%) were ranked among top 5, 221 (23.54%) were ranked among top 10, and 281 (29.93%) were ranked among top 20. In contrast, a random guess procedure that assigns ranks to protein complexes at random was only expected to rank 0.70 (0.07%) test cases at first ($939/1343 \approx 0.7, 1/1343 \approx 0.07\%$), 3.50 (0.37%) among top 5, 6.99 (0.74%) among top 10, and 13.98 (1.49%) among top 20. These results, as illustrated in Figure 2(a), therefore strongly suggest the effectiveness of our method in identifying disease-related protein complexes from a collection of candidates.

We further calculated the proposed evaluation criteria in Algorithm 1 and plotted the ROC curve in Figure 2(b). According to these results, our method achieves a TOP (PRE) of 83 (8.84%), a mean rank (mean relative rank) of 169.04 (12.59%), and an AUC of 88.44%, also supporting the effectiveness of this approach. The ROC curve, as shown in Figure 2(b), climbs fast towards the top-left corner of the plot and again suggests the effectiveness of our method.

A naïve thinking of identifying disease-related protein complex is to quantify the strength of associations between proteins and the query disease and then sum over the scores of member proteins to obtain a score for a protein complex. The main difference between this naïve approach and our method is that when a protein is contained in multiple protein complexes, the score of the protein will be counted multiple times (once for a protein complex) in the naïve approach,

Table 1: Comparison of the proposed approach and the naïve approach.

	Proposed method	Naïve approach
Top (PRE)	83 (8.84%)	75 (7.99%)
MR (MRR)	169.04 (12.59%)	180.49 (13.44%)
AUC	88.44%	87.57%

Table 2: Comparison of different strategies for constructing the protein-protein interaction network.

	Edge reweight	Node removal	HPRD
TOP (PRE)	83 (8.84)%	83 (8.84)%	75 (7.99)%
MR (MRR)	169.04 (12.59%)	168.52 (12.55%)	187.65 (13.97%)
AUC	88.44%	88.49%	87.03%

while with our method, such phenomenon will not happen because the probability of going out from the protein will be distributed uniformly to the multiple protein complexes in the random walk procedure. We performed a comparison between these two methods and showed the results in Table 1. It is clear, according to this table, that our approach outperforms the naïve approach in all of the three criteria. In detail, our method achieves a TOP of 83, a mean rank of about 169.04, and an AUC of 88.44%, while the naïve approach obtains these criteria as 75, 180.49, and 87.57%, respectively, all supporting the conclusion that our method performs better than the naïve approach.

3.3. Comparison of Different Strategies for Constructing the Disease Similarity Layer. We considered two strategies for constructing the disease similarity network at the top layer of the disease-protein-complex network: the k-nearest neighbour (k-NN) strategy and the δ-threshold strategy. In both strategies, we further considered two variations: weighting edges by the original similarity values or treating edges as unweighted. We then conducted a comparative study of these strategies and presented the results in Figure 3.

We first observe that our method with the weighted disease similarity network outperforms that with the unweighted one in terms of the precision of test protein complexes (PRE), and the difference between these two variations is subtle according to the other two criteria (MRR and AUC), though the weighted one slightly outperforms the unweighted one. For example, with the k-NN strategy and the default parameter setting, the PRE, MRR, and AUC are 8.84%, 12.59%, and 88.44% for the weighted variation, respectively, and 7.88%, 12.71%, and 88.32% for the unweighted one, respectively. Using the δ-threshold strategy ($\delta = 0.35$) with the default parameter setting, the PRE, MRR, and AUC are 6.28%, 13.69%, and 87.34% for the weighted variation, respectively, and 5.64%, 13.72%, and 87.30% for the unweighted one, respectively. With these observations, we conjecture that the weighted disease similarity network is preferred by our method and will use this network as the top layer of our disease-protein-complex network in the rest of this paper.

