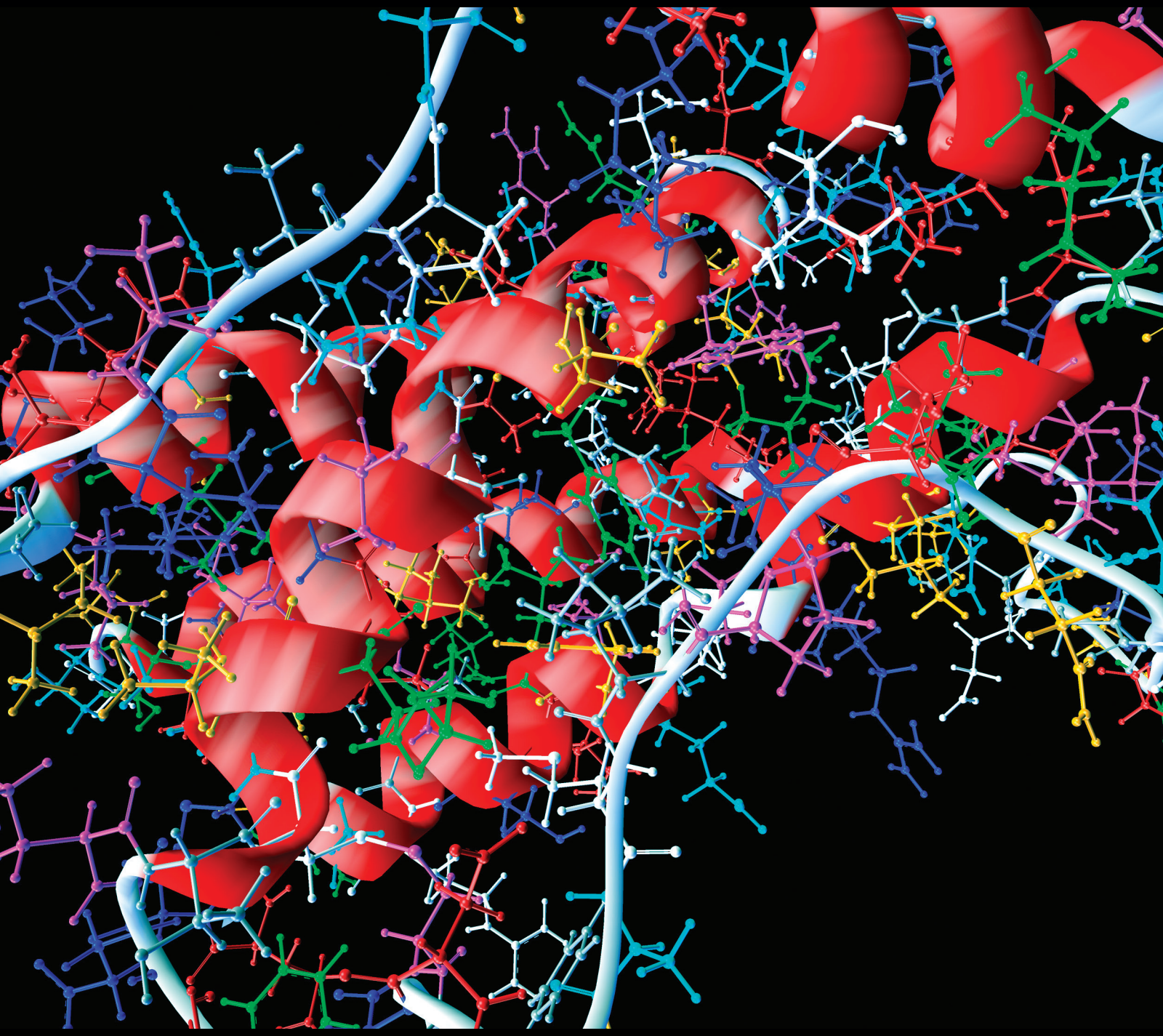
Computational and Mathematical Methods in Medicine

Machine Learning and Network Methods for Biology and Medicine

Guest Editors: Lei Chen, Tao Huang, Chuan Lu, Lin Lu, and Dandan Li



Machine Learning and Network Methods for Biology and Medicine

Computational and Mathematical Methods in Medicine

Machine Learning and Network Methods for Biology and Medicine

Guest Editors: Lei Chen, Tao Huang, Chuan Lu, Lin Lu, and Dandan Li

Editorial Board

Contents

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 915124, 2 pages
http://dx.doi.org/10.1155/2015/915124

Editorial

Machine Learning and Network Methods for Biology and Medicine

Lei Chen,[1] Tao Huang,[2,3] Chuan Lu,[4] Lin Lu,[5] and Dandan Li[6]

[1]*College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*
[2]*Department of Genetics and Genomics Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA*
[3]*Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[4]*Department of Computer Science, Aberystwyth University, Aberystwyth, Ceredigion SY23 3DB, UK*
[5]*Department of Radiology, Columbia University Medical Center, New York, NY 10032, USA*
[6]*Gastrointestinal Medical Department, China-Japan Union Hospital of Jilin University, Changchun 130033, China*

Correspondence should be addressed to Lei Chen; chen_lei1@163.com

Received 12 October 2015; Accepted 12 October 2015

In recent years, many computational methods have been proposed to tackle the problems that arise in analyzing various large-scale high dimensional data in biology and medicine. Useful techniques have been developed by the use of conventional statistical modeling and analysis and have helped to reveal many biological mechanisms. However, with the rapid development of high throughput technologies, biological and medical data generated nowadays are becoming increasingly more heterogeneous and complex. It is therefore necessary to develop more effective and efficient approaches to analyzing such data, requiring more powerful methods like advanced machine learning algorithms and network based methods.

In this special issue, eighteen novel investigations are presented, including a number of newly proposed techniques for up-to-date data analysis and application systems for interesting biological and medical problems.

A computational method was proposed by B. Wang et al. to identify novel candidate genes related to apoptosis. This method first applied shortest path algorithm in a large protein-protein interaction network to search new candidate genes and then the candidate genes were filtered by a permutation test. Twenty-six genes were obtained and analyzed regarding their likelihood of being novel apoptosis-related genes.

F. Yuan et al. proposed a computational method to identify new candidate genes and chemicals based on currently known genes and chemicals related to prostate cancer by applying shortest path approach in a hybrid network which was constructed according to information concerning chemical-chemical interactions, chemical-protein interactions, and protein-protein interactions.

B. Sun et al. designed an analysis pipeline to study the relationships between eight types of damaging protein posttranslational modifications (PTM) and a few human inherited diseases and cancers. The results suggested that some human inherited diseases or cancers might be related to the interactions of damaging PTMs.

Y. Zhan et al. identified a five-gene signature that predicts prognosis in patients with kidney renal clear cell carcinoma (KIRC). The RNA expression data from RNA-sequencing and clinical information of 523 KIRC patients were analyzed. The AUC (area under ROC curve) of the five-gene signature was 0.783 which showed high sensitivity and specificity.

Z. Ji et al. developed a Nonnegative Matrix Factorization (NMF) based feature selection approach (NMFBFS) to identify potential clinical symptoms for HCC patient stratification. The results on 407 HCC patient samples with 57 symptoms showed the effectiveness of the NMFBFS approach in identifying important clinical features, which will be very helpful for HCC diagnosis.

C. Zhang et al. proposed adaptive weight regularized ADSIR for low dose CT reconstruction. Three numerical experiments are carried out for evaluation and comparisons are made with other algorithms.

J. I. Galván-Tejada et al. presented the potential of X-ray based multivariate prognostic models to predict the onset of chronic knee pain. Using X-rays quantitative image-assessments, multivariate models may be used to predict subjects that are at risk of developing knee pain by osteoarthritis.

Y. Cui et al. developed a method called ROC-Boosting to select significant Haar-like features extracted from tongue images for health identification. They analyzed the images of 1,322 tongue cases and selected features focused on the root, top, and side areas of the tongue which can classify the healthy and ill cases.

S. Wang et al. proposed a novel automatic approach for dendritic spine identification in neuron image. The method integrated wavelet based conditional symmetric analysis and regularized morphological shared-weight neural networks. Its good performance and the comparison with existing methods suggest the utility of the method.

S. Yang et al. proposed the use of a combination of edgeR and DESeq to analyze miRNA sequencing data with a large sample size.

R. Hu et al. proposed an automated resource provisioning method, G2LC, for bioinformatics applications in IaaS. It guaranteed applications performance and improved resource utilization. Evaluated on real sequence searching data of BLAST, G2LC saved up to 20.14% of resource.

R. Hu and C. Li proposed an improved PID algorithm based on insulin-on-board estimate using a combinational mathematical model of the dynamics of blood glucose-insulin regulation in the blood system. The simulation results demonstrated that the improved PID algorithm can perform well in different carbohydrate ingestion and different insulin sensitivity situations. Compared with the traditional PID algorithm, the control performance was improved obviously and hypoglycemia can be avoided.

J. G. Rodriguez-Escobedo et al. described the use of the "a priori" algorithm at resolving KIR gene patterns associated with haematological malignancies, previously unrevealed through traditional statistical approaches.

Z. Jiang et al. built a new method to predict chemical toxicities based on ontology information of chemicals. This method was more effective than previous method and provided new insights to study chemical toxicity and other attributes of chemicals.

L. Yuan et al. explored the hidden relationship between miRNAs and imprinted genes in cell pluripotency. They found that the neighbors of imprinted genes on molecular network were enriched in modules such as cancer, cell death and survival, and tumor morphology. The imprinted region may provide a new look for those who are interested in cell pluripotency of hiPSCs and hESCs.

T. Liu et al. reviewed the recent discoveries and advance in the field of evolutional developmental biology in light of the development in large-scale omics studies.

J. A. Vanegas et al. presented a survey on the state-of-the-art text mining approaches to extraction of biomolecular events, which are useful for understanding the underlying biological mechanisms. The popular natural language processing and machine learning methods and tools have been analyzed for this task of phases varied from feature extraction, trigger/edge detection to postprocessing.

Z. Zeng et al. surveyed natural language processing techniques in bioinformatics. First, they searched for knowledge on biology and retrieved references using text mining methods and reconstructed databases. Then, they analyzed the applications of text mining and natural language processing techniques in bioinformatics. Finally, numerous methods and applications are discussed for future use by text mining and natural language processing researchers.

In summary, this special issue collects a number of innovative studies that address various challenging issues in analyzing data in biology and medicine. We hope that this publication will become a landmark in the international development of the relevant literature and also will help encourage more researchers and practitioners to be engaged in this ever increasingly important field.

Lei Chen
Tao Huang
Chuan Lu
Lin Lu
Dandan Li

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 454076, 12 pages
http://dx.doi.org/10.1155/2015/454076

Research Article

Detection of Dendritic Spines Using Wavelet-Based Conditional Symmetric Analysis and Regularized Morphological Shared-Weight Neural Networks

Shuihua Wang,[1,2] Mengmeng Chen,[3,4,5] Yang Li,[1] Yudong Zhang,[2,6] Liangxiu Han,[7] Jane Wu,[3,4] and Sidan Du[1]

[1]*Department of Electronic Engineering, Nanjing University, Nanjing 210024, China*
[2]*School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, China*
[3]*State Key Laboratory of Brain and Cognitive Science, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China*
[4]*Department of Neurology, Lurie Cancer Center, Center for Genetic Medicine, Northwestern University School of Medicine, Chicago, IL 60611, USA*
[5]*University of Chinese Academy of Sciences, Beijing 100101, China*
[6]*Translational Imaging Division, Columbia University, New York, NY 10032, USA*
[7]*School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Manchester M1 5GD, UK*

Correspondence should be addressed to Sidan Du; coff128@nju.edu.cn

Received 17 June 2015; Revised 2 September 2015; Accepted 27 September 2015

Academic Editor: Valeri Makarov

Identification and detection of dendritic spines in neuron images are of high interest in diagnosis and treatment of neurological and psychiatric disorders (e.g., Alzheimer's disease, Parkinson's diseases, and autism). In this paper, we have proposed a novel automatic approach using wavelet-based conditional symmetric analysis and regularized morphological shared-weight neural networks (RMSNN) for dendritic spine identification involving the following steps: backbone extraction, localization of dendritic spines, and classification. First, a new algorithm based on wavelet transform and conditional symmetric analysis has been developed to extract backbone and locate the dendrite boundary. Then, the RMSNN has been proposed to classify the spines into three predefined categories (mushroom, thin, and stubby). We have compared our proposed approach against the existing methods. The experimental result demonstrates that the proposed approach can accurately locate the dendrite and accurately classify the spines into three categories with the accuracy of 99.1% for "mushroom" spines, 97.6% for "stubby" spines, and 98.6% for "thin" spines.

1. Introduction

Dendritic spines are small "doorknob" shaped extensions from neuron's dendrites, which can number thousands to a single neuron. Spines are typically classified into three types based on the shape information: mushroom, stubby, and thin. "Mushroom" spine has a bulbous head with a thin neck; "stubby" spine only has a bulbous head; "thin" spine has a long thin neck with a small head. Research has shown that the changes in shape, length, and size of dendritic spines are closely linked with neurological and psychiatric disorders, such as attention-deficit hyperactivity disorder (ADHD), autism, intellectual disability, Alzheimer's disease, and Parkinson's disease [1–5]. Therefore, the morphology analysis and identification of structure of dendritic spines are critical for diagnosis and further treatment of these diseases [6, 7].

Traditional manual detection approach of dendritic spines detection is costly and time consuming and prone to error due to human subjectiveness. With the recent advances in biomedical imaging, computer-aided semiautomatic or automatic approaches to detect dendritic spines based on

image analysis have shown the efficacy. SynD method proposed by Schmitz et al. [8] is a semiautomatic image analysis routine to analyze dendrite and synapse characteristics in immune-fluorescence images. For the fluorescence imaging, the neurite and soma were captured in the separated imaging channels. In that case, soma and synapse were detected without intervention from neurite [9–11] based on the channel information. However, this method cannot be extended to the images, of which the information is captured in the same channel. Therefore, many other methods were proposed to solve this problem, for instance, ImageJ [12], NeuronStudio [13], NeuronJ [14], and NeuronIQ [15]. However, these methods have some limitations. For example, NeuronIQ was designed for the confocal multiphoton laser scanning. NeuronJ was used to trace the dendrite growing in the condition of manually marking the dendrite first. Koh et al. detected spines from stacks of image data obtained by laser scanning microscopy [16]. The algorithm first extracted the dendrite backbone defined as the medial axis and then geometric information was employed to detect the attached and detached spines according to the shape of each candidate spine region. Features including spine length, volume, density, and shape for static and time-lapse images of hippocampal pyramidal neurons were used as key points for the detection. The disadvantage of this method is that it might lose many spines during the detection because of the thresholding method used in this case. To overcome this problem, Xu et al. proposed a new detection algorithm for the attached spines from the dendrites by two grassfire steps [17]: a global threshold was chosen to segment the image and then the medial axis transform (MAT) was applied to find the centerlines of the dendrites. Then some large spines (noncenterlines) were removed from the centerlines. After the backbone was extracted, two grassfire procedures were applied to separate the spine and dendrite. The results of the proposed method were similar to the results of the manual method. Cheng et al. proposed a method using an adaptive threshold based on the local contrast to determine the foreground, containing the spine and dendrite, and detect attached and detached spines [18]. Fan et al. used the curvilinear structure detector to find the medial axis of the dendrite backbone and spines attached to the backbone [19]. To locate the boundary of dendrite, an adaptive local binary fitting (aLBF) energy level set model was proposed for localization. Zhang et al. extracted the boundaries and the centerlines of the dendrite by estimating the second-order directional derivatives for both the dendritic backbones and spines [20]. Then a classifier based on Linear Discriminate Analysis (LDA) was built to classify the attached spines into true and false types. The accuracy of the algorithm was calculated according to the backbone length, spine number, spine length, and spine density. Janoos et al. used the medial geodesic to extract the centerlines of the dendritic backbone [21]. He et al. proposed a method based on NDE to classify the dendrite and spines [22]. The principle of their method was that spine and dendrite had different shrink rates. Shi et al. proposed a wavelet-based supervised method for classifying 3D dendritic spines from neuron images [23].

Existing work is encouraging. However, the problems remain on how to improve accuracy (e.g., accurate extraction of backbone, accurate detection of attached and detached spines). Different from existing approaches, in this paper, we have proposed new algorithms for efficient detection of dendritic spines using wavelet-based conditional symmetric analysis and regularized morphological shared-weight neural network. Our contributions include the following:

(1) A new extraction model for dendrite backbone and its boundary localization using wavelet-based conditional symmetric analysis and pixel intensity difference, which can allow accurate extraction of backbone, the first important step for dendritic spines.

(2) A new way for spine detection based on regularized morphological shared-weight neural networks (RMSNN) to efficiently detect spines and classify them into right categories, that is, mushroom, thin, and stubby.

The rest of this paper is organized as follows. Section 2 describes the proposed methods including wavelet-based conditional symmetry analysis and pixel intensity difference for the dendrite detection and localization and regularized shared-weight neural networks for the spine detection. In Section 3, we have conducted experimental evaluation and demonstrated the effectiveness of the proposed algorithm. Section 4 discusses the results. Section 5 concludes the proposed approach and highlights the future work.

2. Methods

Figure 1 shows the steps of our proposed approach to dendritic spines. In the image acquisition phase, we demonstrated the process for the neuron culture, label, and imaging. In the second step, we preprocessed the images by reducing the noise and smoothing the background [24, 25]. Then, we extracted the dendrite backbone based on the conditional symmetric analysis and located the dendrite boundary based on the difference of the pixel intensity. Afterwards, the spines were detected, classified, and characterized by RMSNN.

2.1. Image Acquisition. The neurons used for imaging in this paper were cortical neurons, primary cultured from Embryonic 18th- (E18-) day rat and next cultured until the 22nd day in vitro. Then, the neurons were transfected by Lipofectamine 2000 and imaged at the 24th day by Leica SP5 confocal laser scanning microscopy (CLSM) by 63x. The size of the image is 1024×1024, and the resolution is 0.24 um/pixel at the confocal layer. The images used for the morphology analysis were obtained by the maximum intensity projection (MIP) of the original 3D image stack. As the images were captured as Z-stack series, we projected the 3D image stack onto the xy, yz, and zx planes, respectively. Since the slices along the optical direction (z) provided very limited information and the computation time based on the 3D image stacks is highly increased, it was desired to consider only the 2D projection onto the xy plane. The 2D image used for analysis was a maximum intensity projection of

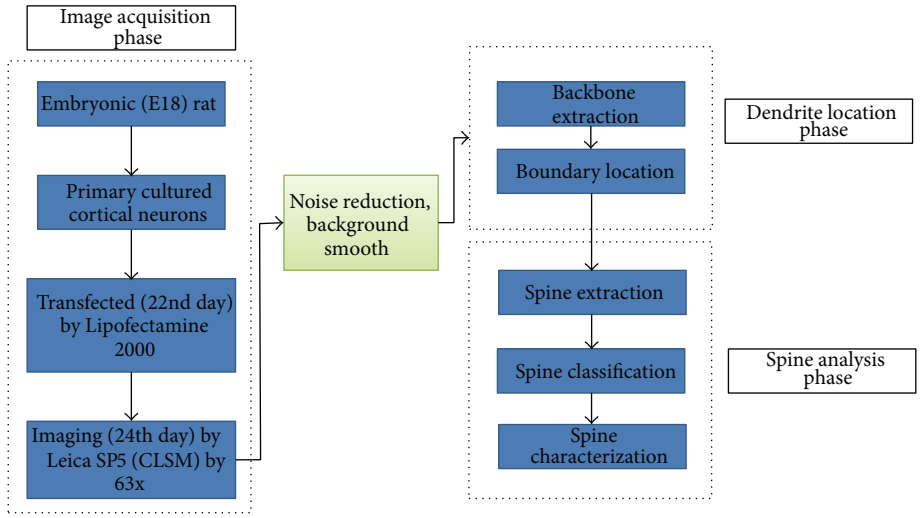


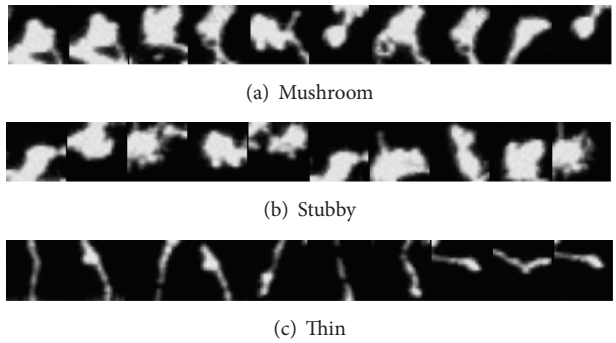
Figure 1: Flowchart of the proposed detection method of the dendritic spines.

the original 3D stack. It was obtained by projecting in the xy plane the voxels with maximum intensity values that fall in the way of parallel rays traced from the viewpoint to the plane of projection.

We randomly selected 15 different images from Leica SP5 confocal laser scanning microscopy to form the spines library to test our algorithm. All images contain distinct spines including mushroom, stubby, and thin types. The typical size of the image is 1024×1024. Most spines in the images are within a rectangle of 20×20 in pixel, but the "thin" spine is within an about 5×20 rectangle in pixel. The spines have variable gray-level intensities. Spines collected from the image library were employed to build an image base library. Spine subimages in the library were taken as samples to test the classification accuracy of RMSNN. In order to cover as many cases as possible, the image base library contains distinct sizes and spines with different orientations.