Second, we also observe that our method is quite robust to the number of neighboring diseases in the k-NN strategy. All of the three criteria only show small fluctuations in a wide range of the parameter k. Focusing on weighted networks, the PRE, MRR, and AUC are in general greater than 3.94% 16.25%, and 84.76%, respectively, when k is greater than 10 and less than 500, with the optimum values of these criteria achieved at $k = 15$, 20, and 20, respectively. For the δ-threshold strategy, our method is also quite robust when the cut-off value δ is not too large. Also focusing on weighted networks, the PRE, MRR, and AUC are in general greater than 3.30%, 20.97%, and 79.99%, respectively, when δ is greater than 0.25 and less than 0.45, with the optimum values of these criteria achieved at $\delta = 0.45$, 0.35, and 0.35, respectively. With these observations, we conclude that the selection of the parameters k and δ is not critical and kind of flexible. To achieve a balance over all of the three criteria, we recommend to select $k = 15$ and $\delta = 0.35$ as default values of these parameters.

Third, we notice that the k-NN strategy gives us higher performance than the δ-threshold does in a wide range of parameter settings. When comparing the performance at the default parameters, the PRE, MRR, and AUC are 8.84%, 12.59%, and 88.44%, respectively, for the k-NN strategy and 6.28%, 13.69%, and 87.34%, respectively, for the δ-threshold strategy. Around these parameters, the k-NN strategy exhibits consistent higher performance than the δ-threshold strategy in terms of both MRR and AUC. Therefore, we recommend the use of the k-NN strategy in the construction of the disease similarity network.

3.4. Comparison of Different Strategies for Constructing the Protein-Protein Interaction Layer. We considered two strategies to construct the tissue-specific PPI network at the middle layer of the disease-protein-complex network: the node removal strategy and the edge reweight strategy. Besides, we also considered the use of a tissue-nonspecific PPI network extracted from the HPRD database as the middle layer. We then performed a comparison study of these strategies and presented the results in Table 2.

We first observe from this table that the difference between the node removal strategy and the edge reweight strategy is subtle. For example, with the default parameter setting, the PRE, MRR, and AUC are 8.84%, 12.55%, and 88.49% for the node removal strategy, respectively, and 8.84%, 12.59%, and 88.44% for the edge reweight strategy, respectively. This observation is consistent with a previous study about relying on a tissue-specific PPI network to prioritize candidate genes [18]. Therefore, following the literature [18], we focus on the edge reweight strategy in our study because the network constructed using this strategy exhibits preferred properties in connectivity.

We then notice from Table 2 that the tissue-specific PPI network gives us a better performance than the tissue-nonspecific one. For example, with the default parameter setting, the PRE, MRR, and AUC are 8.84%, 12.59%, and 88.44% for the tissue-specific PPI with edge removal strategy,

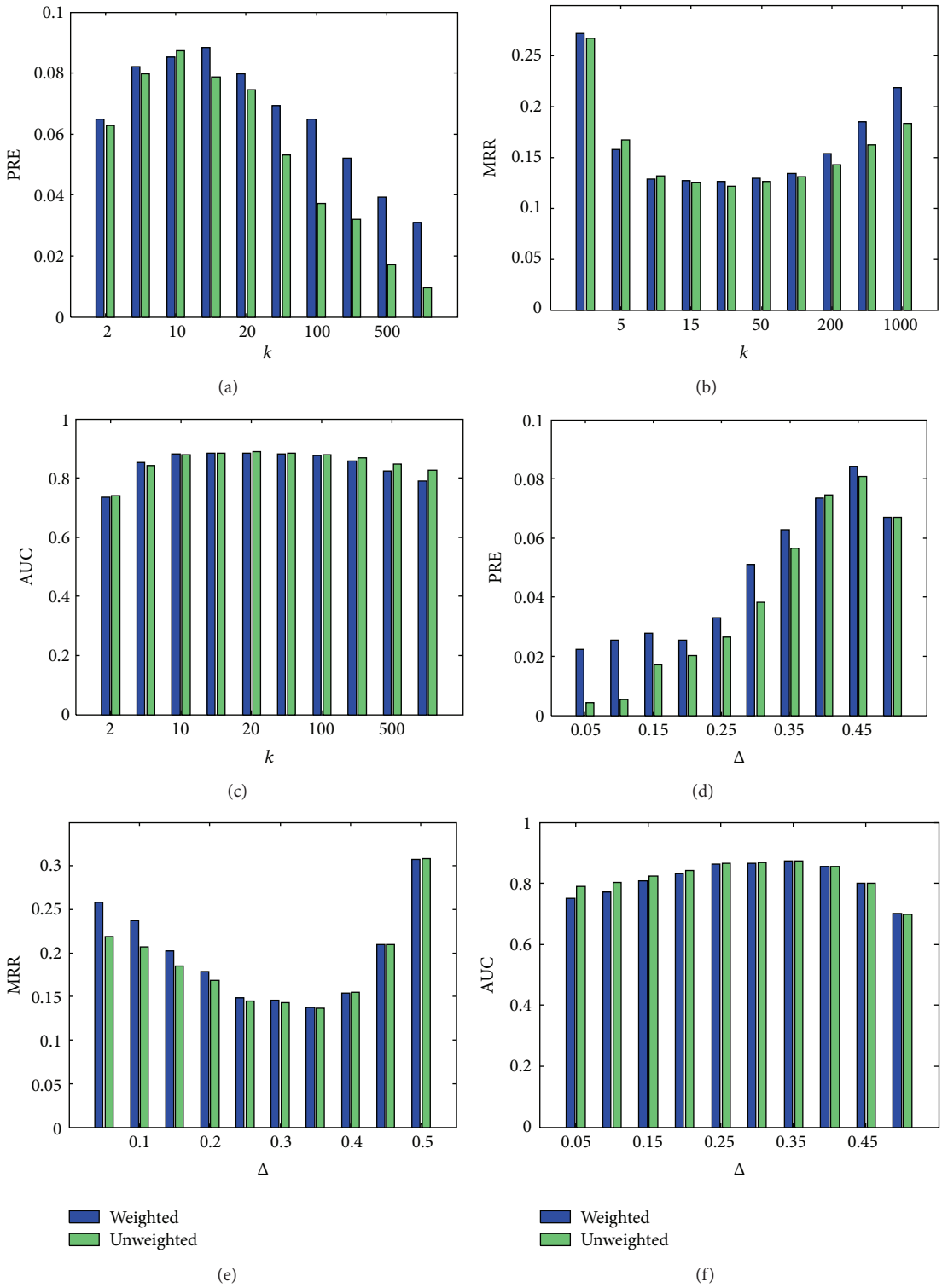


FIGURE 3: Comparison of different strategies for constructing the disease similarity layer. ((a)–(c)) PRE, MRR, and AUC for the k-NN strategy. ((d)–(f)) PRE, MRR, and AUC for the δ-threshold strategy.

respectively, and 7.99%, 13.97%, and 87.03% for the tissue-nonspecific one, respectively. Therefore, we use the tissue-specific PPI network as the middle level of our disease-protein-complex network.

3.5. Robustness to the Parameters Involved. There are three main parameters involved in our method: the weights of the disease-protein connections (α), the weights of the protein-complex connections (β), and the restart probability in