In order to build the golden-standard spine library, five experts in the neuroscience field were employed to manually mark the spines in the collected images and classify the spines into three predefined categories including "mushroom," "stubby," and "thin" types. For the conflict of the manual marking, the minority was supposed to be subordinated to the major. Then according to the marked spines, we computed the maximum width, length, area, and the center point. The randomly selected image base library contains about 2700 subimage samples, 900 for each type of spines. Figure 2 shows some image samples in our image base library. As we can see from the image sample, spines of "mushroom" type contain a thin neck and head, the stubby type connects directly with the dendrite without neck, and the thin type is with the smallest size with only a thin neck and without head.

2.2. Image Preprocessing. Considering the limitation of imaging technique, we have employed the 2D median filter to deal with the noise introduced by the imaging mechanism of the photomultiplier tubes (PMT) and then used the partial differential equation (PDE) proposed by Wang et al. [26] to enhance the image. Figure 3 shows an example of the original image and the preprocessed result.



(a) Mushroom

(b) Stubby

(c) Thin

Figure 2: Samples of the subimages used in the image library.

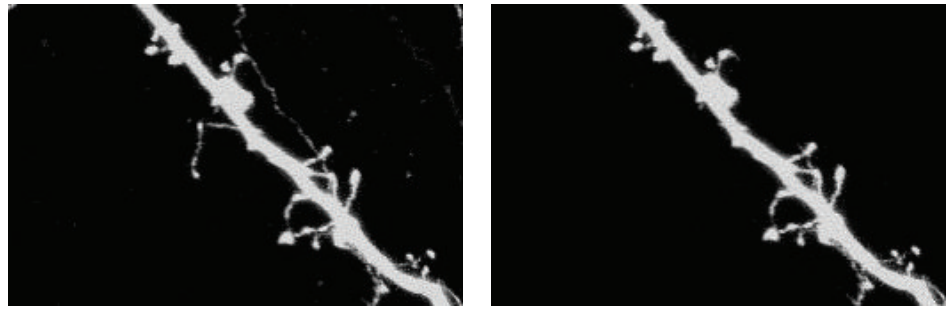
2.3. Backbone Extraction Using the Wavelet Transformation Based Conditional Symmetric Analysis. Considering the attached spines, it is necessary to firstly locate the dendrites in order to segment the spines from the dendrite. The backbone extraction and boundary localization are critical for dendritic spine classification and analysis, which include the following steps.

Step 1. Remove the noise and small isolated point-set.

Step 2. Locate the backbone of the dendrite.

Step 3. Locate the boundary of the dendrite.

The backbone is defined as the thinning of the dendrite. Due to the variance of width of dendrite, attached and detached spines, it is a challenging task to locate the boundary



(a) Original image (b) Preprocessed image

Figure 3: An example of preprocessed image.

of the dendrite directly from the preprocessed images. Therefore, we have developed a new extraction model utilizing wavelet transform based conditional symmetric analysis. The essence of this model is to conduct a local conditional symmetry analysis of the contour of the region of interest (ROI) and then compute the center points to produce the backbone of the dendrite.

Due to the complexity of the dendrites and dendrite spines' distribution, we have employed morphological operation to remove the small isolated point-set for the dendrite in the binary image obtained by local Otsu [27–29] via (1), which could decrease the disconnection rate of the dendrite detection:

$$P = \begin{cases} 1, & \text{more than } n \text{ positive pixels in its 3-by-3 window,} \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

in which n is the threshold of the number of positive pixels. The value of n could be determined by trial and error method and means that the pixel belongs to the major line if there are more than n positive pixels in its 3×3 neighborhood window. Otherwise, the value of the pixel is forced to be 0, treated as the small isolated point-set. The determination of the centerline of the dendrite is based on the conditional symmetric analysis.

The symmetric analysis was accomplished via the wavelet transform. We have applied the wavelet transform to detect a pair of contour curves:

$$\varphi_x(x,y) = \frac{\partial}{\partial x}\theta(x,y) = \phi'\left(\sqrt{x^2+y^2}\right)\frac{x}{\sqrt{x^2+y^2}},$$
$$\varphi_y(x,y) = \frac{\partial}{\partial y}\theta(x,y) = \phi'\left(\sqrt{x^2+y^2}\right)\frac{y}{\sqrt{x^2+y^2}}, \quad (2)$$

in which x and y stand for the coordinate of the contour curve. $\varphi_x(x,y)$ means the partial derivative of x and $\varphi_y(x,y)$ stands for the partial derivative of y, respectively. $\theta(x,y)$ is a low pass filter.

For $\varphi_x(x,y)$ and $\varphi_y(x,y)$, the scale wavelet transform (WT) could be written as the following equations:

$$W_{x,s}f(x,y) = (f * \varphi_{x,s})(x,y) = s\frac{\partial}{\partial x}(f * \theta_s)(x,y),$$
$$W_{y,s}f(x,y) = \left(f * \varphi_{y,s}\right)(x,y) = s\frac{\partial}{\partial y}(f * \theta_s)(x,y). \quad (3)$$

Here, $\theta_s = (1/s^2)\theta(x/s, y/s)$. We can get the modulus of the gradient vector as

$$\nabla W_s f(x,y) = \begin{pmatrix} W_{x,s}f(x,y) \\ W_{y,s}f(x,y) \end{pmatrix}, \quad (4)$$

$$\left|\nabla W_s f(x,y)\right| = \sqrt{\left|W_{x,s}f(x,y)\right|^2 + \left|W_{y,s}f(x,y)\right|^2}, \quad (5)$$

$$A_s f(x,y) = \arctan\left(\frac{W_{y,s}f(x,y)}{W_{x,s}f(x,y)}\right), \quad (6)$$

where ∇ is the gradient vector and the gradient direction is given as (6). The contour points (x,y) are the local maxima of $|\nabla W_s f(x,y)|$ in the direction of $A_s f(x,y)$ at scale s. However, the local maxima modulus is not the exact edge point.

We selected (7) as the basis function. We set $\varphi^-(x) = -\varphi^+(-x)$ and had $\varphi(x) = \varphi^+(x) + \varphi^-(x)$ as the wavelet function, which had the following properties: gray invariant, slope invariant, width invariant, and symmetric [29, 30]. The advantage is to make the extraction of a pair of contours with accurate protrusions. Consider

$$\varphi^{+} = \begin{cases} \frac{2}{\pi}\left(4x\ln\frac{\left(1-8x^2+2\sqrt{1-16x^2}\right)\left(1+\sqrt{1-x^2}\right)}{9x-8x^2+3\sqrt{9-16x^2}} - \frac{1}{2x}\left(\sqrt{1-16x^2}-3\sqrt{9-16x^2}+8\sqrt{1-x^2}\right)\right), & x\in\left(0,\frac{1}{4}\right) \\ \frac{2}{\pi}\left(4x\ln\frac{8x\left(1+\sqrt{1-x^2}\right)}{9+3\sqrt{9-16x^2}} - \frac{1}{2x}\left(3\sqrt{9-16x^2}-8\sqrt{1-x^2}\right)\right), & x\in\left[\frac{1}{4},\frac{3}{4}\right) \\ \frac{2}{\pi}\left(4x\ln\frac{1+\sqrt{1-x^2}}{x} - \frac{4}{x}\sqrt{1-x^2}\right), & x\in\left[\frac{3}{4},1\right) \\ 0, & x\in[1,\infty). \end{cases} \quad (7)$$

The distance between two symmetric points is equal to the scale of the wavelet transform. If the distance between two symmetric points is larger than or equal to the width of regular region, the center point of the symmetric pair can potentially be located outside of the dendrite. The regular region is defined as the dendrite is smooth, where the function has a stable variation along the axis. Thus, we defined the stable symmetry as follows.

If the scale of wavelet transform is larger than or equal to the width of regular region, the modulus maxima points generate two new parallel contours inside the periphery of the dendrite. All the symmetric pairs of the wavelet transforms that do not have a counterpart are defined as the unstable symmetry. In this case, we have considered the width as the constraint condition. In the direction of the perpendicular to the gradient direction, we selected the width nearest to the regular region.

The center of every symmetric pair located on the centerline of the original regular region of the stroke point. Finally, the backbone of the regular region was defined by the curve of all connected symmetric points.

2.4. Boundary Location Based on the Pixel Intensity Difference. The morphological operation of removing noise blurred the boundary. Therefore, after localization of backbone, the boundary of the dendrite was detected via varies of the pixel intensity of the preprocessed image from Section 2.2. We can observe that the pixel intensity of the line pixel changes abruptly at the boundary locations. The boundary location was performed in two steps. In the first step, we have searched the image along the two directions perpendicular to the local line direction until the pixel intensity of the line pixel changed sharply. We set a threshold for each pixel. The local line direction is determined as

$$A_s f(x,y) = \arctan\left(\frac{W_{y,s} f(x,y)}{W_{x,s} f(x,y)}\right). \quad (8)$$

The formulation of each pixel is given by $(\alpha, I(p))$, in which $I(p)$ is the pixel intensity of point p in the original image and α is a predefined pixel intensity value, that is,

$$\text{if } \begin{cases} I(p) \geq \alpha, & p \text{ belongs to the line pixel} \\ I(p) < \alpha, & p \text{ does not belong to the line pixel.} \end{cases} \quad (9)$$

In the second step, some boundary points that were not on the searching path could be missed. The missed boundary points were detected from the neighboring boundary points. Provided that there are two known boundary points, if they are adjacent, there were no other boundary points between them; otherwise, the method proposed by Tang and You [31] was used to find the missed points, which can link the two points into a discrete line with one point as the starting point and the other one as the ending point.

There are several advantages of our proposed algorithms for backbone detection and boundary location. (1) The first are computing efficiency and noise reduction. Our approach uses less computing time than the method based on the derivatives of the Gaussian kernel and is more robust when dealing with the noise. (2) Meanwhile, it reduces the error rate for misclassifying spine pixels as dendrite pixels and sharply reduces the disconnection rate, which means our approach is more robust when dealing with the disturbance information than other methods, such as NDE proposed by He et al. [22].

2.5. Spine Detection Based on Regularized Morphological Shared-Weight Neural Network (RMSNN). Considering the dendritic spine's structure, we have employed the regularized morphological shared-weight neural networks for the detection and classification of spines. The regularized morphological shared-weight neural networks consist of two-phase heterogeneous neural networks in series as shown in Figure 4: the first phase is for feature extraction and the second phase is for classification. In the first phase, it is accomplished via the gray-scale Hit-Miss transform. The feature extraction phase has multiple feature extraction layers. Each layer is composed of one or more feature maps. Each feature map is generated by the Hit-Miss transform with a pair of structure elements (SEs) from the previous layer and is accompanied by a new pair of SEs, in which one is for the erosion and the other one is for the dilation. In the classification stage, it shows a fully connected Feedforward Neural Network (FNN) [32–34]. The input of FNN is the direct output of the feature extraction stage. The output of the classification stage is a three-node layer, in which each node stands for one type of spine. Figure 4 shows the structure of the morphological shared-weight neural network (MSNN) [35]. The MSNN has been widely applied in the following research fields,

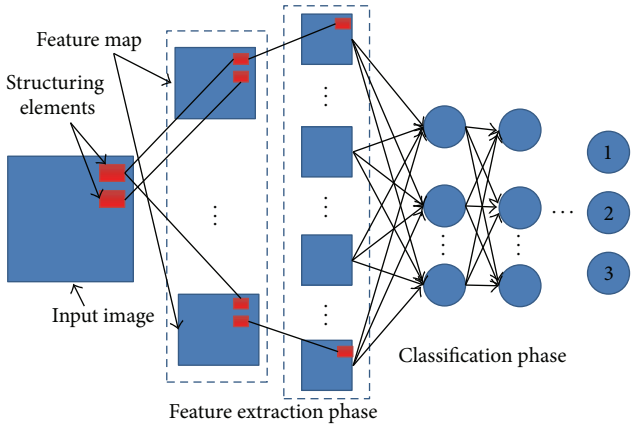


Figure 4: Structure of morphological shared-weight neural network.

including laser radar (LADAR), forward-looking infrared (FLIR), synthetic aperture radar, and visual spectrum image. The existing research demonstrates that the MSNN is robust for detection with rotation, image intensity translation, and occlusion variables [36]. In this paper, we have proposed to apply the regularized morphological shared-weight neural network to spine classification.

Dilation is defined as

$$A \oplus B = \left\{x \mid \left(\widehat{B}\right)_x \cap A \neq \emptyset\right\}, \tag{10}$$

in which A and B are sets in Z^2 and $\widehat{B}$ is the reflection of B. $\emptyset$ is the empty set. Equation (10) is termed the dilation of A by SE B. Dilation is the reflection of B about its origin, then translated by x, with the set of all x, which allow $\widehat{B}$ to intersect A with at least one element.

Erosion is defined as (11) or (12) by the duality of the erosion-dilation relationship:

$$A \ominus B = \left\{x \mid (B)_x \subseteq A\right\}, \tag{11}$$

$$A \ominus B = \left(A^c \oplus \widehat{B}\right)^c, \tag{12}$$

in which A^c is defined as the complement of A.

Hit-Miss transform is defined as an operation that detects a given pattern in a binary image based on a pair of disjoint structure elements, one for Hit and the other one for Miss. The result of the Hit-Miss transform is a set of positions, where the first SE fits in the foreground of the input image and the second SE misses it completely:

$$A \otimes B = (A \ominus X) \cap \left(A^c (W - X)\right), \tag{13}$$

in which X is a SE that consisted from set B, W is an enclosing window of X, and $(W - X)$ is the local background of X. By supposing X as H, the Hit SE, and $(W - X)$ as M, the Miss SE, we can get

$$A \otimes B = (A \ominus H) \cap \left(A^c \ominus M\right), \tag{14}$$

in which $B = (H, M)$ and it can be written as

$$A \otimes B = (A \ominus H) - \left(A^c \oplus \widehat{M}\right). \tag{15}$$

As far as the gray scale is concerned, we assume the image function as $I = f(x, y)$, in which $f(x, y)$ was the intensity value of the point (x, y). Meanwhile, we made the SE $b(x, y)$. The morphological operation can be thought of as a 3D binary set by way of the umbra transform. The umbra of a 3D surface function is defined as

$$U(f) = \left\{(x, y, z) \mid (x, y) \in D_f,\ z \leq f(x, y)\right\}, \tag{16}$$

where we take D_f as the domain of f. Then the gray scale dilation can be defined as

$$\begin{aligned}(f \oplus b)(s, t) = \max \{&f(s - x, t - y) \\ &+ b(x, y) \mid (s - x), (t - y) \in D_f; (x, y) \in D_b\}.\end{aligned} \tag{17}$$

Meanwhile, erosion is defined as

$$\begin{aligned}(f \ominus b)(s, t) = \min \{&f(s + x, t + y) \\ &- b(x, y) \mid (s + x), (t + y) \in D_f; (x, y) \in D_b\}.\end{aligned} \tag{18}$$

The gray scale erosion measures the minimum gap between the image values f and the translated SE values over the domain of x. The gray scale dilation is the dual of the erosion and indirectly measures how well the SEs fit above f. The Hit-Miss transform measures how a shape h fits under f using erosion and how a shape m fits above f via dilation. The high value of Hit-Miss transform means good fit. The gray scale Hit-Miss transform is independent of shifting in gray scale.

2.5.1. The Feature Extraction Phase. There are four elements associated with each layer of feature extraction phase: feature maps, input, and two structure elements. In the first layer, the subimage is used as input, and the last layer's output is the input of the classification stage. In each feature extraction layer, a pair of Hit-Miss SEs is shared within all the feature maps. These SEs are translated as input weights for the feature map nodes in the feature extraction layer. Table 1 shows the input parameters and output parameters related to the feature extraction phase.

According to the above parameters, we can define the Hit-Miss transform as follows:

$$\begin{aligned}\text{net}_y^h &= \min_{x \in D_{t_y}} \left\{a(x) - t_y^h(x)\right\}, \\ \text{net}_y^m &= \max_{x \in D_{t_y}} \left\{a(x) - t_y^{\widehat{m}}\right\}, \\ a_y &= \text{net}_y^h - \text{net}_y^m.\end{aligned} \tag{19}$$

Here, net_y^h stands for the input for Hit operation in node y and h means the Hit operation. net_y^m means the net input for the Miss operation in node y. m and $\widehat{m}$ here mean the Miss operation and reflection of m, respectively. a_y is the result of Hit-Miss transform performed at node y. The learning rule

Table 1: Parameters of the feature extraction phase.

	Parameter	Definition
Input	$a(x)$	The input to a node y from node x
	$t_y(x)$	Connections associating the node y with node x
	$t_y^h(x_y)$	Hit SE associating node y with node x
	$t_y^m(x)$	Miss SE associating node y with x
	$w_y^h(x)$	Weight for Miss SE node y with x
	$w_y^m(x)$	Weight for Hit SE node y with x
Output	a_y	The output of node y

for the Hit and Miss SE is derived based on the gradient decent as

$$\Delta t_y^h = \eta \delta_y \frac{\partial \text{net}_y^h}{\partial t_y^h(x)}, \qquad \Delta t_y^{\hat{m}} = -\eta \delta_y \frac{\partial \text{net}_y^m}{\partial t_y^{\hat{m}}(x)}, \tag{20}$$

where η is the learning rate of the network and δ_y is expressed as

$$\delta_y = \delta(y) = \sum {}_k \delta_k w_k(y). \tag{21}$$

Equation (21) is for the top level or final extraction layer. δ_y for the lower layers of multiple-layer feature extraction is expressed as

$$\delta_y = \delta(y) = \sum {}_k \delta_k \left(\frac{\partial \text{net}_y^h}{\partial a(y)} - \frac{\partial \text{net}_y^m}{\partial a(y)} \right), \tag{22}$$

in which k is the node in the layer next to the node y.

Based on the back-propagation of error from the classification stage with these learning rules, the MSNN learns the optimized SE to extract the features by each set of Hit-Miss transforms. Consider

$$E = \frac{1}{2} \sum_o (t_o - O_o)^2. \tag{23}$$

Here, t_o stands for the target node output and O_o the actual node output:

$$O_j = f(\text{net}_j), \qquad \text{net}_j = \sum_i w_{ji} O_i + \Delta_j, \tag{24}$$

in which w_{ji} is the connection weight strength to node j from node i and Δ_j is the bias output for node j. w_{ji} is typically learned by the back-propagation of error. The update rule of connecting weight for each connection is expressed as follows:

$$\Delta w_{ji} = -\eta \frac{\partial E}{\partial w_{kj}} = \eta \delta_j O_j. \tag{25}$$

For the output layer nodes, w_{kj} stands for the connection strength to node k from node j:

$$\delta_j = (t_j - O_j) f'(\text{net}_j) \tag{26}$$

and for the hidden layer nodes,

$$\delta_j = f'(\text{net}_j) \sum_k \delta_k w_{ji}. \tag{27}$$

2.5.2. The Classification Phase. The classification phase takes the output directly from the last feature extraction layer as its input. The parameters used for the classification phase are predefined in the feature extraction phase. There are three output nodes for the classification stage of our algorithm, indicating which type of spines the subimage contains.

2.5.3. Acceleration of the MSNN Based on the Regularization. In order to accelerate the learning rate and decrease the learning epochs, we employed the regularization factor. Regularization is used to reduce near-zero connection weight value to zero, therefore reducing the complexity of the network. It is defined as

$$R(w) = E_s(w) + \lambda E_c(w), \qquad E_c(w) = \sum_{\forall w \text{ in network}} \frac{(w_i/w_o)^2}{1 + (w_i/w_0)^2}, \tag{28}$$

where $E_s(w)$ is the performance measure of the learning algorithm, the total network error, and $E_c(w)$ is the complexity penalty of the network model. λ is the regularization factor. w_0 is a predefined parameter. Meanwhile, research shows that a network with proper SEs produces better result [36]. Therefore, it is essential to choose the suitable SEs. In this paper, according to the average size of spine and the comparison result in Table 3, we defined the SE as a disk with the radius of 4 pixels.

For the training procedure, the RMSNN takes the subimage as the input and makes one output value for each image. For the testing procedure, our proposed algorithm scans the whole ROI and generates an image named the detection plane, which is based on the outputs from the target class nodes.

3. Experimental Evaluation

3.1. Experiment Design. We have trained neural networks with the back-propagation algorithm. The subimages were submitted to the input nodes of the neural network. The error of the output was propagated through all the connections. The process repeated until the network converged to a stable state with required MSE. When the MSE approximated to a preset value or the maximum epoch was achieved, the algorithm converged and the training would stop. During the training, the RMSNN took each subimage as the input and produced one output value for each of the three categories. Figure 2(a) shows the samples of subimages containing mushroom type

spine. Figure 2(b) shows the samples of the subimages containing the stubby type, and Figure 2(c) shows the samples of thin type subimage.

In the training step, the subimage samples were input to the network sequentially. The median-squared error was employed to measure the training effectiveness. For each subimage, the RMSNN produced one output value, which indicated the type of spine in the subimage. Then, we scanned the entire microscopy image and finally generated a detection plane according to the output nodes of RMSNN.

In order to test the classification accuracy, we randomly selected 900 samples for each type of spine, respectively. Following common convention and ease of stratified cross validation, 10×10-fold stratified cross validation (CV) was used for the dataset to perform an unbiased statistical analysis. The RMSNN was constructed in the form as two feature extraction layers, one hidden layer with ten hidden neurons and one output layer with three neurons. The input subimage size was 20 by 20 pixels, and the size of the structure elements was with the radius of 4 pixels. The initial weight was in the range of $[-1.0, 1.0]$. The learning rate was set to 0.0015. The maximum training epoch was predefined as 15000. The expected output values for mushroom, stubby, and thin type spines were [1 0 0], [0 1 0], and [0 0 1].