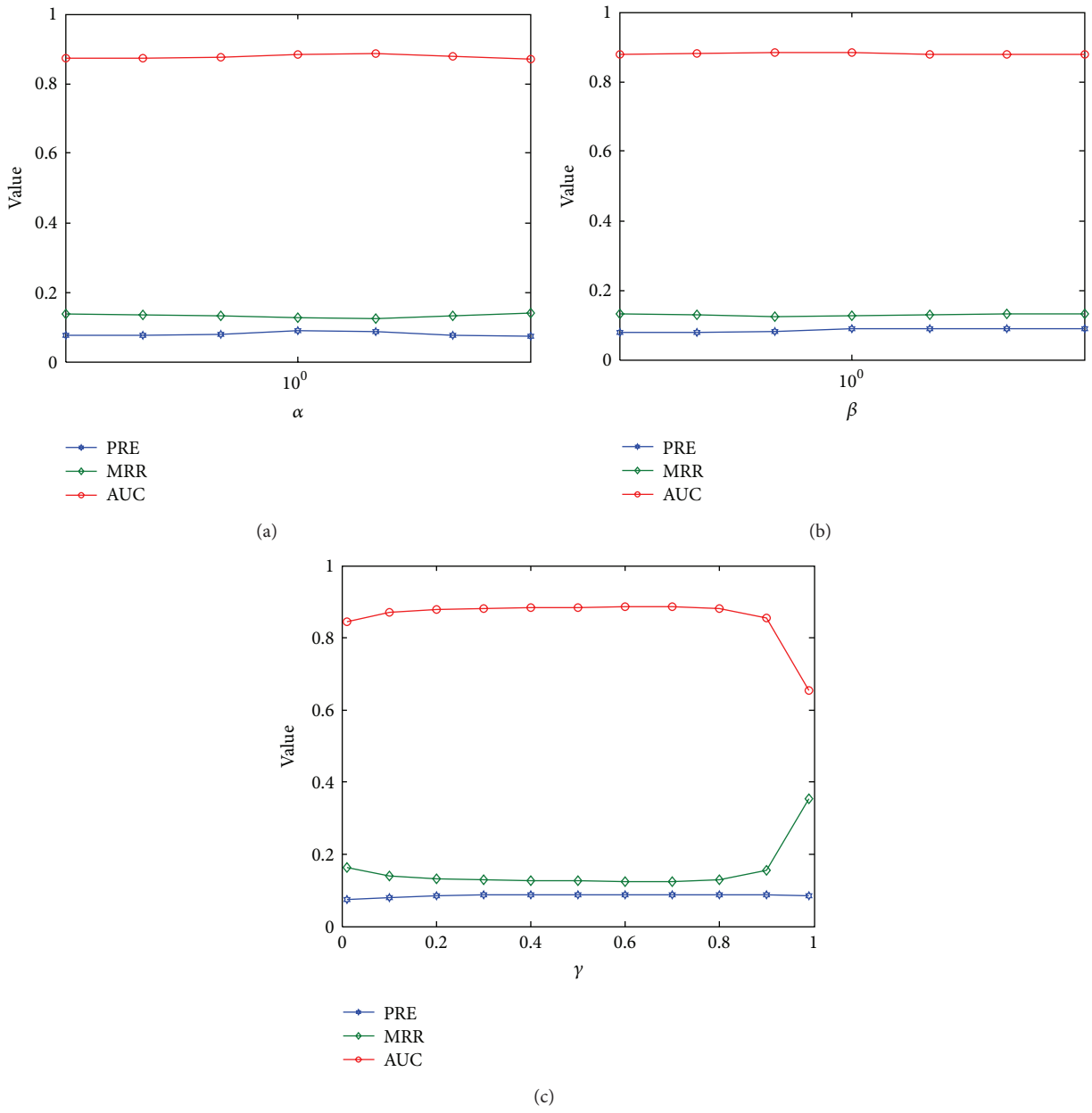


FIGURE 4: Influence of the parameters involved. (a) Influence of the weights of the disease-protein connections (α). (b) Influence of the weights of the protein-complex connections (β). (c) Influence of the restart probability (γ).

the random walk model (γ). To study the influence of these parameters on our method, we performed a comparative study on different values of these parameters and presented the results in Figure 4.

The weights of the disease-protein connections (α) determine the possibility of jumping from the disease layer to the protein layer and vice versa. With a large value of α, it is easier to travel between the two layers, while with a small value of α, it is harder to travel between the two layers. From Figure 4(a), we observe that our method is quite robust to this parameter. In a wide range of this parameter (10^{-3} to 10^3), all of the three criteria show only tiny fluctuations. For example, at the lower end of the spectrum ($\alpha = 10^{-3}$), the PRE, MRR, and AUC are 7.56%, 13.64%, and 87.38%, respectively, while at the higher end of the spectrum ($\alpha = 10^3$), the PRE, MRR, and AUC are 7.24%, 14.07%, and 86.95%, respectively. Moreover, at the optimum point ($\alpha = 1$), the PRE, MRR, and AUC are 8.84%, 12.59%, and 88.44%, respectively. From these observations, we conjecture that the selection of this parameter is not critical to the performance of our method. We hence use $\alpha = 1$ as the default value for this parameter.