3.2. Experiment Results

3.2.1. Backbone Extraction. The extraction result is shown in Figure 5. Figure 5(a) shows the original image. Figure 5(b) shows the extracted backbone, of which the width covers merely one pixel.

3.2.2. Boundary Location. Figure 6(a) shows the mark of the located backbone of the dendrite based on the original image, and Figure 6(b) shows the marked boundary of the dendrite after the backbone is extracted. Figure 6(c) shows the marked dendrite that determines the starting point of the spine.

3.2.3. Spine Analysis. Figure 7 shows a ROI of our sample image, and Figure 7(b) shows the detection result of the spines. The backbone is marked by the purple color and the boundary is marked by the red color. The spines are marked by their periphery of blue color.

Figure 8(a) shows the original image with the marked region of interest. Figure 8(b) shows the classification result based on the features extracted in the first phase. The corresponding SE gets respect features around each pixel, but it is blind for readers to understand which features are obtained. The detected spines contain 8 mushroom types, 8 stubby types, and 4 thin types. The average of the classification accuracy of RMSNN is shown in Table 2 based on the 2700 samples in total. We can find that the detection of the mushroom and thin types has better performance than the stubby type. It is because the stubby type seems connected with the major lines, and the neck of the spine is blurred. Figures 8(c), 8(d), and 8(e) demonstrate partial geometric attributes of the spines, including the area, perimeter, and width. We found that the areas of the spines of the ROI ranged within [10, 23] and the perimeter ranged within [8, 88].

Table 2: Average of the classification accuracy on a 10-by-10 CV.

Spine types	Mushroom	Stubby	Thin
Mushroom	99.1%	1.3%	1.1%
Stubby	0.7%	97.6%	0.3%
Thin	0.2%	1.1%	98.6%

3.3. Optimal Parameter in SE. According to [36], unsuitable SEs will degrade the performance of the RMSNN; hence, it is critical to choose the proper SEs. According to the average size of the spines as 20 by 20 pixels, we selected SEs with different sizes and shapes to test the performance. The comparison of classification accuracies based on the 2700 samples is shown in Table 3. We can find that the disk with a radius of 4 pixels reaches the best performance. Therefore, we finally defined the SEs as a disk with the radius of 4 pixels.

3.4. Algorithm Comparison. To further validate the efficacy of our proposed approach, we have compared the proposed algorithm with Cheng et al.'s method [18] and the manual method. In Cheng et al.'s paper, the authors employed the adaptive threshold to segment the image and Chen and Molloi's algorithm [37] to extract the backbone and then used the local SNR for the detection of the detached spine and local spine morphology for the detection of the attached spines. The comparison results based on ROI1 in Figure 8 and 15 images collected in our database are shown in Table 4. It is found from Figure 9 that Cheng et al.'s method missed some small protrusions whose number of pixels is more than 5. The number of detected spines via our algorithm is 19, 13 by Cheng et al.'s method, and 20 via the manual method as shown in Table 4. Cheng et al.'s method is robust at dealing with the spines detached from the dendrite but weak at spines attached with the dendrite. However, the detached spines from the dendrite are caused by the deconvolution to denoise the image. Our proposed algorithm overcomes the problem of detecting attached spines.

4. Discussion

In this paper, we have proposed new algorithms using conditional symmetric analysis and regularized morphological shared-weight neural network to detect and analyze the dendrite and dendritic spines.

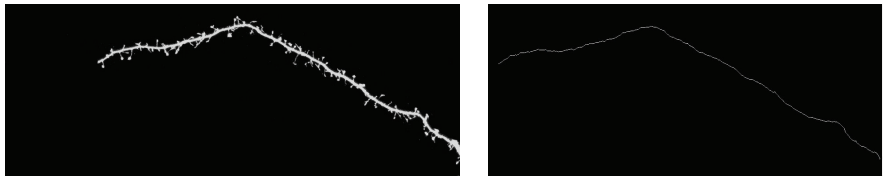
Figure 5 shows that backbone extraction result based on the conditional symmetry analysis. Compared to the second-order directional derivatives method in [14], our proposed algorithms reduced the computation time of linking the breaking point of the backbone.

Figure 6 shows the result of the marked backbone and the boundary of the dendrite, which is used to determine the starting point of the spines.

Table 2 shows the classification result of the different types of spines. The row in Table 2 stands for the actual class and the column in Table 2 stands for the predicted class. The "mushroom" type has an obvious head and thin neck. The "stubby" type lacks obvious neck, and the "thin" type lacks obvious head. In Table 2, the detection accuracy of

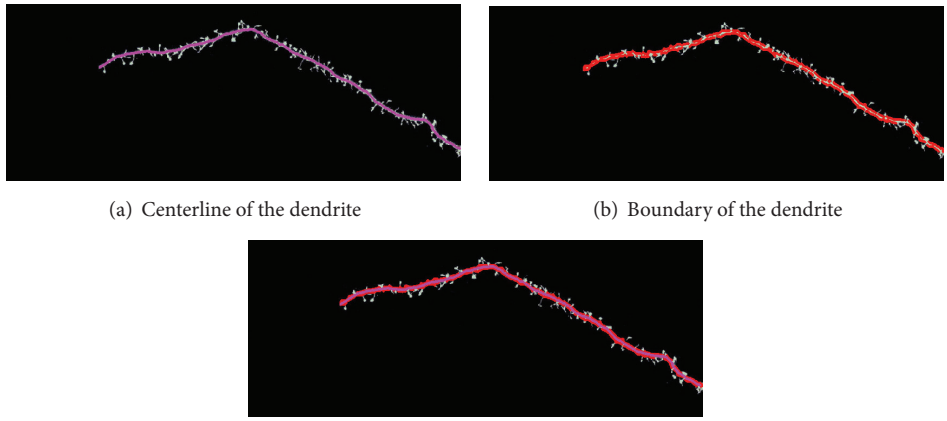
Table 3: Classification accuracy by different SEs (unit is in pixel, bold denotes the best, r is radius, and w is width).

	Disk ($r = 5$)	Disk ($r = 4$)	Disk ($r = 3$)	Square ($w = 3$)	Square ($w = 4$)
Mushroom	98.7%	**99.1**%	95.4%	85.3%	89.2%
Stubby	96.2%	**97.6**%	94.1%	87.2%	91.2%
Thin	94.3%	**98.6**%	96.2%	79.1%	75.3%



(a) Original image (b) Extracted backbone

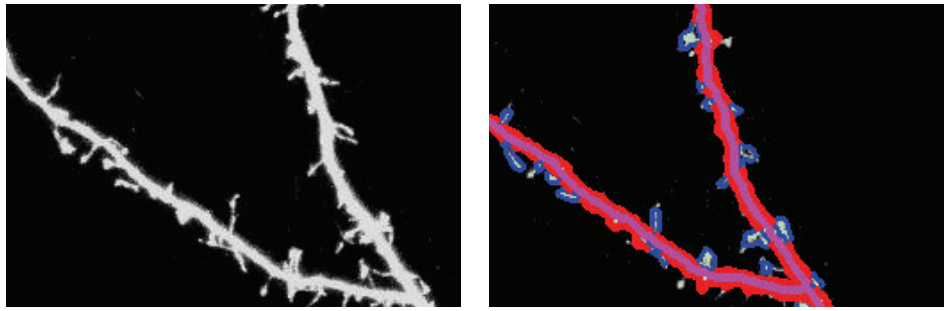
Figure 5: Backbone extraction result.



(a) Centerline of the dendrite (b) Boundary of the dendrite

(c) Dendrite

Figure 6: Dendrite location results.



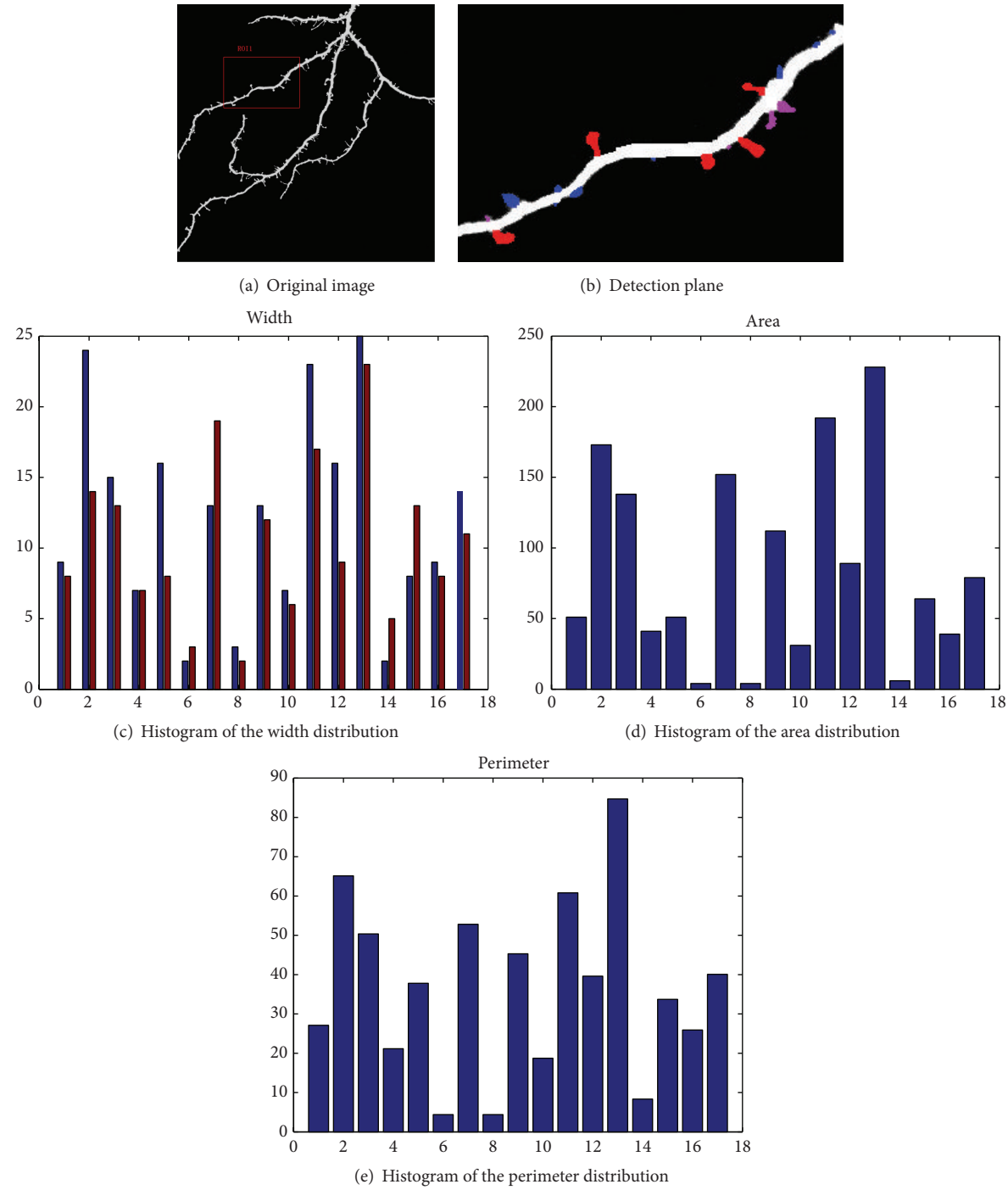
(a) (b)

Figure 7: (a) ROI of the original Image. (b) Detection result of the spines.

Table 4: Detection result of ROI1 in Figure 8 and 15 images in our database.

Methods	ROI1	15 images
Manual	20	2021
ALS [18]	13	1750
SRMSNN (proposed)	19	1987

the mushroom type is higher than the other two types, and part of the stubby type is misclassified into mushroom and thin types as its head and neck ratio is at the level of average. A percent of 1.1 of thin spines are misclassified into mushroom type and 0.3% into stubby type, which is caused by the similar size of the head and neck. Table 4 shows the result of detected spines of Figure 8, respectively, by manual, ALS [18], and our proposed method SRMSNN. The results demonstrate



(a) Original image

(b) Detection plane

(c) Histogram of the width distribution

(d) Histogram of the area distribution

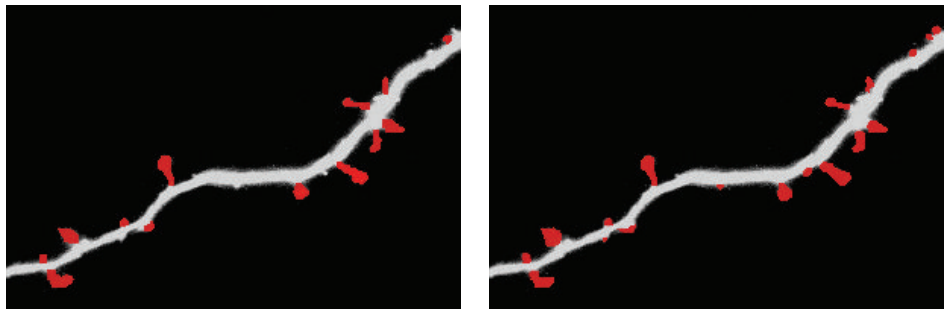
(e) Histogram of the perimeter distribution

Figure 8: Experiment result with corresponding parameters for characterization.

that our algorithm has better performance than the other two methods for the images obtained by the confocal laser scanning microscopy.

5. Conclusion

In this paper, we proposed a new automatic approach to accurately identify dendritic spines with different shapes. The novelty of this approach includes (1) a new model using wavelet-based conditional symmetry analysis for dendrite backbone extraction and localization, which is the first step towards identification of dendritic spins; (2) a new algorithm based on regularized morphological shared-weight neural networks for classification of spines into the right classes (i.e., mushroom, stubby, and thin), entitled "RMSNN." This research was based on our collected microscopy images. We



(a) ALS [18] (b) SRMSNN

Figure 9: Detection result based on ALS and SRMSNN.

have applied our approach to image base library containing around 2700 subimage samples, 900 for each type of spines, and have compared the proposed method with the existing methods. The experimental results demonstrate that our algorithm outperforms existing methods with a significant improvement in accuracy in terms of classifying spines into the different spine categories. The classification accuracy is 99.1% for mushroom spines, 97.6% for stubby spines, and 98.6% for thin spines.

The future work will be focusing on further validation of the robustness of the algorithms through collecting more samples and testing on different datasets. A user-friendly interface will be also built for usability improvement and enhancement. Meanwhile, we will be focusing on reducing the computation time while improving the classification accuracy based on the 3D image stacks. Other feature extraction tools (such as wavelet packet analysis [38], wavelet entropy [39], and 3D-DWT [40]) and other advanced classification tools [41, 42] will be tested. Besides, swarm intelligence method will be used to find optimal parameters [43].

Conflict of Interests

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

Acknowledgment

This work was financially supported by the National Natural Science Foundation of China (no. 61271231).

References

[1] J. L. Krichmar, S. J. Nasuto, R. Scorcioni, S. D. Washington, and G. A. Ascoli, "Effects of dendritic morphology on CA3 pyramidal cell electrophysiology: a simulation study," *Brain Research*, vol. 941, no. 1-2, pp. 11–28, 2002.

[2] D. Johnston and S. M.-S. Wu, *Foundations of Cellular Neurophysiology*, MIT Press, Cambridge, Mass, USA, 1995.

[3] Z. F. Mainen and T. J. Sejnowski, "Influence of dendritic structure on firing pattern in model neocortical neurons," *Nature*, vol. 382, no. 6589, pp. 363–366, 1996.

[4] N. Keren, N. Peled, and A. Korngreen, "Constraining compartmental models using multiple voltage recordings and genetic algorithms," *Journal of Neurophysiology*, vol. 94, no. 6, pp. 3730–3742, 2005.

[5] B. Van Calster, D. Timmerman, C. Lu et al., "Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods," *Ultrasound in Obstetrics and Gynecology*, vol. 29, no. 5, pp. 496–504, 2007.

[6] K. M. Stiefel and T. J. Sejnowski, "Mapping function onto neuronal morphology," *Journal of Neurophysiology*, vol. 98, no. 1, pp. 513–526, 2007.

[7] S. Wang, H. Pan, C. Zhang, and Y. Tian, "RGB-D image-based detection of stairs, pedestrian crosswalks and traffic signs," *Journal of Visual Communication and Image Representation*, vol. 25, no. 2, pp. 263–272, 2014.

[8] S. K. Schmitz, J. J. J. Hjorth, R. M. S. Joemai et al., "Automated analysis of neuronal morphology, synapse number and synaptic recruitment," *Journal of Neuroscience Methods*, vol. 195, no. 2, pp. 185–193, 2011.

[9] T. M. Liu, G. Li, J. X. Nie et al., "An automated method for cell detection in zebrafish," *Neuroinformatics*, vol. 6, no. 1, pp. 5–21, 2008.

[10] W. Yu, H. K. Lee, S. Hariharan, W. Bu, and S. Ahmed, "Evolving generalized voronoi diagrams for accurate cellular image segmentation," *Cytometry Part A*, vol. 77, no. 4, pp. 379–386, 2010.

[11] M. K. Bashar, K. Komatsu, T. Fujimori, and T. J. Kobayashi, "Automatic extraction of nuclei centroids of mouse embryonic cells from fluorescence microscopy images," *PLoS ONE*, vol. 7, no. 5, Article ID e35550, 2012.

[12] J. L. Martiel, A. Leal, L. Kurzawa et al., "Measurement of cell traction forces with ImageJ," in *Methods in Cell Biology*, E. K. Paluch, Ed., vol. 125, chapter 15, pp. 269–287, Academic Press, 2015.

[13] D. L. Dickstein, A. Rodriguez, A. B. Rocher et al., "NeuronStudio: an automated quantitative software to assess changes in spine pathology in Alzheimer models," *Alzheimer's & Dementia*, vol. 6, no. 4, article S410, 2010.

[14] E. Meijering, M. Jacob, J.-C. F. Sarria, P. Steiner, H. Hirling, and M. Unser, "Design and validation of a tool for neurite tracing and analysis in fluorescence microscopy images," *Cytometry Part A*, vol. 58, no. 2, pp. 167–176, 2004.

[15] J. Cheng, X. B. Zhou, B. L. Sabatini, and S. T. C. Wong, "NeuronIQ: a novel computational approach for automatic dendrite spines detection and analysis," in *Proceedings of the IEEE/NIH Life Science Systems and Applications Workshop (LISA '07)*, pp. 168–171, IEEE, Bethesda, Md, USA, November 2007.

[16] I. Y. Y. Koh, W. B. Lindquist, K. Zito, E. A. Nimchinsky, and K. Svoboda, "An image analysis algorithm for dendritic spines," *Neural Computation*, vol. 14, no. 6, pp. 1283–1310, 2002.

[17] X. Y. Xu, J. Cheng, R. M. Witt, B. L. Sabatini, and S. T. C. Wong, "A shape analysis method to detect dendritic spine in 3D optical microscopy image," in *Proceedings of the 3rd IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 554–557, Arlington, Va, USA, April 2006.

[18] J. Cheng, X. Zhou, E. Miller et al., "A novel computational approach for automatic dendrite spines detection in two-photon laser scan microscopy," *Journal of Neuroscience Methods*, vol. 165, no. 1, pp. 122–134, 2007.

[19] J. Fan, X. Zhou, J. G. Dy, Y. Zhang, and S. T. C. Wong, "An automated pipeline for dendrite spine detection and tracking of 3D optical microscopy neuron images of in vivo mouse models," *Neuroinformatics*, vol. 7, no. 2, pp. 113–130, 2009.

[20] Y. Zhang, X. B. Zhou, R. M. Witt, B. L. Sabatini, D. Adjeroh, and S. T. C. Wong, "Dendritic spine detection using curvilinear structure detector and LDA classifier," *NeuroImage*, vol. 36, no. 2, pp. 346–360, 2007.

[21] F. Janoos, K. Mosaliganti, X. Xu, R. Machiraju, K. Huang, and S. T. C. Wong, "Robust 3D reconstruction and identification of dendritic spines from optical microscopy imaging," *Medical Image Analysis*, vol. 13, no. 1, pp. 167–179, 2009.

[22] T. He, Z. Xue, and S. T. C. Wong, "A novel approach for three dimensional dendrite spine segmentation and classification," in *Medical Imaging 2012: Image Processing*, vol. 8314 of *Proceedings of SPIE*, San Diego, Calif, USA, February 2012.

[23] P. Shi, Y. Huang, and J. Hong, "Automated three-dimensional reconstruction and morphological analysis of dendritic spines based on semi-supervised learning," *Biomedical Optics Express*, vol. 5, no. 5, pp. 1541–1553, 2014.

[24] S. Reid, C. Lu, I. Casikar et al., "Prediction of pouch of Douglas obliteration in women with suspected endometriosis using a new real-time dynamic transvaginal ultrasound technique: the sliding sign," *Ultrasound in Obstetrics & Gynecology*, vol. 41, no. 6, pp. 685–691, 2013.

[25] S. Reid, C. Lu, I. Casikar et al., "The prediction of pouch of Douglas obliteration using offline analysis of the transvaginal ultrasound 'sliding sign' technique: inter-and intra-observer reproducibility," *Human Reproduction*, vol. 28, no. 5, pp. 1237–1246, 2013.

[26] Y.-H. Wang, W.-N. Liu, A.-H. Chen, and Y. Wang, "Nonlinear dim target enhancement algorithm based on partial differential equation," *Journal of Dalian Maritime University*, vol. 34, no. 2, pp. 57–60, 2008.

[27] L. Chen, J. H. Zhang, S. Y. Chen, Y. Lin, C. Y. Yao, and J. W. Zhang, "Hierarchical mergence approach to cell detection in phase contrast microscopy images," *Computational and Mathematical Methods in Medicine*, vol. 2014, Article ID 758587, 10 pages, 2014.

[28] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[29] P.-S. Liao, T.-S. Chen, and P.-C. Chung, "A fast algorithm for multilevel thresholding," *Journal of Information Science and Engineering*, vol. 17, no. 5, pp. 713–727, 2001.

[30] L. H. Yang, X. You, R. M. Haralick, I. T. Phillips, and Y. Y. Tang, "Characterization of Dirac edge with new wavelet transform," in *Proceedings of the 2nd International Conference on Wavelets and Applications*, vol. 1, pp. 872–878, Hong Kong, December 2001.

[31] Y. Y. Tang and X. G. You, "Skeletonization of ribbon-like shapes based on a new wavelet function," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1118–1133, 2003.

[32] Y. D. Zhang, S. H. Wang, G. L. Ji, and P. Phillips, "Fruit classification using computer vision and feedforward neural network," *Journal of Food Engineering*, vol. 143, pp. 167–177, 2014.

[33] S. Wang, Y. Zhang, Z. Dong et al., "Feed-forward neural network optimized by hybridization of PSO and ABC for abnormal brain detection," *International Journal of Imaging Systems and Technology*, vol. 25, no. 2, pp. 153–164, 2015.

[34] G. Yang, Y. Zhang, J. Yang et al., "Automated classification of brain images using wavelet-energy and biogeography-based optimization," *Multimedia Tools and Applications*, 2015.

[35] D. Guo, Y. Zhang, Q. Xiang, and Z. Li, "Improved radio frequency identification indoor localization method via radial basis function neural network," *Mathematical Problems in Engineering*, vol. 2014, Article ID 420482, 9 pages, 2014.

[36] X. Jin and C. H. Davis, "Vehicle detection from high-resolution satellite imagery using morphological shared-weight neural networks," *Image and Vision Computing*, vol. 25, no. 9, pp. 1422–1431, 2007.

[37] Z. Chen and S. Molloi, "Automatic 3D vascular tree construction in CT angiography," *Computerized Medical Imaging and Graphics*, vol. 27, no. 6, pp. 469–479, 2003.

[38] Y. Zhang, Z. Dong, S. Wang, G. Ji, and J. Yang, "Preclinical diagnosis of magnetic resonance (MR) brain images via discrete wavelet packet transform with tsallis entropy and generalized eigenvalue proximate support vector machine (GEPSVM)," *Entropy*, vol. 17, no. 4, pp. 1795–1813, 2015.

[39] Y. Zhang, S. Wang, P. Sun et al., "Pathological brain detection based on wavelet entropy and Hu moment invariants," *Bio-Medical Materials and Engineering*, vol. 26, supplement 1, pp. S1283–S1290, 2015.

[40] Y. Zhang, S. Wang, P. Phillips, Z. Dong, G. Ji, and J. Yang, "Detection of Alzheimer's disease and mild cognitive impairment based on structural volumetric MR images using 3D-DWT and WTA-KSVM trained by PSOTVAC," *Biomedical Signal Processing and Control*, vol. 21, pp. 58–73, 2015.

[41] S. Wang, Y. Zhang, G. Ji, J. Yang, J. Wu, and L. Wei, "Fruit classification by wavelet-entropy and feedforward neural network trained by fitness-scaled chaotic ABC and biogeography-based optimization," *Entropy*, vol. 17, no. 8, pp. 5711–5728, 2015.

[42] Y. Zhang, Z. Dong, P. Phillips et al., "Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning," *Frontiers in Computational Neuroscience*, vol. 9, article 66, 15 pages, 2015.

[43] S. Wang, X. Yang, Y. Zhang, P. Phillips, J. Yang, and T.-F. Yuan, "Identification of green, oolong and black teas in China via wavelet packet entropy and fuzzy support vector machine," *Entropy*, vol. 17, no. 10, pp. 6663–6682, 2015.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 571381, 19 pages
http://dx.doi.org/10.1155/2015/571381

Review Article

An Overview of Biomolecular Event Extraction from Scientific Documents

Jorge A. Vanegas,[1] Sérgio Matos,[2] Fabio González,[1] and José L. Oliveira[2]

[1]*MindLab Research Laboratory, Universidad Nacional de Colombia, Bogotá, Colombia*
[2]*DETI/IEETA, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal*

Correspondence should be addressed to Sérgio Matos; aleixomatos@ua.pt

Received 13 May 2015; Revised 10 August 2015; Accepted 18 August 2015

Academic Editor: Chuan Lu

This paper presents a review of state-of-the-art approaches to automatic extraction of biomolecular events from scientific texts. Events involving biomolecules such as genes, transcription factors, or enzymes, for example, have a central role in biological processes and functions and provide valuable information for describing physiological and pathogenesis mechanisms. Event extraction from biomedical literature has a broad range of applications, including support for information retrieval, knowledge summarization, and information extraction and discovery. However, automatic event extraction is a challenging task due to the ambiguity and diversity of natural language and higher-level linguistic phenomena, such as speculations and negations, which occur in biological texts and can lead to misunderstanding or incorrect interpretation. Many strategies have been proposed in the last decade, originating from different research areas such as natural language processing, machine learning, and statistics. This review summarizes the most representative approaches in biomolecular event extraction and presents an analysis of the current state of the art and of commonly used methods, features, and tools. Finally, current research trends and future perspectives are also discussed.

1. Introduction

The scientific literature is the most important medium for disseminating new knowledge in the biomedical domain. Thanks to advances in computational and biological methods, the scale of research in this domain has changed remarkably, reflected in an exponential increase in the number of scientific publications [1]. This has made it harder than ever for scientists to find, manage, and exploit all relevant studies and results related to their research field [1]. Because of this, there is growing awareness that automated exploitation tools for this kind of literature are needed [2]. To address this need, natural language processing (NLP) and text mining (TM) techniques are rapidly becoming indispensable tools to support and facilitate biological analyses and the curation of biological databases. Furthermore, the development of this kind of tools has enabled the creation of a variety of applications, including domain-specific semantic search engines and tools to support the creation and annotation of pathways or for automatic population and enrichment of databases [3–5].

Initial efforts in biomedical TM focused on the fundamental tasks of detecting mentions of entities of interest and linking these entities to specific identifiers in reference knowledge bases [6, 7]. Although entity normalization remains an active research challenge, due to the high level of ambiguity in entity names, some existing tools offer performance levels that are sufficient for many information extraction applications [6]. In recent years there has been increased interest in the identification of interactions between biologically relevant entities, including, for instance, drug-drug [8] or protein-protein interactions (PPIs) [9]. Amongst these, the identification of PPIs mentioned in the literature has received most attention, encouraged by their importance in systems biology and by the necessity to accelerate the population of numerous PPI databases.

Following the advances achieved in PPI extraction, it became relevant to automatically extract more detailed descriptions of protein related events that depict protein characteristics and behavior under certain conditions. Such events, including expression, transcription, localization,

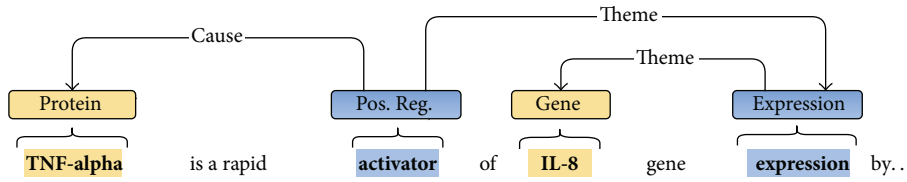


Figure 1: Example of complex biomolecular event extracted from a text fragment. A recursive structure, composed of two types of events, is presented: Positive Regulation and Expression.

binding, or regulation, among others, play a central role in the understanding of biological processes and functions and provide insight into physiological and pathogenesis mechanisms. Automatically creating structured representations of these textual descriptions allows their use in information retrieval and question answering systems, for constructing biological networks composed of such events [2] or for inferring new associations through knowledge discovery. Unfortunately, extraction of this kind of biological information is a challenging task due to several factors: firstly, the biological processes described are generally complex, involving multiple participants which may be individual entities such as genes or proteins, groups, or families, or even other biological processes; sentences describing these processes are long and in many cases have long-range dependencies; and, finally, biological text is also rich in higher level linguistic phenomena, such as speculation and negation, which may cause misinterpretation of the text if not handled properly [1, 9].

This review summarizes the different approaches used to address the extraction and formalization of biomolecular events described in scientific texts. The downstream impact of these advances, namely, for network extraction, for pharmacogenomics studies, and in systems biology and functional genomics, has been highlighted in recent reviews [2, 4, 10], which have also described various end-user systems developed on top of these technologies. This review focuses on the methodological aspects, describing the available resources and tools as well as the features, algorithms, and pipelines used to address this information extraction task, and specifically for protein related events, which have received the most attention in this perspective. We present and discuss the most representative methods currently available, describing the advantages, disadvantages, and specific characteristics of each strategy. The most promising directions for future research in this area are also discussed.

The contents of this paper are organized as follows: we start by introducing biomolecular events and defining the event extraction task; we then describe the event extraction steps, present commonly used frameworks, text processing, and NLP tools and resources, and compare the different approaches used to address this task; in the following section we compare the performance of the proposed methods and systems, followed by a discussion regarding the most relevant aspects; finally, we present some concluding remarks in the last section.

2. Biomolecular Events

In the biomedical domain, an event refers to the change of state of one or more biomedical entities, such as proteins, cells, and chemicals [11]. In their textual description, an event is typically referenced through a trigger expression that specifies the event and indicates its type. These triggers are generally verbal forms (e.g., "stimulates") or nominalizations of verbs (e.g., "expression") and may occur as a single word or as a sequence of words. This textual description also includes the entities involved in the event, referred to as participants, and possibly additional information that further specifies the event, such as a particular cell type in which the described event was observed. Biomolecular events may describe the change of a single gene or protein, therefore having only one participant denoting the affected entity, or may have multiple participants, such as the biomolecules involved in a binding process, for example. Additionally, an event may act as participant in a more complex event, as in the case of regulation events, requiring the detection of recursive structures.

Extraction of event descriptions from scientific texts has attracted substantial attention in the last decade, namely, for those events involving proteins and other biomolecules. This task requires the determination of the semantic types of the events, identifying the event participants, which may be entities (e.g., proteins) or other events, their corresponding semantic role in the event, and finally the encoding of this information using a particular formalism. This structured definition of events is associated with an ontology that defines the types of events and entities, semantic roles, and also any other attributes that may be assigned to an event. Examples of ontologies for describing biomolecular events include the GENIA Event Ontology [11] and Gene Ontology [12].

Figure 1 presents an example of a complex event described in the text fragment "*TNF-alpha is a rapid activator of IL-8 gene expression by...*." From this fragment we can construct a recursive structure composed of two events: a first event, of type *Expression* denoted by the trigger word "*expression*" that has a single argument ("IL-8") with the role *Theme* (denoting that this is the participant affected by the event), and a second event of type *Positive Regulation*, defined by the trigger word "*activator.*" This second event has two participants: the protein "TNF-alpha" with the role *Cause* (defining that this protein is the cause of the event) and the first event with the role *Theme*.

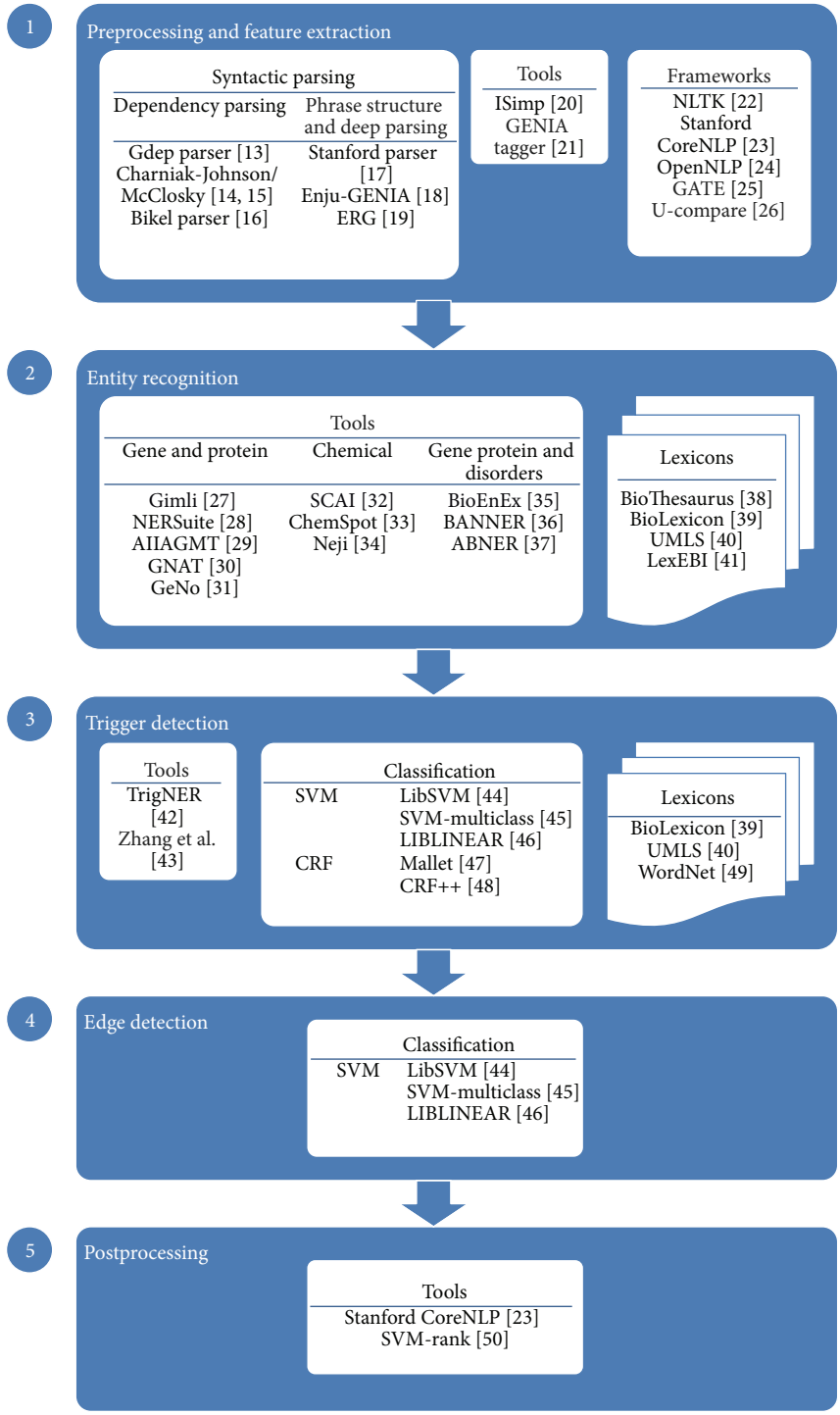


Figure 2: Overall pipeline of a biomedical event extraction solution. Joint prediction methods merge steps 3 and 4 in a single step. The corresponding reference paper for each tool and method is also identified [13–50].

3. Event Extraction

Figure 2 illustrates a common event extraction pipeline, identifying the most popular tools, models, and resources used in each stage. The two initial stages are usually preprocessing and feature extraction, followed by the identification of named entities. The next step is to perform event detection. This step is frequently divided into two separate stages: trigger detection, which consists of the identification of event triggers and their type, and edge detection (or event construction), which is focused on associating event triggers with their arguments. Some authors, on the other hand,

have addressed event detection in a single, joint prediction step. These approaches tackle the cascading errors that occur with the two-stage methods and have commonly shown improved performance. Finally, a postprocessing stage is usually present, to refine and complete the candidate event structures. Negation or speculation detection may also be included in this final step. This section describes each phase, presenting the most commonly used approaches.

3.1. Corpora for Event Extraction. The development and improvement of information extraction systems usually requires the existence of manually annotated text collections, or corpora. This is mostly true for supervised machine learning methods, but annotated data can also be exploited for inferring patterns to be used in rule-based approaches. In the case of biomedical event extraction, various corpora have been compiled, including corpora annotated with protein-protein interactions.

3.1.1. GENIA Event Corpus. The GENIA Event corpus contains human-curated annotations of complex, nested, and typed event relations [51, 52]. The GENIA corpus [53] is composed of 1,000 paper abstracts from Medline. It contains 9,372 sentences from which 36,114 events are identified. This corpus is provided by the organizers of BioNLP shared task to participants as the main resource for training and evaluation and is publicly available online (http://www.nactem.ac.uk/aNT/genia.html).

3.1.2. BioInfer Corpus. BioInfer (Biomedical Information Extraction Resource) (http://www.it.utu.fi/BioInfer) [54] is a public resource providing manually annotated corpus and related resources for information extraction in the biomedical domain.

The corpus contains sentences from abstracts of biomedical research articles annotated for relationships, named entities, and syntactic dependencies. The corpus is annotated with proteins, genes, and RNA relationships and serves as a resource for the development of information extraction systems and their components such as parsers and domain analyzers. The corpus is composed of 1100 sentences from abstracts of biomedical research articles.

3.1.3. Gene Regulation Event Corpus. The Gene Regulation Event Corpus (GREC) (http://www.nactem.ac.uk/GREC/) [55] consists of 240 MEDLINE abstracts, in which events relating to gene regulation and expression have been annotated by biologists. This corpus has the particularity that not only core relations between entities that are annotated, but also a range of other important details about these relationships, for example, location, temporal, manner, and environmental conditions.

3.1.4. GeneReg Corpus. The GeneReg Corpus [56] consists of 314 MEDLINE abstracts containing 1770 pairwise relations denoting gene expression regulation events in the model organism *E. coli*. The corpus annotation is compatible with the GENIA event corpus and with in-domain and out-of-domain lexical resources.

3.1.5. PPI Corpora. Although not as richly annotated as event corpora, protein-protein interaction corpora may be considered for complementing the available training data. The most relevant PPI corpora are the LLL corpus [57], the AIMed corpus [58], and the BioCreative PPI corpus [7].

3.2. Preprocessing and Feature Extraction. Preprocessing is a required step in any text mining pipeline. This includes reading the data from its original format to an internal representation, and extracting features, which usually involves some level of text or language processing. In the specific case of event extraction, preprocessing may also involve resolving coreferences [59] or applying some form of sentence simplification [60], for example, by expanding conjunctions, in order to improve the extraction results.

3.2.1. Preprocessing Tools

Frameworks. In order to derive a feature representation from texts, it is necessary to perform text processing involving a set of common NLP tasks, going from sentence segmentation and tokenization, to part-of-speech tagging, chunking, and linguistic parsing. Various text processing frameworks exist that support these tasks, among which the following stand out: NLTK (http://www.nltk.org/), Apache OpenNLP (https://opennlp.apache.org/), and Stanford CoreNLP (http://nlp.stanford.edu/software/corenlp.shtml) (Figure 2).

Syntactic Parsers. A syntactic parser assigns a tree or graph structure to a free text sentence. These structures establish relations or dependencies between the organizing verb and its dependent arguments and have been useful for many applications like negation detection and disambiguation among others. Syntactic parsers can be categorized in three groups: dependency parsers, phase structure parsers, and deep parsers [61]. The aim of dependency parsers is to compute a tree structure of a sentence where nodes are words, and edges represent the relations among words; phrase structure parsers focus on identifying phrases and their recursive structure, and deep parsers express deeper relations by computing theory-specific syntactic/semantic structures. For the task of event extraction several implementations of each parser groups have been used, as shown in Figure 2.

3.2.2. Features. One of the main requirements of a good event extraction system is a rich feature representation. Most event extraction systems present a complex set of features extracted from tokens, sentences, dependency parsing trees, and external resources. Table 1 summarizes the features commonly extracted in this processing stage and indicates their use in the event extraction process.

(i) Token-based features capture specific knowledge regarding each token, such as syntactic or linguistic features, namely, part-of-speech (POS) and

Table 1: Most common features used in the main event detection stages.

Feature groups	Features	Trigger recognition	Edge detection
Token	Part-of-speech	X	X
	Lemma		
	Orthographic	X	
	Char n-grams	X	
	Word shape	X	
	Prefixes/suffixes	X	
Sentence and local context	Number of entities	X	
	BoW counts	X	
	Windows or conjunctions of features	X	
Dependency	Number and type of dependency edges	X	
	Words, lemmas, or POS tags in dependency path	X	X
	N-grams in dependency path	X	X
External resources	WordNet lemmas	X	X
	Trigger lexicon	X	X
	Entity lexicon	X	X

the lemma of each token, and features based on orthographic (e.g., presence of capitalization, punctuation, and numeric or special characters) [42, 43, 62–68] and morphological information, namely, prefixes, suffixes, and character n-grams [42, 43, 64, 67, 69–72].

(ii) Contextual features provide general characteristics of the sentence or neighborhood where the target token is present. Features extracted from sentences include the number of tokens in the sentence [42], the number of named entities in the sentence, and bag-of-word counts of all words [43, 64]. Local context is usually encoded through windows or conjunctions of features, including POS tags, lemmas, and word n-grams, extracted from the words around the target token [42, 63, 65, 73].

(iii) Dependency parsing provides information about grammatical relationships involving two words, extracted from a graph representation of the dependency relations in a sentence. Commonly used features include the number or type of dependency hops between two tokens, and the sequence or n-grams of words, lemmas, or POS tags in the dependency path between two tokens [65, 68, 72, 74]. These features are usually extracted between two entities in a sentence [64, 75], or between a candidate trigger and an entity [75].