Similarly, the weights of the protein-complex connections (β) determine the possibility of jumping from the protein layer to the complex layer and vice versa. With a large value

of β, it is easier to travel between the two layers, while with a small value of β, it is harder to travel between the two layers. From Figure 4(b), we observe that our method is also quite robust regarding this parameter. In a wide range of this parameter (10^{-3} to 10^{3}), all of the three criteria show only tiny fluctuations. For example, at one end of the spectrum ($\beta = 10^{-3}$), the PRE, MRR, and AUC are 7.88%, 13.29%, and 87.74%, respectively, while at the other end of the spectrum ($\beta = 10^{3}$), the PRE, MRR, and AUC are 8.84%, 13.26%, and 87.76%, respectively. Moreover, at the optimum point ($\beta = 10$), the PRE, MRR, and AUC are 9.05%, 13.05%, and 87.94%, respectively. From these observations, we conclude that the selection of this parameter is not critical to the performance of our method. Therefore, we use $\beta = 1$ as the default value for this parameter.

The restart probability (γ) determines the possibility of jumping from any node in the network back to the starting point of the query disease. With a large value of γ, a random walker cannot go far away from the starting point and thus will mainly explore neighbouring nodes of this point, while with a small value of γ, the random walker is able to explore areas far away from the starting query disease. From Figure 4(c), we observe that our method is robust regarding this parameter, except for extreme values. In a wide range of this parameter (0.1 to 0.8), all of the three criteria show only tiny fluctuations. For example, at one end of the spectrum ($\gamma = 0.1$), the PRE, MRR, and AUC are 8.09%, 13.92%, and 87.1%, respectively, while at the other end of the spectrum ($\gamma = 0.8$), the PRE, MRR, and AUC are 8.73%, 12.86%, and 88.18%, respectively. At the optimal point ($\gamma = 0.6$), the PRE, MRR, and AUC are 8.63%, 12.49%, and 88.54% respectively. Moreover, at the middle point of the spectrum ($\gamma = 0.5$), the PRE, MRR, and AUC are 8.84%, 12.59%, and 88.44%, respectively, not very different from the optimum point. From these observations, we conclude that the selection of this parameter is not critical to the performance of our method. Therefore, we seek for the simplicity to select $\gamma = 0.5$ as the default value for this parameter.

3.6. Robustness to the Network Structure. There are four types of connections in the heterogeneous network: edges between diseases, connecting diseases and proteins, between proteins, and connecting proteins and protein complexes. These connections determine the structure of the disease-protein-complex network. We then studied how the performance of our method changed with the addition or removal of a proportion of edges and presented the results in Figure 5.

From the figure, we see that our method is quite robust to the addition of edges. For example, when adding 10% edges between diseases into the network, the PRE, MRR, and AUC change from 8.84%, 12.59%, and 88.44% to 8.39%, 13.43%, and 87.59%, respectively. When adding other types of edges, we observe similar robust pattern. Particularly, the performance of our method is quite robust to the noise in the protein-protein interaction network, because the criteria only change slightly with the addition of this type of edges. These observations suggest the robustness of our method to false positive edges in the network.

Our method is also robust to the removal of edges. For example, when removing 10% edges connecting diseases and proteins from the network, the PRE, MRR, and AUC change from 8.84%, 12.59%, and 88.44% to 8.54%, 12.90%, and 88.14%, respectively. When removing 10% edges connecting proteins and protein complexes from the network, the PRE, MRR, and AUC change to 8.22%, 13.66%, and 87.36%, respectively. Again, the performance of our method is quite robust to the noise in the protein-protein interaction network, because the criteria only change slightly with the removal of this type of edges. These observations suggest that our method is also robust to false negative connections in the network.