(iv) Finally, it is also common to encode domain knowledge as features using external resources such as lexicons of possible trigger words and of gene and protein names to indicate the presence of a candidate trigger or entity [27, 76–78]. Also, the token representation is often expanded with related words according to some semantic relations such as WordNet hypernyms [27, 77, 79].

3.3. Entity Recognition. Entity recognition consists of the detection of references (or mentions) to entities, such as genes or proteins, in natural language text and labeling them with their location and type. Named-entity recognition in the biomedical domain is generally considered to be more difficult than in other domains, for several reasons: first, there are millions of entity names in use [71] and new ones are added constantly, implying that dictionaries cannot be sufficiently comprehensive; second, the biomedical field is evolving too quickly to allow reaching a consensus on the name to be used for a given entity [80] or even regarding the exact concept defined by the entity itself. So the same name or acronym can be used for different concepts [81].

Several entity recognition systems for the biomedical domain have been developed in the last decade. Much of this work has focused on the recognition of gene and protein names and, more recently, chemical compounds [82]. In these cases, machine learning strategies using rich sets of features have provided the best results, with performances in the order of 85% F-measure [83].

The most popular entity recognition tools are shown in Figure 2, which also lists the biomedical lexicons that are commonly used, either in dictionary-matching approaches or as features for machine learning. Some of these tools, namely, BANNER [36] and Gimli [27], offer simple interfaces for training new models and have been applied to the recognition of various entity types such as chemical compounds and diseases.

3.4. Trigger Detection. Trigger word detection is the event extraction task that has attracted most research interest. It is a crucial task, since the effectiveness of the following tasks strongly depends on the information generated in this step. This task consists of identifying the chunk of text that triggers the event and serves as predicate. Although trigger words are not restricted to a particular set of part-of-speech tags, verbs (e.g., "activates") and nouns (e.g., "expression") are the most

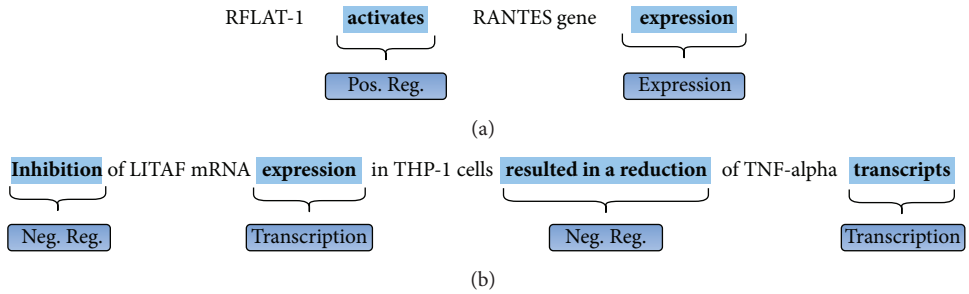


Figure 3: Trigger detection for two example sentences: (a) "RFLAT-1 activates RANTES gene expression" and (b) "Inhibition of LITAF mRNA expression in THP-1 cells resulted in a reduction of TNF-alpha transcripts."

Table 2: Most relevant work addressing the problem of trigger detection. Studies are listed in chronological order and the different approaches are classified in three main groups: rule-based, dictionary-based, and ML-based strategies.

Approach						Reference
Rule-based	Dictionary-based	ML-based				
		SVM	CRF	VSM	MEMM	
X	X					Kilicoglu and Bergler 2009 [84]
	X		X			MacKinlay et al. 2009 [85]
		X (structural)	X			Björne et al. 2009 [86]
	X					Miwa et al. 2010 [87]
X	X					Le Minh et al. 2011 [70]
X	X	X				Kilicoglu and Bergler 2011 [79]
X						Casillas et al. 2011 [88]
X		X (L, R)				Van Landeghem et al. 2011 [74]
		X (P)	X	X (CS)		Martinez and Baldwin 2011 [73]
					X	Zhou and He 2011 [89]
		X (L)				Miwa et al. 2012 [75]
		X (L)				Björne et al. 2012 [64]
		X (C)				Qian and Zhou 2012 [90]
		X (L)				Wang et al. 2013 [65]
		X (L)				Hakala et al. 2013 [91]
		X (L)				Zhang et al. 2013 [43]
		X (L)				Liu et al. 2013 [72]
			X			Campos et al. 2014 [42]
		X (L)				Xia et al. 2014 [92]

L: linear kernel; R: radial basis function kernel; P: polynomial kernel; C: convolution tree kernel; CS: cosine similarity.

common. Furthermore, a trigger may consist of multiple consecutive words.

Figure 3 illustrates the expected results of the trigger detection process in two example sentences. As we can see in Figure 3, trigger detection involves the identification of event triggers and their type, as specified by the selected ontology. In sentence (a), two different kinds of events are identified: the trigger word *activates* defines an event of type *Positive Regulation* and the trigger word *expression* defines an event of type *Gene Expression*. Sentence (b) illustrates the difficulty of this task: it shows that short sentences can contain various related events; that triggers may be expressed in diverse ways (two event of type *Negative Regulation* are defined with different trigger words); and, finally, that the same trigger word (*expression*) may indicate different types of event, depending on the context.

The various approaches proposed for trigger detection can be roughly categorized in three types: rule-based, dictionary-based, and machine learning-based. These approaches are summarized in Table 2 and presented in the remainder of this section.

3.4.1. Patterns and Matching Rules for Trigger Detection. There are several strategies based on patterns [70, 93] and matching rules. Rule-based methods commonly follow some manually defined linguistic patterns, which are then augmented with additional constraints based on word forms and

syntactic categories to generate better matching precision. The main advantage of this kind of approach is that they usually require little computational effort. Rule-based event extraction systems consist of a set of rules that are manually defined or generated from training data. For instance, Casillas et al. [88] present a strategy based on Kybots (Knowledge Yielding Robots), which are abstract patterns that detect actual concept instances and relations in a document. These patterns are defined in a declarative format, which allows definition of variables, relations, and events. Vlachos et al. [76] present a domain-independent approach based on the output of a syntactic parser and standard linguistic processing (namely, stemming, lemmatization, and part-of-speech (POS) tagging, among others), augmented by rules acquired from the development data in an unsupervised way, avoiding the need to use explicitly annotated training data.

In the dictionary-based approach, a dictionary containing trigger words with their corresponding classes (event types) is used to identify and assign event triggers. Van Landeghem et al. [74] proposed a strategy following this approach, using a set of manually cleaned dictionaries and a formula to calculate the importance of each trigger word for a particular event. This is required since the same word may be associated with events of different types [66]. For instance, in the BioNLP'09 Shared Task dataset [51], the token "overexpression" appears as trigger for the gene expression event in about 30% of its occurrences, while the other 70% of occurrences are triggers for positive or negative regulation events.

Many strategies combine both approaches. For instance, Le Minh et al. [70] present a strategy where rule-based and dictionary-based approaches are combined. First, they select tokens that have appropriate POS tags and occur near a protein mention and then apply heuristic rules extracted from a training corpus to identify candidate triggers. Finally, a dictionary built from the training corpus and containing trigger words and their corresponding classes is used to classify candidate triggers. For ambiguous trigger classes, the class with the highest rate of occurrence is selected. Kilicoglu and Bergler [93] also present a combined strategy based on a linguistically inspired rule-based and syntax-driven methodology, using a dictionary based on trigger expressions collected from the training corpus. Events are then fully specified through syntactic dependency based heuristics, starting from the triggers detected by the dictionary-matching step.

Pattern-based methods usually present low recall rates, since defining comprehensive patterns would require extensive efforts, and because the most common patterns are too rigid to capture semantic/syntactic paraphrases.

3.4.2. Machine Learning-Based Approach to Trigger Detection. The most recent and successful approaches to trigger word detection are based on machine learning methods [72], with most work defining this as a sequence-labeling problem. The definition of event types, on the other hand, is addressed as a multiclass task, where candidate event triggers are classified into one of the predefined types of biomedical events. In order to address these problems, several probabilistic techniques have been proposed, using, for example, Hidden Markov Models (HMMs), Maximum Entropy Markov Models (MEMMs), Conditional Random Fields (CRFs) [94, 95], and Support Vector Machines (SVMs).

For instance, Zhou and He [89] proposed treating trigger identification as a sequence-labeling problem and use the Maximum Entropy Markov Model (MEMM) to detect trigger words. MEMM is based on the concept of a probabilistic finite state model such as HMM but consists of a discriminative model that assumes the unknown values to be learnt are connected in a Markov chain rather than being conditionally independent of each other. Similarly, various strategies based on Conditional Random Fields (CRFs) have been proposed [42, 73, 85, 86]. CRFs have become a popular method for sequence-labeling problems, justified mainly by the fact that CRFs avoid the label bias problem present in MEMMs [96] but preserve all the other advantages. Unlike Hidden Markov Models (HMMs), CRF is a discriminant model. So CRFs use conditional probability for inference, meaning that they maximize $p(y \mid x)$ directly, where x is the input sequence and y is the sequence of output labels, unlike HMMs, which maximize the joint probability $p(x, y)$. This relaxes strong independence assumptions required to learn the parameters of generative models.

The most recent proposals for trigger detection are based on Support Vector Machines (SVMs). SVMs do not follow a probabilistic approach but are instead maximum margin classifiers that try to find the maximal separation between classes. This classifier has presented very good results, showing a higher generalization performance than CRFs. However, training complex SVM models may require excessive computational time and memory overhead. Several strategies using different SVM implementations and kernels have been proposed.

The general approach is to classify initial candidate triggers as positive or not, based on a set of carefully selected features and a training set with annotated events. For instance, Björne et al. [80, 86, 97] proposed a solution based on the SVM-multiclass (http://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html) implementation with a linear kernel, optimized by exploring in an exhaustive grid search the C-parameter that maximizes the F-score in trigger detection. In this study only linear kernels were used since the size and complexity of the training set, composed of over 30 thousand instances and nearly 300 thousand features, hinders the application of more computationally demanding alternatives, namely, radial basis function kernels.

In addition to purely supervised learning, which depends on the amount and quality of annotated data, semisupervised approaches have also been proposed. Wang et al. [65] combined labeled data with large amounts of unlabeled data, using a rich representation based on semantic features (such as walk subsequence features and n-gram features, among others) and a new representation based on Event Feature Coupling Generalization (EFCG). EFCG is a strategy to produce higher-level features based on two kinds of original features: class-distinguishing features (CDFs) which have

the ability to distinguish the different classes and example-distinguishing features (EDFs) that are good at indicating the specific examples. EFCG generates a new set of features by combining these two kinds of features and taking into account a degree of relatedness between them.

A different strategy was followed by Martinez et al., who presented a solution based on word-sense disambiguation (WSD) using a combined CRF-VSM (Vector Space Model) classifier, where the output of VSM is incorporated as a feature into the CRF [73]. This approach significantly improved the performance of each method separately.

3.5. Edge Detection. Edge detection (also known as event theme construction or event argument identification) is the task of predicting arguments for an event, which may be named entities (i.e., genes and proteins) or another event, represented by another trigger word. Event arguments are graphically represented through directed edges from the trigger word for the event and the argument. These edges also express the semantic role that a participant (entity or event) plays in a given event. In Figure 4, sentence (a) illustrates a basic event defined by the trigger word *Phosphorylation* that denotes an event of type *Phosphorylation*. The directed edge between this trigger word and the entity *TRAF2*, denoting a relation of type "Theme," indicates that this entity is the affected participant in this event. It is important to note that events can act as participants in other events, thus allowing the construction of complex conceptual structures. For example, consider the sentence (c), where two events are mentioned: a first event of type *Expression* and a second event of type *Positive Regulation*. The directed edge from the trigger word *activator* and the trigger word *expression* denotes that the event *Expression* is affected directly by the event *Positive Regulation*. Similarly, the edge of type cause between *activator* and the entity *TNFalpha* indicates that this is the causing participant for this event.

Different approaches have been suggested to tackle the edge detection task, including rule and dictionary-based strategies and machine learning-based methods. These are summarized in Table 3 and described in the following subsections.

3.5.1. Patterns and Matching Rules for Edge Detection. These strategies are based on the identification of edges according to a set of rules that can be manually defined or generated from training data. Among the most basic approaches, we find the strategy proposed by MacKinlay et al. [85], in which a specific set of hand-coded grammars, supported by specific domain knowledge like named entity annotations and lexicons, is defined for each type of event. In the case of basic events a simple distance criterion is applied, assigning the closest protein as the theme of the event, while extra criteria is required for more complex events. For instance, to assign the *Theme* arguments for binding events, the maximum distance away from the trigger event word(s), and the maximum number of possible themes are estimated, and for regulation events, in addition to the maximum distance, some priority rules are used to define *Cause* or *Theme* arguments.

Kilicoglu and Bergler [93] present another rule-based approach, where identification of the event participants and corresponding roles (e.g., *Theme* or *Cause*) is primarily achieved based on a grammar created from dependency relations between event trigger expressions and event arguments in the training corpus. This strategy is based on the Stanford syntactic parser [98], which was applied to automatically extract dependency relation paths between event triggers and their corresponding event arguments. These paths were manually filtered, preserving only the correct and sufficiently general ones.

Le Minh et al. [70] follow a similar strategy by generating pattern lists from training data using the dependency graphs resulting from application of a deep syntactic parser.

Bui et al. [99] present one of the most recent studies based on dictionaries and patterns automatically generated from a training set. In this work, less than one minute was required to process a training set composed of about 950 abstracts on a computer with 4 gigabytes of memory, illustrating a main advantage of rule-based systems. Unfortunately, despite the low computational requirements, this kind of approach usually shows modest performance in terms of recall, due to the difficulty in modeling more complex relationships and in defining rules capable of generalizing.

3.5.2. Machine Learning-Based Approach to Edge Detection. In recent years, similarly to trigger detection, there has been a clear tendency to approach the edge detection task using machine learning methods. Most works agree on addressing this problem as a supervised multiclass classification problem by defining a limited number of edge classes.

As can be seen in Table 3, most approaches are based on SVMs. Miwa et al. [87] presented one such approach, dividing the task into two different classification problems: edge detection between two triggers and edge detection between a trigger and a protein. For this purpose a set of annotated instances is constructed from a training set, as follows: for each event found in the training set, a list of annotated edges is constructed using as label the combination of the corresponding event class and the edge type (e.g., Binding: Theme). Using these extracted annotated edges, an unbalanced classification problem is then solved using one-versus-rest linear SVMs. Björne et al. [64] and Wang et al. [65] followed similar approaches, using multiclass SVMs in which two kinds of edges are annotated: trigger-trigger and trigger-protein. Each example is classified as *Theme*, *Cause,* or *Negative* denoting the absence of an edge between the two nodes. Each edge is predicted independently, so that the classification is not affected by positive or negative classification of other edges.

Roller and Stevenson [68] evaluated a similar strategy, using a polynomial kernel. The classification of the relations is carried out in three stages. The first consists of the identification of basic events by defining the trigger and a theme referring to a protein; the second stage seeks to identify regulation events by defining the trigger and a theme referring to a trigger from a previously identified basic event; and the final stage tries to identify additional arguments.

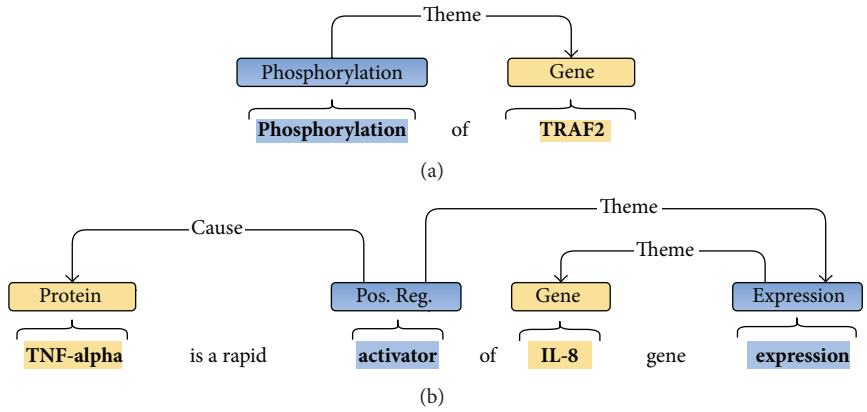


Figure 4: Event extraction from two example sentences: (a) "phosphorylation of TRAF2" and (b) "TNF-alpha which is a rapid activator of IL-8 gene expression."

Table 3: Most relevant work addressing the problem of edge detection. Studies are listed in chronological order and the different approaches are classified in three main groups: rule-based, dictionary-based, and ML-based strategies.

Approach					Reference
Rule-based	Dictionary-based	ML-based			
		SVM	CRF	HVS	
X	X				Kilicoglu and Bergler 2009 [84]
X		X			Björne et al. 2009 [86]
X					MacKinlay et al. 2009 [85]
		X (L)			Miwa et al. 2010 [87]
X					Le Minh et al. 2011 [70]
X					Kilicoglu and Bergler 2011 [79]
		X (L)		X	Zhou and He 2011 [89]
			X		Martinez and Baldwin 2011 [73]
		X (L)			Miwa et al. 2012 [75]
		X (L)			Björne et al. 2012 [64]
		X (L)			Wang et al. 2013 [65]
		X (L)			Hakala et al. 2013 [91]
		X (L)			Xia et al. 2014 [92]

L: linear kernel.

Hakala et al. [91] proposed a reranking approach that uses the prediction scores of a first SVM classifier and information about the event structure as inputs for a new SVM model focused on optimizing the ranking of the predicted edges. For this new model, polynomial and radial basis kernels were evaluated, showing an improvement in the overall precision of the system.

A different strategy was used by Zhou and He [89], who proposed a method based on a Hidden Vector State model, called HVS-BioEvent. Although this method presented lower performance in basic events, compared to systems based on SVM classifiers, it achieved better performance in complex events due to the hierarchical hidden state structure. This structure is indeed more suitable for complex event extraction since it can naturally model embedded structural context in sentences.

Van Landeghem et al. [74] proposed an approach that processes each type of event in parallel using binary SVMs. All predictions are assembled in an integrated graph, on which heuristic postprocessing techniques are applied to ensure global consistency. Linear and radial base function (RBF) kernels were evaluated by performing parameter tuning via 5-fold cross-validation. Van Landeghem et al. made an interesting exploration about feature selection; they applied fully automated feature selection techniques aimed at identifying a subset of the most relevant features from a large initial set of features. An analysis of the results showed that up to 50% of all features can be removed without losing more than one percentage point in F-score, while at the same time creating faster classification models.

3.5.3. Hybrid Approaches. In the literature, we can find many studies that combine ML-based with rule-based and dictionary-based strategies. This combination is often performed in two ways: (1) in an ensemble strategy, each method

is performed independently and the final output is obtained by combining the results of each method, either through rules or by using some classification or regression model; and (2) in a stacked strategy, the output of one method is used as input for the following one that performs a filtering and refining process to produce a more accurate final output.

As an example of the first kind of approach, Pham et al. [100] proposed a hybrid system that combines both rule-based and machine learning-based approaches. In this method, the final list of predicted events is given by the combination of the events extracted by rule-based methods based on syntactic and dependency graphs and those extracted via SVM classifiers. In the second kind of approach, several studies [68, 80, 97] have used a rule-based postprocessing step to refine the initial resulting graph generated by ML-based classifiers by eliminating duplicate nodes and separating their edges into valid combinations based on the syntax of the sentences and the conditions in argument type combinations, taking into account the characteristics and peculiarities of each kind of event.

3.5.4. Structured Prediction and Joint Models. To address the potential cascading errors that originate from two-stage approaches described above, some authors have proposed the joint prediction of triggers, event participants, and connecting edges. Riedel et al. [101] and Poon and Vanderwende [102] proposed two methods based on Markov logic. Markov logic is an extension to first-order logic in which a probabilistic weight is attached to each clause [103]. Instead of using the relational structures over event entities, as represented in Figure 4, Riedel et al. represent these as labeled links between tokens of the sentence and apply link prediction over token sequences. As stated by the authors, this link-based representation simplifies the design of the Markov Logic Network (MLN). Poon and Vanderwendle, on the other hand, used Markov logic to model the dependency edges obtained with the Stanford dependency parser. The resulting MLN therefore jointly predicts if a token is a trigger word, the corresponding event type, and which of the token's dependency edges connect to (Theme or Cause) event arguments. This allows using a simpler set of features in the MLN, which leads to a more computationally efficient solution without sacrificing the prediction performance. The authors used heuristics to fix two typical parsing errors, namely, propositional phrase attachment and coordination, and showed that this had an important impact on the final results.

Riedel and McCallum [104] proposed another approach in which the problem is decomposed in three submodels: one for extracting event triggers and outgoing edges, one for event triggers and incoming edges, and one for protein-protein bindings. The optimization methods for the three submodels are combined via dual decomposition [105], with three types of constraints enforced to achieve a joint prediction model. Links between tokens are represented through a set of binary variables as in Riedel et al. [101].

McClosky et al. [98] proposed a different approach, in which event structures are converted into dependencies between event triggers and event participants. Various dependency parsers are trained using features from these dependency trees as well as features extracted from the original sentences. In recognition phase, the parsing results are converted back to event structures and ranked by a maximum-entropy reranker component.