3.7. Predicted Landscape of Associations between Diseases and Protein Complexes. With the performance and robustness of our method demonstrated, we further applied our method to a total of 926 diseases with tissue association information in our data set and predicted associations between these diseases and a total of 1,343 protein complexes. The lists of diseases, protein complexes, and the predicted score for each pair of disease and protein complexes are available for free downloading at our website http://bioinfo.au.tsinghua.edu.cn/jianglab/complex.

4. Conclusions and Discussion

In this paper, we have proposed a method for the identification of protein complexes that are related to a query disease via random walking on a heterogeneous network that is composed of a disease layer, a protein layer, and a protein complex layer. We have shown the high performance of our approach via a large-scale leave-one-out cross-validation experiment and have demonstrated the robustness of our approach to the parameters involved. As an application of our approach, we have predicted a landscape of associations between diseases and protein complexes.

Our method has the following advantages. First, in the disease layer, a disease is connected to its neighboring diseases with similar phenotype properties. Therefore, our method is capable of predicting associations for a query disease whose genetic basis is unknown by borrowing information from its neighboring diseases. Second, our method allows the inclusion of the recent discovery about the tissue specificity of protein-protein interactions, leading to high accuracy in making predictions. Finally, our method shows great robustness to the parameters involved, and hence it is easy to be adapted to the analysis of other data.

Certainly, our method can further be extended from the following directions. First, the disease similarity network plays a key role in our method. Besides the phenotype similarity profile derived from MeSH, there are also alternative profiles derived from the unified medical language system (UMLS) [32] and the human phenotype ontology (HPO) [33]. It has been shown that integrated use of these profiles provides a more comprehensive view of correlations in clinic properties of human diseases [34]. The way to integrate these