Vlachos and Craven [106] applied the search-based structured prediction framework (SEARN) to the problem of event extraction. This approach decomposes event extraction into jointly learning classifiers for a set of classification tasks, in which each model can incorporate features that represent the predictions made by the other ones. Moreover, the loss function incorporates all predictions, which means that the models are jointly learned and a structured prediction is achieved. For this specific task, models were trained to classify each token as a trigger or not and to classify each possible pair of trigger-theme and trigger-cause in a sentence.

3.6. Modality Detection. Modality detection refers to the crucial part of identifying negations and speculations [107]. The aim of this task is to avoid opposite meanings and to distinguish when a sentence can be interpreted as subjective or as a nonfactual statement. The detection of speculations (also referred to as hedging) in the biomedical literature has been the focus of several recent studies, since the ability to distinguish between factual and uncertain information is of vital importance for any information extraction task [108].

In many approaches, modality detection is addressed as an extra phase following the edge detection process. Most approaches address this problem in two steps: first speculation/negation cues (which may be words such as "may," "might," "suggest," "suspect," and "seem,") are detected, and, next, the scope of the cues is analyzed. Most of the initial systems were rule-based and relied on lexical or syntactic information, but recent studies have looked at solving this problem using binary classifiers [64, 78, 85] trained with generated instances annotated as negation, speculation, or negative (see Table 4).

4. Comparison of Existing Methods

In this section we present a comparative analysis of the different approaches and systems described in this review. To achieve a consistent comparison, we use the results achieved by the different systems on the standard datasets from the BioNLP shared tasks on event extraction [51, 52, 109]. These datasets provide a direct point of comparison and are commonly used to validate and evaluate new approaches and development, which endorses their use in this comparative analysis. The datasets are based on the GENIA corpus [53], consisting of a training set with 800 abstracts and a development set with 150 abstracts. The test data, composed of 260 abstracts, comes from an unpublished portion of the corpus. For the second edition of the challenge, this initial dataset was extended with 15 full-text articles, equally divided into training, development, and test portions. Evaluation is performed with standard recall, precision, and F-score metrics.

Table 4: Modality detection. Most relevant work addressing the problem of modality detection classified in rule-based, dictionary-based, and ML-based strategies.

Approach				Reference
Rule-based	Dictionary-based	ML-based		
		SVM	CRF	
X	**X**			Kilicoglu and Bergler 2009 [84]
		X (L)		Björne et al. 2009 [86]
			X	MacKinlay et al. 2009 [85]
		X (L)		Miwa et al. 2010 [87]
	X			Kilicoglu and Bergler 2011 [79]
		X (L)		Miwa et al. 2012 [75]
		X (L)		Björne et al. 2012 [64]
		X (L)		Van Landeghem et al. 2013 [110]
		X (L)		Xia et al. 2014 [92]

L: linear kernel.

4.1. BioNLP Shared Task on Event Extraction. The BioNLP shared task series is the main community-wide effort to address the problem of event extraction, providing a standardized dataset and evaluation setting to compare and verify the evolution in performance of different methods. Since its initial organization in 2009, the BioNLP-ST series has defined a number of fine-grained information extraction (IE) tasks motivated by bioinformatics projects. In this analysis, we focus on the main task, GENIA Event Extraction (GE). This task focuses on the recognition of biomolecular events defined in the GENIA Event Ontology, from scientific abstracts or full papers. From the first edition three separate subtasks have been defined, each addressing the event extraction with a different level of specificity.

Task 1. Core event extraction: it consists of the identification of trigger words, associated with 9 events related to protein biology. The annotation of protein occurrences in the text, used as arguments for event triggers, is provided in both the training and the test sets.

Task 2. Event enrichment: it is recognition of secondary arguments that further specify the events extracted in Task 1.

Task 3. Negation/speculation detection: it is detection of negations and speculation statements concerning extracted events.

4.1.1. Target Event Types. The shared task defined a subset of nine biomolecular events from the GENIA Event Ontology, classified in three kinds with different levels of complexity: basic events, binding events, and regulation events. Basic events are the simplest to fully resolve, because these only require the specification of a primary argument. Five types of events are categorized in this group: *gene expression*, *transcription*, *protein catabolism*, *phosphorylation,* and *localization*. *Binding* events, on the other hand, require the detection of at least two arguments. Finally, *regulation* events, including *Negative* and *Positive Regulation*, are the most difficult to fully specify, because these involve the definition of another argument, which may be an entity or another event, requiring identification of a recursive structure.

4.2. Comparative Analysis

4.2.1. Core Event Extraction. Table 5 summarizes the performance achieved by the most representative strategies addressing the core event extraction subtask (Task 1). The best results achieved during the first edition of the BioNLP-ST were obtained through machine learning techniques, formulating the problems of trigger and edge detection as different multiclass classification problems, solved by using linear SVM classifiers [86]. Using the same approach, Miwa et al. [87] reported improvements over these results by adding a set of shortest path features between triggers and proteins for the edge detection problem. As can be observed from the table, a considerable improvement was obtained for binding events, with an increase of over 12 percentage points in recall and 3 points in precision.

In BioNLP-ST 2011, the datasets were extended to include full text articles, but the abstract collection used for the first edition was maintained in order to measure the progress between the two editions. The best result in the second edition, an F-score of 57.46% when considering only the abstracts, was obtained by the FAUST system. This corresponds to a substantial increase of more than four percentage points over the previous best system, resulting from an improvement in the recognition of simple events but especially from a much better recognition of complex regulation events, with an increase of over 11 percentage points in precision and 3 points in recall.

The FAUST system consists of a stacked combination of two models: the Stanford event parser [98] was used for constructing dependency trees that were then used as additional input features for the second model, the UMass model [104]. The main distinction of the UMass model is that it performs joint prediction of triggers, arguments, and event structures, therefore overcoming the cascading errors that occur in the common pipeline approaches when, for example,

Table 5: Core event extraction performance comparison. BioNLP shared task comparison results in recall/precision/*F*-score (%) on the test set for Task 1 (core event extraction). (A) abstracts only and (F) full papers. Data extracted from BioNLP-ST 2009, BioNLP-ST 2011, and BioNLP-ST 2013 overviews [51, 52, 109].

Year	System		Event type			Total
			Simple	Binding	Regulation	
2009	UTurku Björne et al. [86]	(A)	64.21/77.45/70.21	40.06/49.82/44.41	35.63/45.87/40.11	46.73/58.48/51.95
2010	Miwa Miwa et al. [87]	(A)	65.31/76.44/70.44	52.16/53.08/52.62	35.93/46.66/40.60	48.62/58.96/53.29
2011	FAUST Riedel et al. [111]	(A) (F)	66.16/81.04/72.85 75.58/78.23/76.88	45.53/58.09/51.05 40.97/44.70/42.75	39.38/58.18/46.97 34.99/48.24/40.56	50.00/67.53/57.46 47.92/58.47/52.67
	UMass Riedel and McCallum [104]	(A) (F)	64.21/80.74/71.54 75.58/83.14/79.18	43.52/60.89/50.76 41.67/47.62/44.44	38.78/55.07/45.51 34.72/47.51/40.12	48.74/65.94/56.05 47.84/59.76/53.14
2013	EVEX Hakala et al. [91]	(F)	73.83/79.56/76.59	41.14/44.77/42.88	32.41/47.16/38.41	45.44/58.03/50.97
	TEES-2.1 Björne and Salakoski [97]	(F)	74.19/79.64/76.82	42.34/44.34/43.32	33.08/44.78/38.05	46.17/56.32/50.74
	BioSEM Bui et al. [99]	(F)	67.71/86.90/76.11	47.45/52.32/49.76	28.19/49.06/35.80	42.47/62.83/50.68

a trigger is not correctly predicted in the first stage [111]. In this model, the problem of event extraction is divided into smaller simple subproblems that are solved individually, with each subproblem presenting a set of penalties that are added to an objective function. The final solution is found via an iterative tuning of the penalties until all individual solutions are consistent with each other. When used separately, the UMass model achieved the second best-performing results in this edition and was the top performing system when considering just full-texts. In its third edition, BioNLP-ST focused on simulating a more realistic scenario. For this reason, a new dataset was constructed using only recent full papers, so that the extracted information could represent up-to-date knowledge of the domain. Unfortunately, the collection of abstracts used in the first two editions (BioNLP-ST 2009 and BioNLP-ST 2011) was removed from the official evaluation and the full text collection used in the 2011 edition corresponds only to a small part of dataset used in this edition, making it difficult to compare against previous results and measure the progress of the community.

In this latest edition of the shared task the best-performing systems were EVEX [91] and TEES [97]. TEES, an evolution of the UTurku system and also mainly based on SVM classifiers, introduces an automated annotation scheme learning system that derives task-specific event rules and constraints from the training data. In turn, EVEX is a combined system that takes the outputs predicted by TEES and tries to reduce false positives by applying a reranking that assigns a numerical score to events and removing all events that are below a defined threshold. For this reranking, SVM^{rank} is used with a set of features based on confidence scores (i.e., maximum/minimum trigger confidence and maximum/minimum argument confidence, among others) and features describing the structure of the event (i.e., event type of the root trigger and paths in the event from root to arguments, among others). This reranking and filtering approach provided a small overall improvement, achieved through a better precision in the definition of regulation events, which constitute a substantial fraction of the annotated data [105].

BioSEM [99], a rule-based system based on patterns automatically derived from annotated events also achieved high performance results, with only marginal differences to the machine learning approaches described above. BioSEM learns patterns of relations between an event trigger and its arguments defined at three different levels: chunk, phrase, and clause. Notably, this system presents significantly greater precision than ML-based systems, especially considering simple and binding events with improvements of more than seven percentage points. While in the case of simple events this was accompanied by a decrease in recall, for binding events this rule-based system achieved the best results with a difference of over six percent in *F*-score. These results indicate that although ML methods still produce the best generalization, rule-based systems can approximate those results with much better precision and further suggests the combination of the two approaches.

4.2.2. Event Enrichment. Table 6 shows the results obtained in the BioNLP-ST Task 2, which consists of the recognition of secondary event arguments. These secondary arguments depend on the type of event and include *Location* arguments (i.e., *AtLoc* or *ToLoc*) that define the source or destination of an event and *Site* arguments (i.e., *Site* or *Csite*) that indicate domains or regions to better specify the Theme or Cause of an event. The settings of this subtask changed between editions, not only in terms of the dataset used, but also in terms of the sites to be predicted as secondary arguments. This means that the results shown in the table are not directly comparable, namely, for the last edition of the challenge in which sites for different protein modification and regulation events were also considered. Nevertheless, these results were included for reference.

Table 6: Event enrichment performance comparison. BioNLP shared task comparison results in recall/precision/F-score (%) on the test set for Task 2 (event enrichment). (A) abstracts only and (F) full papers. Data extracted from BioNLP-ST 2009, BioNLP-ST 2011, and BioNLP-ST 2013 overviews [51, 52, 109].

Year	System		Site	Localization	Total
2009[a]	UTurku + DBCLS09 Björne et al. [86]	(A)	71.43/71.43/71.43	23.08/88.24/36.59	32.14/72.41/44.52
2011[b]	FAUST Riedel et al. [111]	(A) (F)	43.51/71.25/54.03 17.58/69.57/28.07	36.92/77.42/50.00 —	41.33/72.97/52.77 17.39/66.67/27.59
	UMass Riedel and McCallum (b) [104]	(A) (F)	42.75/70.00/53.08 16.48/75.00/27.03	36.92/77.42/50.00 —	40.82/72.07/52.12 16.30/75.00/26.79
2013[c]	TEES-2.1 Björne and Salakoski [97]	(F)	20.68/59.82/30.73	36.67/78.57/50.00	22.03/61.90/32.50
	EVEX Hakala et al. [91]	(F)	19.44/59.43/29.30	36.67/78.57/50.00	20.90/61.67/31.22

[a]Only phosphorylation sites were considered.
[b]The results are for overall binding and phosphorylation sites.
[c]The task included the prediction of sites for other protein modification and regulation events.

Table 7: Negation and speculation detection performance comparison. BioNLP shared task comparison results in recall/precision/F-score (%) on the test set for Task 3 (negation/speculation detection). (A) abstracts only and (F) full papers only. Data extracted from BioNLP-ST 2009, BioNLP-ST 2011, and BioNLP-ST 2013 overviews [51, 52, 109].

Year	System		Negation	Speculation	Total
2009	ConcordU09 Kilicoglu and Bergler [84]	(A)	14.98/50.75/23.13	16.83/50.72/25.27	15.86/50.74/24.17
2011	UTurku Björne et al. [64, 77]	(A) (F)	22.03/49.02/30.40 25.76/48.28/33.59	19.23/38.46/25.64 15.00/23.08/18.18	20.69/43.69/28.08 19.28/30.85/23.73
	ConcordU11 Kilicoglu and Bergler [93]	(A) (F)	18.06/46.59/26.03 21.21/38.24/27.29	23.08/40.00/29.27 17.00/34.69/22.82	20.46/42.79/27.68 18.67/36.14/24.63
2013	TEES-2.1 Björne and Salakoski [97]	(F)	21.68/36.84/27.30	18.46/33.96/23.92	19.53/35.59/25.22
	EVEX Hakala et al. [91]	(F)	20.98/38.03/27.04	18.46/32.73/23.61	19.82/34.41/25.15

Considering the analysis of abstracts, the table shows an evident improvement on the results achieved by the top performing systems in the first and second editions. More interestingly, there is a considerable difference between the results achieved over full-texts and the results obtained on abstracts. This is an indication that, as expected, the language used for describing the events is much more complex in the main body of the articles, where events are specified in more detail, than in the abstracts. Moreover, while the events are predicted with acceptable levels of precision, the recall is much lower, especially in full-texts.

4.2.3. Negation and Speculation Detection. Table 7 shows the best-performing systems in Task 3, corresponding to the identification of negations and speculations. In the second edition only two teams participated in this task, both presenting an important improvement over the best result of 2009 (ConcordU09 [84]), with UTurku [64, 77] showing a better performance in extracting negated events, and ConcordU11 [93] showing a better performance in extracting speculated events and better overall results in terms of full-texts. As can be directly seen from lower precision and recall rates achieved, this task is considerably more difficult than the extraction of secondary arguments. Although the dataset is different, preventing direct comparison, the results achieved for full-texts on the last edition of the task were similar to the previous results.

5. Discussion and Future Research Directions

Biomolecular event extraction consists of identifying alterations in the state of a biomolecule or interactions between two or more biomolecules, described in natural language text in the scientific literature. These events constitute the building blocks of biological processes and functions, and automatically mining their descriptions has the potential of providing insights for the understanding of physiological and pathogenesis mechanisms. Event extraction has been addressed through multiple approaches, starting from basic pattern matching and parsing techniques to machine learning methods.

Despite the steady progress shown over the last decade, the current state-of-the-art performance clearly shows that extracting events from biomedical literature still presents

various challenges. While performance results close to 80% in F-score have been achieved in the recognition of simpler events, the extraction of more complex events such as binding and regulation events is still limited. Although substantial efforts have been made for the recognition of these events, the best performance achieved remains 30%–40% lower than that for simple events.

5.1. Patterns and Matching Rules versus Machine Learning-Based Approaches. Biomedical event extraction has been moving from purely rule-based and dictionary-based approaches towards ML-based solutions, due to the difficulty in creating sufficiently rich rules that capture the variability and ambiguity of natural language, leading to limited generalization capability and lower recall. Nonetheless, the automatic extraction of rules from annotated data may help in obtaining richer rules. In the third edition of the BioNLP-ST, for instance, the rule-based BioSEM system presented significantly higher precision than the best ML approaches, although with a lower recall.

On the other hand, and despite showing the best performance results in a shared task setting, machine learning approaches present important drawbacks, namely, their dependence on sufficiently large and high-quality training datasets. Another important limitation is that even if such a dataset exists, as in the case of evaluation tasks, its focus may be too restricted which could mean that a model trained on these data would be well tuned for extracting information from similar documents but could become unusable in a slightly different domain. Many recent advances in this task have come from the combination of different systems and approaches. For example, rule-based systems have been applied to derive constrains from the manually annotated data that are then used to correct or filter the results of the machine learning-based event extraction. Another option is to combine the results of rule-based and ML-based methods in an ensemble approach.

5.2. Feature Selection and Feature Reduction. The feature extraction process generates a wide range of features of different nature. In many studies, the generation of the final data representation consists of extracting as many features as possible and integrating them in a basic way. This produces a high dimensional space that does not take into account multiple aspects regarding the nature of the data, such as redundancy, noisy information, or the complexity of its representation space. Although some studies have tried to address this problem, this has mainly been from the point of view of reducing the dimensionality. Some works have shown that an analysis of the contribution of features and appropriate selection of these can significantly reduce the computational requirements. For instance, Campos et al. [42] proposed a solution that chooses the features that better reflect the linguistic characteristics of the triggers for a particular event type; these features are automatically selected via an optimization problem. Also, Van Landeghem et al. [74] showed that a similar overall performance could be achieved using less than 50% of the originally extracted features. Another important consideration is that this reduction not only avoids extra processing time but also helps to avoid undesirable noise [92].

5.3. Current Trends and Challenges. Most event extraction strategies split the problem into two main steps: a first step consisting of the identification of trigger words that indicate the events and a second step (edge detection) that fully specifies the events by adding the corresponding arguments. This makes trigger word detection a crucial task in event extraction, since the second step is commonly performed over the results of that process. In fact, some studies have shown that missing triggers cause about 70% of all errors in event detection [89]. To address these cascading errors, some authors have proposed the joint prediction of triggers and edges connecting these triggers to participants in the event [101, 102, 104, 106, 112]. As shown by the comparative results, this joint inference allowed the most significant advances in terms of prediction performance and constitutes the state-of-the-art approach for event detection. Structured prediction and jointly trained models have also been applied successfully in other biomedical information extraction tasks. Berant et al. [113], for example, used event extraction in order to improve fine-grained information extraction for question answering, applying the structured averaged perceptron algorithm to jointly extract the event triggers and arguments. Kordjamshidi et al. [114] applied structured prediction to the task of extracting information on bacteria and their locations (e.g., host organism) by jointly identifying mentions of entities, organisms, and habitats and corresponding localization relationship. They used a set of local and contextual features for words and phrases and for pairs of phrases and trained structured SVMs for jointly extracting the information.

The use of postprocessing rules to filter and refine the results of model predictions has proved to be an essential step in event extraction. These rules are usually automatically obtained from annotated data and reflect restrictions or likelihoods for the creation of edges between triggers and participants in the construction of the events. On the other hand, the application of automatically extracted rules, on their own, has also shown positive results as shown by the BioSEM system. The ensemble combination of this strategy with the results from ML models could provide a way of balancing the precision and recall of each approach.

While the initial efforts in this task focused on the analysis of abstracts, this greatly limits the amount of information that can be extracted and therefore the impact of these methods on downstream applications, such as question answering, network construction and curation, or knowledge discovery. The latest attempts have therefore focused on mining full-text documents but, as expected, the precision of event extraction using the full body is lower due to the more complex language used in the main text of the publications. Interestingly, the results obtained have shown that while the recognition of complex events becomes more difficult in full-texts, the recognition performance for simple events is higher.

Improving the extraction of complex events, namely, from full-text documents, either through rules, ML, or hybrid

approaches, may depend on the amount and quality of the training data. However, the construction of a fully annotated large-scale dataset that covers the wide variety of linguistic patterns would be a very demanding and unfeasible task. To overcome this, repositories with large amounts of nonannotated data, such as PubMed, could be exploited by unsupervised and semisupervised machine learning methods, to construct richer text representations that can better model complex relations between words. This is a very promising research direction due to the large amount of available data [1] but, unfortunately, very few studies try to take advantage of this unstructured information (i.e., raw text without annotations). Another interesting aspect that could also be further explored is the incorporation of domain information in resources such as dictionaries, thesaurus, and ontologies. Related concepts and semantic relations obtained from these resources could be used to enrich the representation of textual instances or to aid in the generation of filtering and postprocessing rules.

Another major challenge for event extraction is related to coreferences and anaphoric expressions, which make the correct identification of event participants more difficult. This is a very active research field in computational linguistics and natural language processing and has also been vastly studied in the specific case of biomedical text mining [75, 115, 116]. The second edition of the BioNLP-ST included coreference resolution as a supporting task, in which the best participants obtained results ranging from 55% to 73% in precision, for a recall varying between 19% and 22%. These results show that there is still much room for improvement in this area, which would also enhance the event extraction results.

Additionally to the extraction of events, respective types, and participants, a more complete specification of events requires the identification of additional arguments, such as specific binding sites, protein regions, or domains. This extraction of fine-grained information is inherently more difficult than the primary identification of events, as can be seen from the current state-of-the-art performance. However, this information is required if the automatically extracted events are to be used for constructing biological networks [2]. Similarly, the identification of negation and speculation, also addressed by various works and evaluated in the BioNLP-ST setting, still represents a very difficult challenge. Nonetheless, even if current limitations still hinder the direct extraction of reliable biological networks from scientific texts, the existing methods can serve as an efficient aid to accelerate the process of network extraction, when integrated in curation pipelines that allow simple and user-friendly revision, correction, and completion of the extracted information.

6. Conclusions

This paper presents a review of the state-of-the-art in biomolecular event extraction, which is a challenging task due to the ambiguity and variability of scientific documents, and the complexity of the biological processes described. Over the last decades a wide range of approaches have been proposed, ranging from basic pattern matching and parsing techniques to sophisticated machine learning methods.

Current state-of-the-art methods use a stacked combination of models, in which the second model either uses rules to refine the initial predictions or applies reranking to select the best event structures. Additionally, the joint prediction of the full event structure as opposed to a two- or three-stage approach has shown to produce improved results.