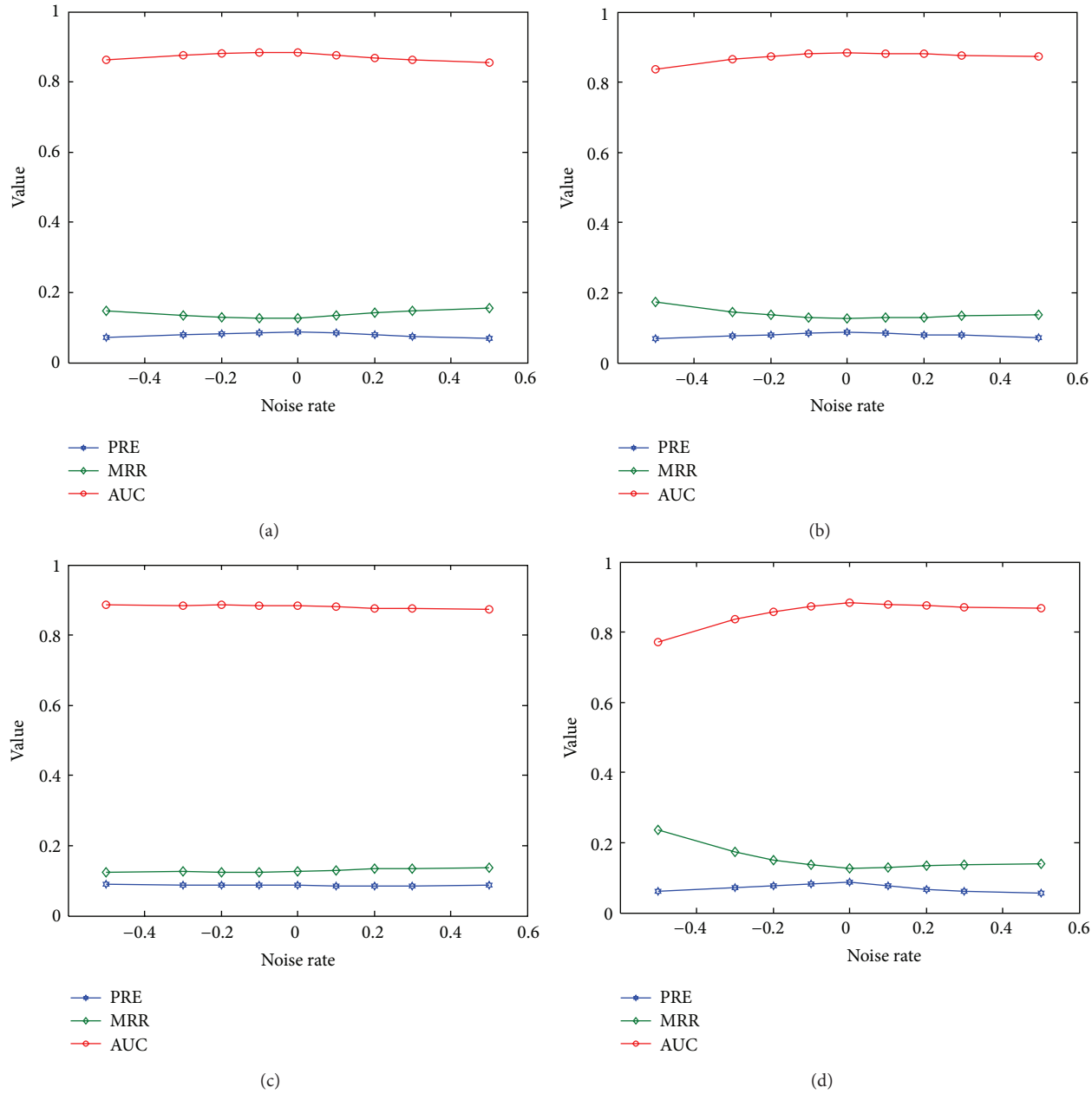


FIGURE 5: Influence of the addition or removal of edges. Results are the performance of our method with the addition (>0) or removal (<0) of a proportion of edges (a) between diseases, (b) connecting diseases and proteins, (c) between proteins, and (d) connecting proteins and protein complexes. All results are average of 5 independent runs.

similarity profiles in our current heterogeneous network will be a direction worth exploring.

Second, although the PPI network provides a systematic view of functional similarities between genes, such genomic information as transcriptional regulation, noncoding RNA regulation, functional annotation, pathway annotation, and structure domain annotation also provides useful assessments on functional similarities between genes. Integrating such genomic information with tissue-specific gene expression data to obtain a more comprehensive characterization of tissue-specific functional similarities between genes and further enhance the performance of our method will be one of our future research directions.

Third, protein complexes represent higher level functional units than proteins. Besides, gene modules such as pathways can be thought of as even higher level function units. Therefore, it also matters to pursue the goal of identifying pathways or gene modules that are associated with a given query disease. In technology, our method can be directly applied to solve this problem.

Finally, the predicted genome-wide landscape of associations between human diseases and protein complexes

provides a rich resource in understanding genetic bases of human inherited diseases. Using these prediction results to facilitate the analysis of prevalent genetic data such as single nucleotide polymorphisms identified in traditional genome-wide association studies or recent exome sequencing studies will also be a goal worth pursuing.

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), the National Natural Science Foundation of China (61175002), and the Open Research Fund of State Key Laboratory of Bioelectronics, Southeast University.

References

[1] T. A. Manolio, "Genomewide association studies and assessment of the risk of disease," *The New England Journal of Medicine*, vol. 363, no. 2, pp. 166–176, 2010.

[2] The Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–678, 2007.

[3] M. Choi, U. I. Scholl, W. Ji et al., "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 45, pp. 19096–19101, 2009.

[4] L. G. Biesecker, "Exome sequencing makes medical genomics a reality," *Nature Genetics*, vol. 42, no. 1, pp. 13–14, 2010.

[5] G. B. Ehret, P. B. Munroe, K. M. Rice et al., "Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk," *Nature*, vol. 478, no. 7367, pp. 103–109, 2011.

[6] Y. Bromberg, "Chapter 15: disease gene prioritization," *PLoS Computational Biology*, vol. 9, no. 4, Article ID e1002902, 2013.

[7] Y. Chen, W. S. Zhang, M. X. Gan, and R. Jiang, "Constructing human phenome-interactome networks for the prioritization of candidate genes," *Statistics and Its Interface*, vol. 5, no. 1, pp. 137–148, 2012.