Important challenges still exist, namely, in the extraction of complex regulation events, in the resolution of coreferences, and in the identification of negation and speculation. Nonetheless, current methods can be used in text-mining-assisted curation pipelines, for network construction and population of knowledge bases.

Conflict of Interests

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

References

[1] M. S. Simpson and D. Demner-Fushman, "Biomedical text mining: a survey of recent progress," in *Mining Text Data*, pp. 465–517, Springer, New York, NY, USA, 2012.

[2] C. Li, M. Liakata, and D. Rebholz-Schuhmann, "Biological network extraction from scientific literature: state of the art and challenges," *Briefings in Bioinformatics*, vol. 15, no. 5, pp. 856–877, 2014.

[3] A. Manconi, E. Vargiu, G. Armano, and L. Milanesi, "Literature retrieval and mining in bioinformatics: state of the art and challenges," *Advances in Bioinformatics*, vol. 2012, Article ID 573846, 10 pages, 2012.

[4] S. Ananiadou, P. Thompson, R. Nawaz et al., "Event-based text mining for biology and functional genomics," *Briefings in Functional Genomics*, vol. 14, no. 3, pp. 213–230, 2015.

[5] L. Hirschman, G. A. P. C. Burns, M. Krallinger et al., "Text mining for the biocuration workflow," *Database: The Journal of Biological Databases and Curation*, vol. 2012, Article ID bas020, 2012.

[6] D. Campos, S. Matos, and J. L. Oliveira, "Current methodologies for biomedical named entity recognition," in *Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data*, pp. 839–868, John Wiley & Sons, 2013.

[7] C. N. Arighi, Z. Lu, M. Krallinger et al., "Overview of the biocreative III workshop," *BMC Bioinformatics*, vol. 12, supplement 8, article S1, 2011.

[8] I. Segura-Bedmar, P. Martínez, and M. Herrero-Zazo, "Semeval-2013 task 9: extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013)," in *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval '13)*, pp. 341–350, June 2013.

[9] S. Ananiadou, S. Pyysalo, J. Tsujii, and D. B. Kell, "Event extraction for systems biology by text mining the literature," *Trends in Biotechnology*, vol. 28, no. 7, pp. 381–390, 2010.

[10] U. Hahn, K. B. Cohen, Y. Garten, and N. H. Shah, "Mining the pharmacogenomics literature—a survey of the state of the art," *Briefings in Bioinformatics*, vol. 13, no. 4, pp. 460–494, 2012.

[11] J.-D. Kim, T. Ohta, and J. Tsujii, "Corpus annotation for mining biomedical events from literature," *BMC Bioinformatics*, vol. 9, article 10, 2008.

[12] 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.

[13] K. Sagae and J. Tsujii, "Dependency parsing and domain adaptation with LR models and parser ensembles," in *Proceedings of the CoNLL Shared Task of EMNLP-CoNLL*, pp. 1044–1050, Prague, Czech Republic, June 2007.

[14] E. Charniak and M. Johnson, "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05)*, pp. 173–180, June 2005.

[15] D. McClosky, *Any domain parsing: automatic domain adaptation for natural language parsing [Ph.D. thesis]*, Brown University, Providence, RI, USA, 2010.

[16] D. M. Bikel, "Intricacies of collins' parsing model," *Computational Linguistics*, vol. 30, no. 4, pp. 479–511, 2004.

[17] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL '03)*, vol. 1, pp. 423–430, ACM, July 2003.

[18] T. Hara, Y. Miyao, and J. Tsujii, "Evaluating impact of retraining a lexical disambiguation model on domain adaptation of an HPSG parser," in *Proceedings of the 10th International Conference on Parsing Technologies (IWPT '07)*, pp. 11–22, Prague, Czech Republic, June 2007.

[19] A. A. Copestake and D. Flickinger, "An open source grammar development environment and broad-coverage English grammar using HPSG," in *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC '00)*, Athens, Greece, 2000.

[20] Y. Peng, C. O. Tudor, M. Torii, C. H. Wu, and K. Vijay-Shanker, "iSimp in BioC standard format: enhancing the interoperability of a sentence simplification system," *Database*, vol. 2014, Article ID bau038, 2014.

[21] Y. Tsuruoka, Y. Tateishi, J.-D. Kim et al., "Developing a robust part-of-speech tagger for biomedical text," in *Advances in Informatics*, vol. 3746 of *Lecture Notes in Computer Science*, pp. 382–392, Springer, Berlin, Germany, 2005.

[22] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009.

[23] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, Baltimore, Md, USA, June 2014.

[24] The opennlp project, 2005, http://opennlp.apache.org/index.

[25] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva, "Getting more out of biomedical documents with GATE's full lifecycle open source text analytics," *PLoS Computational Biology*, vol. 9, no. 2, Article ID e1002854, 2013.

[26] Y. Kano, W. A. Baumgartner, L. McCrohon et al., "U-compare: share and compare text mining tools with UIMA," *Bioinformatics*, vol. 25, no. 15, pp. 1997–1998, 2009.

[27] D. Campos, S. Matos, and J. L. Oliveira, "Gimli: open source and high-performance biomedical name recognition," *BMC Bioinformatics*, vol. 14, article 54, 2013.

[28] NERsuite: A Named Entity Recognition toolkit, 2015, http://nersuite.nlplab.org/.

[29] C.-N. Hsu, Y.-M. Chang, C.-J. Kuo, Y.-S. Lin, H.-S. Huang, and I.-F. Chung, "Integrating high dimensional bi-directional parsing models for gene mention tagging," *Bioinformatics*, vol. 24, no. 13, pp. i286–i294, 2008.

[30] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, and G. Gonzalez, "Inter-species normalization of gene mentions with GNAT," *Bioinformatics*, vol. 24, no. 16, pp. i126–i132, 2008.

[31] J. Wermter, K. Tomanek, and U. Hahn, "High-performance gene name normalization with GeNo," *Bioinformatics*, vol. 25, no. 6, pp. 815–821, 2009.

[32] R. Klinger, C. Kolářik, J. Fluck, M. Hofmann-Apitius, and C. M. Friedrich, "Detection of IUPAC and IUPAC-like chemical names," *Bioinformatics*, vol. 24, no. 13, pp. i268–i276, 2008.

[33] T. Rocktäschel, M. Weidlich, and U. Leser, "Chemspot: a hybrid system for chemical named entity recognition," *Bioinformatics*, vol. 28, no. 12, pp. 1633–1640, 2012.

[34] D. Campos, S. Matos, and J. L. Oliveira, "A document processing pipeline for annotating chemical entities in scientific documents," *Journal of Cheminformatics*, vol. 7, supplement 1, article S7, 2015.

[35] M. Chowdhury and M. Faisal, "Disease mention recognition with specific features," in *Proceedings of the Workshop on Biomedical Natural Language Processing*, pp. 83–90, Uppsala, Sweden, July 2010.

[36] R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition," in *Proceedings of the 13th Pacific Symposium on Biocomputing*, pp. 652–663, January 2008.

[37] B. Settles, "ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics*, vol. 21, no. 14, pp. 3191–3192, 2005.

[38] H. Liu, Z.-Z. Hu, J. Zhang, and C. Wu, "BioThesaurus: a web-based thesaurus of protein and gene names," *Bioinformatics*, vol. 22, no. 1, pp. 103–105, 2006.

[39] Y. Sasaki, S. Montemagni, P. Pezik, D. Rebholz-Schuhmann, J. McNaught, and S. Ananiadou, "BioLexicon: a lexical resource for the biology domain," in *Proceedings of the 3rd International Symposium on Semantic Mining in Biomedicine (SMBM '08)*, pp. 109–116, September 2008.

[40] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, pp. D267–D270, 2004.

[41] D. Rebholz-Schuhmann, J.-H. Kim, Y. Yan et al., "Evaluation and cross-comparison of lexical entities of biological interest (lexebi)," *PLoS ONE*, vol. 8, no. 10, Article ID e75185, 2013.

[42] D. Campos, Q.-C. Bui, S. Matos, and J. L. Oliveira, "TrigNER: automatically optimized biomedical event trigger recognition on scientific documents," *Source Code for Biology and Medicine*, vol. 9, article 1, 2014.

[43] Y. Zhang, H. Lin, Z. Yang, J. Wang, and Y. Li, "Biomolecular event trigger detection using neighborhood hash features," *Journal of Theoretical Biology*, vol. 318, pp. 22–28, 2013.

[44] 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, article 27, 2011.

[45] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2002.

[46] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: a library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[47] MALLET: A Machine Learning for Language Toolkit, 2002, http://mallet.cs.umass.edu.

[48] T. Kudo, "CRF++: Yet another CRF toolkit," Software, 2005, http://crfpp.sourceforge.net.

[49] M. M. Stark and R. F. Riesenfeld, "Wordnet: an electronic lexical database," in *Proceedings of the 11th Eurographics Workshop on Rendering*, p. 21, Brno, Czech Republic, 1998.

[50] T. Joachims, "Training linear SVMs in linear time," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 217–226, August 2006.

[51] J. D. Kim, T. Ohta, S. Pyysalo et al., "Overview of BioNLP'09 shared task on event extraction," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task (BioNLP '09)*, pp. 1–9, Association for Computational Linguistics, Boulder, Colo, USA, 2009.

[52] J.-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J. Tsujii, "Overview of BioNLP shared task 2011," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 1–6, Association for Computational Linguistics, Stroudsburg, Pa, USA, June 2011.

[53] J.-D. Kim, T. Ohta, K. Oda, and J.-I. Tsujii, "From text to pathway: corpus annotation for knowledge acquisition from biomedical literature," in *Proceedings of the Asia-Pacific Bioinformatics Conference (APBC '08)*, pp. 165–176, Imperial College Press, Kyoto, Japan, January 2008.

[54] S. Pyysalo, F. Ginter, J. Heimonen et al., "BioInfer: a corpus for information extraction in the biomedical domain," *BMC Bioinformatics*, vol. 8, article 50, 2007.

[55] P. Thompson, S. A. Iqbal, J. McNaught, and S. Ananiadou, "Construction of an annotated corpus to support biomedical information extraction," *BMC Bioinformatics*, vol. 10, article 349, 2009.

[56] E. Buyko, E. Beisswanger, and U. Hahn, "The genereg corpus for gene expression regulation events—an overview of the corpus and its in-domain and out-of-domain interoperability," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10)*, N. Calzolari, K. Choukri, B. Maegaard et al. et al., Eds., p. 1921, European Language Resources Association (ELRA), Valletta, Malta, 2010.

[57] The LLL corpus, 2015, http://genome.jouy.inra.fr/texte/LLLchallenge/.

[58] The AIMed corpus, 2015, ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/.

[59] K. Raghunathan, H. Lee, S. Rangarajan et al., "A multi-pass sieve for coreference resolution," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*, pp. 492–501, October 2010.

[60] Y. Peng, M. Torii, C. H. Wu, and K. Vijay-Shanker, "A generalizable NLP framework for fast development of pattern-based biomedical relation extraction systems," *BMC Bioinformatics*, vol. 15, article 285, 2014.

[61] R. S. T. Y. Miyao, K. Sagae, T. Matsuzaki, and J. Tsujii, "Task-oriented evaluation of syntactic parsers and their representations," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Columbus, Ohio, USA, June 2008.

[62] S. Pyysalo, T. Ohta, M. Miwa, H.-C. Cho, J. Tsujii, and S. Ananiadou, "Event extraction across multiple levels of biological organization," *Bioinformatics*, vol. 28, no. 18, pp. i575–i581, 2012.

[63] D. Okanohara, Y. Miyao, Y. Tsuruoka, and J. Tsujii, "Improving the scalability of semi-Markov conditional random fields for named entity recognition," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 465–472, Association for Computational Linguistics, Sydney, Australia, 2006.

[64] J. Björne, F. Ginter, and T. Salakoski, "University of turku in the bionlp'11 shared task," *BMC Bioinformatics*, vol. 13, supplement 11, article S4, 2012.

[65] J. Wang, Q. Xu, H. Lin, Z. Yang, and Y. Li, "Semi-supervised method for biomedical event extraction," *Proteome Science*, vol. 11, article S17, 2013.

[66] S. Riedel, R. Sãtre, H.-W. Chun, T. Takagi, and J. Tsujii, "Biomolecular event extraction with Markov logic," *Computational Intelligence*, vol. 27, no. 4, pp. 558–582, 2011.

[67] L. R. McGrath, K. Domico, C. D. Corley, and B.-J. Webb-Robertson, "Complex biological event extraction from full text using signatures of linguistic and semantic features," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 130–137, Association for Computational Linguistics, Portland, Ore, USA, June 2011.

[68] R. Roller and M. Stevenson, "Identification of genia events using multiple classifiers," in *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 125–129, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.

[69] D. Campos, S. Matos, and J. L. Oliveira, "A modular framework for biomedical concept recognition," *BMC Bioinformatics*, vol. 14, article 281, 2013.

[70] Q. Le Minh, S. N. Truong, and Q. H. Bao, "A pattern approach for biomedical event annotation," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 149–150, Association for Computational Linguistics, Stroudsburg, Pa, USA, 2011.

[71] L. Tanabe, N. Xie, L. H. Thom, W. Matten, and W. J. Wilbur, "GENETAG: a tagged corpus for gene/protein named entity recognition," *BMC Bioinformatics*, vol. 6, supplement 1, article S3, 2005.

[72] X. Liu, A. Bordes, and Y. Grandvalet, "Biomedical event extraction by multi-class classification of pairs of text entities," in *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 45–49, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.

[73] D. Martinez and T. Baldwin, "Word sense disambiguation for event trigger word detection in biomedicine," *BMC Bioinformatics*, vol. 12, supplement 1, article S4, 2011.

[74] S. Van Landeghem, B. De Baets, Y. de Peer, and Y. Saeys, "High-precision bio-molecular event extraction from text using parallel binary classifiers," *Computational Intelligence*, vol. 27, no. 4, pp. 645–664, 2011.

[75] M. Miwa, P. Thompson, and S. Ananiadou, "Boosting automatic event extraction from the literature using domain adaptation and coreference resolution," *Bioinformatics*, vol. 28, no. 13, Article ID bts237, pp. 1759–1765, 2012.

[76] A. Vlachos, P. Buttery, D. Ó. Séaghdha, and T. Briscoe, "Biomedical event extraction without training data," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 37–40, Boulder, Colo, USA, 2009.

[77] J. Björne and T. Salakoski, "Generalizing biomedical event extraction," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 183–191, ACM, Portland, Ore, USA, June 2011.

[78] M. Miwa, S. Pyysalo, T. Ohta, and S. Ananiadou, "Wide coverage biomedical event extraction using multiple partially overlapping corpora," *BMC Bioinformatics*, vol. 14, no. 1, article 175, 2013.

[79] H. Kilicoglu and S. Bergler, "Effective bio-event extraction using trigger words and syntactic dependencies," *Computational Intelligence*, vol. 27, no. 4, pp. 583–609, 2011.

[80] J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski, "Complex event extraction at pubmed scale," *Bioinformatics*, vol. 26, no. 12, pp. i382–i390, 2010.

[81] G. Zhou, J. Zhang, J. Su, D. Shen, and C. Tan, "Recognizing names in biomedical texts: a machine learning approach," *Bioinformatics*, vol. 20, no. 7, pp. 1178–1190, 2004.

[82] M. Krallinger, O. Rabal, F. Leitner et al., "The CHEMDNER corpus of chemicals and drugs and its annotation principles," *Journal of Cheminformatics*, vol. 7, supplement 1, article S2, 2015.

[83] D. Campos, S. Matos, and J. L. Oliveira, "Biomedical named entity recognition: a survey of machine-learning tools," in *Theory and Applications for Advanced Text Mining*, chapter 8, pp. 175–195, InTech, Rijeka, Croatia, 2012.

[84] H. Kilicoglu and S. Bergler, "Syntactic dependency based heuristics for biological event extraction," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 119–127, Association for Computational Linguistics, Boulder, Colo, USA, 2009.

[85] A. MacKinlay, D. Martinez, and T. Baldwin, "Biomedical event annotation with CRFs and precision grammars," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 77–85, Boulder, Colo, USA, June 2009.

[86] J. Björne, J. Heimonen, F. Ginter et al., "Extracting complex biological events with rich graph-based feature sets," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 10–18, 2009.

[87] M. Miwa, R. Sætre, J.-D. Kim, and J. Tsujii, "Event extraction with complex event classification using rich features," *Journal of Bioinformatics and Computational Biology*, vol. 8, no. 1, pp. 131–146, 2010.

[88] A. Casillas, A. D. de Ilarraza, K. Gojenola, M. Oronoz, and G. Rigau, "Using kybots for extracting events in biomedical texts," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 138–142, Portland, Ore, USA, June 2011.

[89] D. Zhou and Y. He, "Biomedical events extraction using the hidden vector state model," *Artificial Intelligence in Medicine*, vol. 53, no. 3, pp. 205–213, 2011.

[90] L. Qian and G. Zhou, "Tree kernel-based protein-protein interaction extraction from biomedical literature," *Journal of Biomedical Informatics*, vol. 45, no. 3, pp. 535–543, 2012.

[91] K. Hakala, S. Van Landeghem, T. Salakoski et al., "EVEX in ST'13: application of a large-scale text mining resource to event extraction and network construction," in *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 26–34, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.

[92] J. Xia, A. C. Fang, and X. Zhang, "A novel feature selection strategy for enhanced biomedical event extraction using the Turku system," *BioMed Research International*, vol. 2014, Article ID 205239, 12 pages, 2014.

[93] H. Kilicoglu and S. Bergler, "Adapting a general semantic interpretation approach to biological event extraction," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 173–182, Association for Computational Linguistics, Portland, Ore, USA, June 2011.

[94] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning (ICML '01)*, pp. 282–289, Williamstown, Mass, USA, June-July 2001.

[95] H. M. Wallach, "Conditional random fields: an introduction," CIS Technical Report MS-CIS-04-21, 2004.

[96] P. Le-Hong, X. H. Phan, and T. T. Tran, "On the effect of the label bias problem in part-of-speech tagging," in *Proceedings of the IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF '13)*, pp. 103–108, Hanoi, Vietnam, 2013.

[97] J. Björne and T. Salakoski, "TEES 2.1: automated annotation scheme learning in the bionlp 2013 shared task," in *Proceedings of the Bionlp Shared Task 2013 Workshop*, pp. 16–25, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.

[98] D. McClosky, M. Surdeanu, and C. D. Manning, "Event extraction as dependency parsing," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11)*, vol. 1, pp. 1626–1635, Association for Computational Linguistics, Portland, Ore, USA, 2011.

[99] Q.-C. Bui, D. Campos, E. van Mulligen, and J. Kors, "A fast rule-based approach for biomedical event extraction," in *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 104–108, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.

[100] X. Q. Pham, M. Q. Le, and B. Q. Ho, "A hybrid approach for biomedical event extraction," in *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 121–124, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.

[101] S. Riedel, H.-W. Chun, T. Takagi, and J. Tsujii, "A Markov logic approach to bio-molecular event extraction," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task (BioNLP '09)*, pp. 41–49, Stroudsburg, Pa, USA, 2009.

[102] H. Poon and L. Vanderwende, "Joint inference for knowledge extraction from biomedical literature," in *Proceedings of the Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*, pp. 813–821, Association for Computational Linguistics, 2010.

[103] M. Richardson and P. Domingos, "Markov logic networks," *Machine Learning*, vol. 62, no. 1-2, pp. 107–136, 2006.

[104] S. Riedel and A. McCallum, "Robust biomedical event extraction with dual decomposition and minimal domain adaptation," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 46–50, Association for Computational Linguistics, Stroudsburg, Pa, USA, June 2011.

[105] N. Komodakis, N. Paragios, and G. Tziritas, "MRF optimization via dual decomposition: message-passing revisited," in *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV '07)*, pp. 1–8, IEEE, Rio de Janeiro, Brazil, October 2007.

[106] A. Vlachos and M. Craven, "Biomedical event extraction from abstracts and full papers using search-based structured prediction," *BMC Bioinformatics*, vol. 13, supplement 11, article S5, 2012.

[107] N. Konstantinova, S. C. M. de Sousa, and J. A. Sheila, "Annotating negation and speculation: the case of the review domain," in *Proceedings of the 2nd Student Research Workshop Associated with RANLP (RANLPStud '11)*, pp. 139–144, Hissar, Bulgaria, September 2011.

[108] R. Morante and C. Sporleder, "Modality and negation: an introduction to the special issue," *Computational Linguistics*, vol. 38, no. 2, pp. 223–260, 2012.

[109] J. D. Kim, Y. Wang, and Y. Yasunori, "The genia event extraction shared task, 2013 edition—overview," in *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 8–15, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.

[110] S. Van Landeghem, J. Björne, C.-H. Wei et al., "Large-scale event extraction from literature with multi-level gene normalization," *PLoS ONE*, vol. 8, no. 4, Article ID e55814, 2013.

[111] S. Riedel, D. McClosky, M. Surdeanu, A. McCallum, and C. D. Manning, "Model combination for event extraction in BioNLP 2011," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 51–55, Association for Computational Linguistics, Portland, Ore, USA, June 2011.

[112] H. Liu, L. Hunter, V. Kešelj, and K. Verspoor, "Approximate subgraph matching-based literature mining for biomedical events and relations," *PLoS ONE*, vol. 8, no. 4, Article ID e60954, 2013.

[113] J. Berant, V. Srikumar, P.-C. Chen et al., "Modeling biological processes for reading comprehension," in *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP '14)*, October 2014.