[8] D. Altshuler, M. Daly, and L. Kruglyak, "Guilt by association," *Nature Genetics*, vol. 26, no. 2, pp. 135–138, 2000.

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

[10] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Molecular Systems Biology*, vol. 4, no. 1, article 189, 2008.

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

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

[13] X. Wu, Q. Liu, and R. Jiang, "Align human interactome with phenome to identify causative genes and networks underlying disease families," *Bioinformatics*, vol. 25, no. 1, pp. 98–104, 2009.

[14] Y. Chen, T. Jiang, and R. Jiang, "Uncover disease genes by maximizing information flow in the phenome-interactome network," *Bioinformatics*, vol. 27, no. 13, Article ID btr213, pp. i167–i176, 2011.

[15] K. Lage, E. O. Karlberg, Z. M. Størling et al., "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nature Biotechnology*, vol. 25, no. 3, pp. 309–316, 2007.

[16] R. Jiang, M. X. Gan, and P. He, "Constructing a gene semantic similarity network for the inference of disease genes," *BMC Systems Biology*, vol. 5, supplement 2, article S2, 2011.

[17] W. Zhang, F. Sun, and R. Jiang, "Integrating multiple protein-protein interaction networks to prioritize disease genes: a bayesian regression approach," *BMC Bioinformatics*, vol. 12, no. 1, article S11, 2011.

[18] O. Magger, Y. Y. Waldman, E. Ruppin, and R. Sharan, "Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks," *PLoS Computational Biology*, vol. 8, no. 9, Article ID e1002690, 2012.

[19] Y. Guan, D. Gorenshteyn, M. Burmeister et al., "Tissue-specific functional networks for prioritizing phenotype and disease genes," *PLoS Computational Biology*, vol. 8, no. 9, Article ID e1002694, 2012.

[20] B. Jiang, J. Wang, J. Xiao, and Y. Wang, "Gene prioritization for type 2 diabetes in tissue-specific protein interaction networks," *Lecture Notes in Operations Research*, vol. 11, pp. 319–328, 2009.

[21] A. D. D'Andrea, "The fanconi anemia/BRCA signaling pathway: disruption in cisplatin-sensitive ovarian cancers," *Cell Cycle*, vol. 2, no. 4, pp. 290–292, 2003.

[22] P. Yang, X. Li, M. Wu, C.-K. Kwoh, and S.-K. Ng, "Inferring gene-phenotype associations via global protein complex network propagation," *PLoS ONE*, vol. 6, no. 7, Article ID e21502, 2011.

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

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

[25] A. I. Su, T. Wiltshire, S. Batalov et al., "A gene atlas of the mouse and human protein-encoding transcriptomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 16, pp. 6062–6067, 2004.

[26] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach et al., "CORUM: the comprehensive resource of mammalian protein complexes," *Nucleic Acids Research*, vol. 36, supplement 1, pp. D646–D650, 2008.

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

[28] K. Lage, N. T. Hansena, E. O. Karlberg et al., "A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 52, pp. 20870–20875, 2008.

[29] R. Rentzsch and C. A. Orengo, "Protein function prediction—the power of multiplicity," *Trends in Biotechnology*, vol. 27, no. 4, pp. 210–219, 2009.

[30] H. J. Lowe and G. O. Barnett, "Understanding and using the Medical Subject Headings (MeSH) vocabulary to perform literature searches," *Journal of the American Medical Association*, vol. 271, no. 14, pp. 1103–1108, 1994.

[31] D. Smedley, S. Haider, B. Ballester et al., "BioMart—biological queries made easy," *BMC Genomics*, vol. 10, no. 1, article 22, 2009.

[32] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, supplement 1, pp. D267–D270, 2004.

[33] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The human phenotype ontology: a tool for annotating and analyzing human hereditary disease," *American Journal of Human Genetics*, vol. 83, no. 5, pp. 610–615, 2008.

[34] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," *Molecular Systems Biology*, vol. 7, no. 1, article 496, 2011.