[114] P. Kordjamshidi, D. Roth, and M.-F. Moens, "Structured learning for spatial information extraction from biomedical text: bacteria biotopes," *BMC Bioinformatics*, vol. 16, article 129, 2015.

[115] N. Nguyen, J.-D. Kim, M. Miwa, T. Matsuzaki, and J. Tsujii, "Improving protein coreference resolution by simple semantic classification," *BMC Bioinformatics*, vol. 13, article 304, 2012.

[116] K. Yoshikawa, S. Riedel, T. Hirao et al., "Coreference based event-argument relation extraction on biomedical text," *Journal of Biomedical Semantics*, vol. 2, article S6, 2011.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 846942, 12 pages
http://dx.doi.org/10.1155/2015/846942

Research Article

NMFBFS: A NMF-Based Feature Selection Method in Identifying Pivotal Clinical Symptoms of Hepatocellular Carcinoma

Zhiwei Ji,[1,2] Guanmin Meng,[3] Deshuang Huang,[1] Xiaoqiang Yue,[4] and Bing Wang[1,5,6]

[1] *Machine Learning & Systems Biology Lab, School of Electronics and Information Engineering, Tongji University, 4800 Caoan Road, Shanghai 201804, China*
[2] *School of Information Engineering, Zhejiang A&F University, 88 Huancheng North Road, Linan 311300, China*
[3] *Department of Clinical Laboratory, Tongde Hospital of Zhejiang Province, 234th Gucui Road, Hangzhou 310012, China*
[4] *Department of Traditional Chinese Medicine, Changzheng Hospital, Second Military Medical University, 415 Fengyang Road, Shanghai 200003, China*
[5] *The Advanced Research Institute of Intelligent Sensing Network, Tongji University, 4800 Caoan Road, Shanghai 201804, China*
[6] *The Key Laboratory of Embedded System and Service Computing, Tongji University, 4800 Caoan Road, Shanghai 201804, China*

Correspondence should be addressed to Xiaoqiang Yue; yuexiaoqiang@163.com and Bing Wang; wangbing@ustc.edu

Received 22 April 2015; Revised 20 June 2015; Accepted 2 July 2015

Academic Editor: Tao Huang

Background. Hepatocellular carcinoma (HCC) is a highly aggressive malignancy. Traditional Chinese Medicine (TCM), with the characteristics of syndrome differentiation, plays an important role in the comprehensive treatment of HCC. This study aims to develop a nonnegative matrix factorization- (NMF-) based feature selection approach (NMFBFS) to identify potential clinical symptoms for HCC patient stratification. *Methods*. The NMFBFS approach consisted of three major steps. Firstly, statistics-based preliminary feature screening was designed to detect and remove irrelevant symptoms. Secondly, NMF was employed to infer redundant symptoms. Based on NMF-derived basis matrix, we defined a novel *similarity measurement of intersymptoms*. Finally, we converted each group of redundant symptoms to a new single feature so that the dimension was further reduced. *Results*. Based on a clinical dataset consisting of 407 patient samples of HCC with 57 symptoms, NMFBFS approach detected 8 irrelevant symptoms and then identified 16 redundant symptoms within 6 groups. Finally, an optimal feature subset with 39 clinical features was generated after compressing the redundant symptoms by groups. The validation of classification performance shows that these 39 features obviously improve the prediction accuracy of HCC patients. *Conclusions*. Compared with other methods, NMFBFS has obvious advantages in identifying important clinical features 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 the complication of cirrhosis [1, 2]. The occurrence of HCC is larvaceous and short of specific symptoms [3, 4]. Its diagnosis depends on biopsy, imaging examination such as Doppler ultrasound, computed tomography, magnetic resonance imaging, and blood test [5, 6]. Once the patients with HCC see doctors, the disease has often entered its late stage, losing the chance of resection. Hence, seeking simple methods to predict HCC and its clinical stage is very meaningful and helpful to improve the diagnosis of HCC.

As one of the most popular complementary and alternative medicine modalities, Traditional Chinese Medicine (TCM) plays an active role in treatment of malignant tumors including HCC in Chinese and some East Asian countries [7, 8]. Unlike modern medicine, the diagnosis and treatment of TCM depend on the analysis of symptoms and signs of HCC collected by inspection, auscultation and olfaction, inquiry, and pulse taking and palpation [8]. TCM regards specific combination of symptoms and signs as a TCM syndrome, which is the main basis for treatment; and it can be also

used to guide clinical diagnosis of HCC. Our previous work proposed a hierarchical feature selection (PSOHFS) model to quickly identify the potential HCC syndromes from a TCM clinical dataset [9], by which all the original symptoms were classified into several groups according to the categories of clinical observations, and each symptom group was then converted into a syndrome signature to reduce the searching space of feature selection. But the limitation of this method is that the interactions among symptoms which belong to different categories (aspects) were ignored. Therefore, the current challenge is to design an efficient feature selection approach for high-dimensional TCM data with consideration of clinical significance.

In this study, a nonnegative matrix factorization- (NMF-[10]) based feature selection (NMFBFS) method was proposed to select pivotal clinical symptoms for HCC diagnoses. A TCM clinical dataset was used in this work, which consisted of 407 HCC patients with 57 clinical symptoms. Each patient sample is labeled with a clinical-staging symbol which indicates the severity of certain patient. Firstly, the preliminary screening with statistical method was designed to detect irrelevant symptoms from the full symptom set. Secondly, the process of NMF was implemented after eliminating the irrelevant symptoms. Based on the NMF-derived basis matrix, we defined a similarity measure to infer redundant symptoms by calculating the distance and correlation among the symptoms. Finally, the secondary dimension reduction was implemented based on the inferred groups of redundant symptoms. We converted each symptom group to a new feature (named "mixed feature") if these symptoms represent similar distribution patterns on the sample space. The experiment results show that 39 novel features inferred by NMFBFS obviously improve the accuracy of diagnosis of HCC clinical samples. Moreover, NMFBFS-derived 39 optimal clinical features included some well-known common symptoms of HCC patients. Comparing to three representative feature selection methods (ReliefF [11], mRMR [12], and Elastic Net [13]), our proposed approach showed the best performance to identify optimal clinical features for HCC patients.

2. Materials and Methods

2.1. Experimental Data

2.1.1. Description. In this work, the questionnaire survey dataset of HCC includes 407 samples within two years, and each patient was observed on 57 clinical symptoms (Table 1). Each patient sample is labeled with a symbol of clinical stage, which is related to TCM pattern of syndrome and indicates the severity degree of HCC. According to the international staging system [14], there are three stages and two substages in each phase in this dataset. The aim of our work is to identify the symptom signatures, which are related to three clinical stages: phases I, II, and III, and the larger value indicates that stronger positive symptom occurred. Within our dataset, all the original symptoms are described by two types of data: binary (0 or 1) or integer (0, 1, 2, 3, ...). For example, the type of symptom "tinnitus" is binary (0 or 1), which means two possible states: occurrence (positive) or nonoccurrence (nonpositive). Another example is "sleeplessness" whose value can be 0, 1, 2, or 3. The larger the value is, the stronger the positive state will be. A symptom does not appear positive if its value equals zero.

Table 1: The description of original clinical data of HCC patients.

Sex	Phase I (82)		Phase II (195)		Phase III (130)	
	Phase IA	Phase IB	Phase IIA	Phase IIB	Phase IIIA	Phase IIIB
Male	33	27	50	115	95	10
Female	12	10	10	20	16	9

2.1.2. Data Preprocessing

Refinement of Feature Set. Our original dataset consists of 407 HCC patient samples (Table 1). The first step of preprocessing is to remove the useless features because they provide no useful information for the following classification. If a feature is constant on all the observed samples, it can be considered as useless feature. For our dataset, some symptoms, such as "pale tongue" and "slow pulse," were removed out because there is no any observed patient positive on these symptoms. After removing this kind of features, a refined clinical dataset with 407 samples and 57 symptoms $(V_1, \ldots, V_{57})$ can be obtained.

Simplification of Clinical Staging. The clinical staging of HCC patients in our original dataset was marked with collections "IA," "IB," "IIA," "IIB," "IIIA," and "IIIB." For identifying the symptom signatures related to three clinical stages, all the samples would be relabeled as three classes. Here, we remarked class label "1" for the samples labeled "IA" and "IB." In a similar way, class label "2" is used for "IIA" and "IIB" and "3" is for "IIIA" and "IIIB." Finally, all the 407 clinical samples can be distributed in three categories: 82 samples in phase I, 195 in phase II, and 130 in phase III. The details of the refined dataset were described in Table 1.

2.2. Feature Selection. Feature selection can be 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 are independent of the model [15]. 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 [16]. However, one of the drawbacks of ranking methods is that the selected subset might not be optimal in that a redundant subset might be obtained. Wrapper methods involve combination searches through the feature space, guided by the predicting performance of a model [17]. Heuristic search is widely used in wrapper methods as searching strategy which can produce good results and is computationally feasible; however, they often yield local optimum results. For an embedded method, the feature search process is embedded into classification algorithm, so that the learning process and the feature selection process cannot be separated [18].

2.3. Nonnegative Matrix Factorization. Nonnegative matrix factorization (NMF) aims to obtain a linear representation of multivariate data under nonnegativity constraints. These constraints lead to a part-based representation because only additive, not subtractive, combinations of the original data are allowed [19]. In general, NMF can be used to describe hundreds to thousands of features in a dataset in terms of a small number of metafeatures, particularly in gene expression profiles analysis [20–22].

Let X be $n \times p$ nonnegative matrix; that is, each element $x_{ij} \geq 0$ in X. Nonnegative matrix factorization (NMF) consists in finding an approximation

$$X \approx WH, \tag{1}$$

where the *basis matrix* W and the *mixture coefficient matrix* H are $n \times r$ and $r \times p$ nonnegative matrices, respectively, where $r > 0$ and $r \ll \min(n, p)$. The objective behind the small value of r is to summarize and split the information contained in X into r factors (also called "basis" or "metafeature"). The matrix H has the same number of samples but much smaller number of features rather than matrix X. Therefore, the metafeature expression patterns in H usually provide a robust clustering of samples [22].

The main approach to NMF is for solving estimate matrices W and H as a local minimum:

$$\min_{W,H \geq 0} [D(X, WH) + R(W, H)], \tag{2}$$

where D is a loss function that measures the quality of the approximation which is usually based on either the Frobenius distance or the Kullback-Leibler divergence [19]. R is an optional regularization function, defined to enforce desirable properties on matrices W and H, such as smoothness or sparse [23, 24].

In our study, the loss function in NMF is based on Kullback-Leibler divergence [25]. The above function R was defined as follows:

$$R(W, H) = F_1(W) + F_2(H), \tag{3}$$

where $F_1(W)$ and $F_2(H)$ are regulation functions for W and H, respectively. Here, we applied Tikhonov smoothness regularization [26] for W in

$$F_1(W) = \frac{1}{2} \sum_{i,j} \left([W]_{ij} - c \right)^2, \tag{4}$$

where c is a constant positive or zero. In addition, we applied sparsity-enforcing regularization [26] for H in

$$F_2(H) = \frac{1}{2} \sum_{j} \left(\left\| [H]_{\cdot j} \right\|_2^2 - \alpha^2 \left\| [H]_{\cdot j} \right\|_1^2 \right)^2. \tag{5}$$

In formula (5), $[H]_{\cdot j}$ is jth row of H. $\|[H]_{\cdot j}\|_2^2$ and $\|[H]_{\cdot j}\|_1^2$ define the l_2-norm and l_1-norm of $[H]_{\cdot j}$. The algorithm proposed by Lee is a well-established method to solve the optimization of NMF [27].

2.4. NMF-Based Feature Selection. In this study, our proposed NMF-based feature selection (NMFBFS) approach can be seen as a two-stage filter method. In the first stage, preliminary screening is implemented to detect irrelevant symptoms and remove them from the whole feature set. In the second stage, NMF clusters the redundant symptoms which potentially have similar patterns into different groups, and each group is then transformed into new single features to reduce the dimension. Obviously, the process of NMFBFS is independent of classifier and can quickly infer the optimal feature subset even in the high-dimensional dataset. The flowchart of NMFBFS is shown in Figure 1.

2.4.1. Removing the Irrelevant Symptoms. In our questionnaire, all the symptoms were defined by clinical doctors, which covered many aspects of patients. However, the relevance weight of each feature for distinguishing samples among the clinical stages was not quantitatively studied. In machine learning, the irrelevant features provide no useful information in any context and always scarcely contribute to patient stratification [28]. If the sample size is large, it is meaningful to quickly detect the irrelevant symptoms by calculating the frequencies of positive. Here, we calculated the ratio (frequency) of presence (positive) of each symptom on the samples in every clinical stage. If the frequencies of certain symptom in all the clinical stages are very low, which indicates that this symptom hardly appears positive in most of patients, therefore it is considered as an irrelevant symptom. After removing the irrelevant symptoms from the original dataset, the rest of symptoms are considered as relevant features, which are potentially related to at least one class of patients (or one clinical stage).

2.4.2. Identifying Redundant Symptoms Based on NMF. After the irrelevant symptoms had been removed, nonnegative matrix factorization was applied on the dataset X ($n \times p$). For a given rank r, the matrix X can be decomposed to *basis matrix* W and *coefficient matrix* H. Usually, the value of rank r is much smaller than the number of features (n) and the sample number (p), so that there is at least one dimension in both W and H being very small. The widespread appliances of NMF in biclustering further indicate that basis matrix W can be used for feature clustering and coefficient matrix H is used for sample clustering, respectively [20, 21]. In our study, the number of samples is much larger than the dimensionality; hence, directly calculating distance or correlation to measure the similarity between original features (symptoms) on all the samples will lead to biases because some features might represent local similar patterns on a part of samples. Fortunately, the basis matrix W represents the compressed sample space of matrix X, which facilitates uncovering the difference between features. Here, we introduced two features (v_i and v_j) in original dataset X as an example to clarify the basic idea of this step. According to the definition of NMF, we can easily know

$$\begin{aligned} x_i &= w_i \times H, \\ x_j &= w_j \times H, \end{aligned} \tag{6}$$

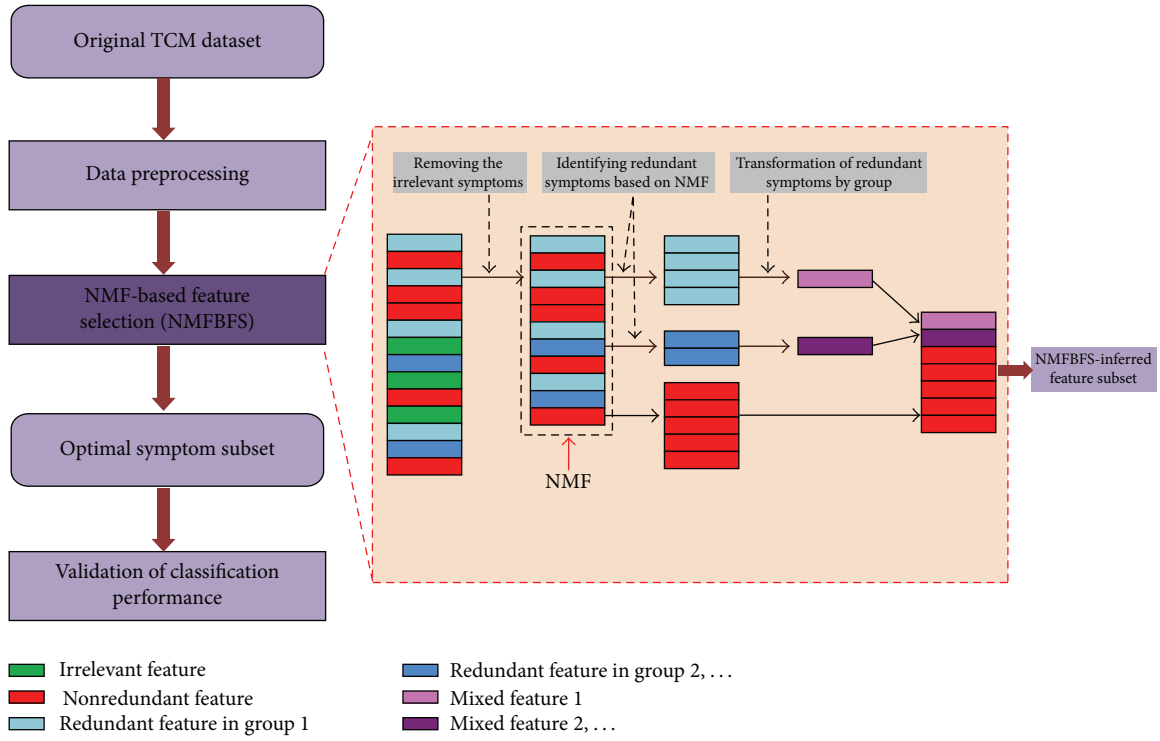


Figure 1: The flowchart of the proposed approach.

where x_i and x_j are ith and jth rows of matrix X; w_i and w_j are ith and jth rows of matrix W. The following can be easily found. (1) If $w_i \approx w_j$, then $x_i \approx x_j$; (2) if $w_i = kw_j$, then $x_i = kx_j$, where k is a constant. Furthermore, if ith row w_i in matrix W is very close to w_j, the feature v_i might have a similar pattern as v_j on all the samples. Therefore, we defined a novel *similarity measurement* in formula (7) to approximately evaluate redundancy between the two original symptoms via matrix W:

$$\begin{aligned} \operatorname{sim}\left(v_i, v_j\right) &\approx \operatorname{sim}\left(w_i, w_j\right) \\ &= \frac{\operatorname{sim\_dist}\left(w_i, w_j\right) + \operatorname{sim\_corr}\left(w_i, w_j\right)}{2}, \end{aligned} \tag{7}$$

where

$$\operatorname{sim\_dist}\left(w_i, w_j\right) = 1 - \frac{\sqrt{\left(w_i - w_j\right) \times \left(w_i - w_j\right)^T}}{\operatorname{Max} D}, \tag{8}$$

$$\operatorname{sim\_corr}\left(w_i, w_j\right) = \left| \frac{\left(w_i - \overline{w}\right) \times \left(w_j - \overline{w}\right)^T}{\sqrt{\left[\left(w_i - \overline{w}\right) \times \left(w_i - \overline{w}\right)^T\right] \times \left[\left(w_j - \overline{w}\right) \times \left(w_j - \overline{w}\right)^T\right]}} \right|. \tag{9}$$

Formula (8) uses *distance-based similarity*, which indicates how two corresponding features are close to each other; and formula (9) adopts *correlation-based similarity*, which describes similar patterns of two original features. Hence, our developed similarity measurement considered distance and correlation between features at the same time. Max D in formula (8) is the maximal distance value in all pairs of (w_i, w_j). Based on the above definition of similarity, we further calculated the similarity matrix SMX using all the basis rows in W (SMX(i, j) = sim(v_i, v_j)), where element SMX(i, j) denotes the similarity between original features i and j. Given a threshold θ $(0 < \theta < 1)$, we can screen all the redundant features by groups with SMX$(i, j) > \theta$.

2.4.3. Transformation of Redundant Symptoms by Group. In the above section, all the redundant symptoms were screened out and were organized into different groups. For each symptom group, a new mixed feature was extracted as the representation of the whole group and replaced all the original features within this group. Therefore, NMFBFS-inferred optimal feature subset includes two parts: nonredundant original features and new generated mixed features (see Figure 1). There are two strategies that can be used to transform the redundant symptom groups to mixed features.

(1) Calculate the mean vector from all the redundant symptoms as in

$$x_{NF} = \operatorname{mean}\left(x_{r1}, x_{r2}, \ldots, x_{rn}\right), \tag{10}$$

where $x_{r1}, x_{r2}, \ldots$, and x_{rn} are the feature vectors of original dataset X and are determined as redundant symptoms in a group. n denotes the number of inferred redundant symptoms in a group. The vector x_{NF} of new single feature v_{NF} was averaged on that group.

(2) Randomly select a vector from one of redundant symptoms as

$$x_{NF} \in \{x_{r1}, x_{r2}, \ldots, x_{rn}\}. \quad (11)$$

In our study, we transformed the groups of redundant symptoms to new mixed features by using formula (10). After this step, the feature space of the clinical dataset was further reduced so that the optimal feature subset rarely included redundant features.

3. Simulation Design

Firstly, we calculated the frequencies of each original symptom appearing positive at each clinical stage and then removed the irrelevant symptoms if their frequency values were very low.

Secondly, a representative sample set was screened out for NMF analysis. In our dataset, the number of samples in three phases of HCC varies a lot, that is, from 82, 130 to 195. If the whole dataset is used, a class imbalance problem will be caused [29–31]. In addition, the sex ratio of patients is also seriously unbalanced in the original dataset (Table 1). For avoiding the bias caused by imbalance of samples, we selected 40 samples from each clinical phase with equal proportion of male and female (20 : 20) to construct a representative clinical dataset D_R (120 samples in total) for the following NMF analysis. Considering the fact that each original sample has a class label which corresponds to clinical stage of that patient, for all the original samples (407), we can actually get a preliminary participation of samples as three clusters, which can also be considered as a trained KNN clustering model [32]. We then defined the center of each cluster, which is the mean vector of all the samples in the same cluster. Given a large value of K, we input each center of cluster into the above KNN model and keep the output consistent with the corresponding class label of the center. Based on the K-nearest neighbors, we can finally screen out 40 representative samples (20 males and 20 females) of each clinical stage according to Euclidean distance.

Finally, several redundant symptom groups were identified. Then we transformed each redundant symptom group into a new mixed feature. Combining all the nonredundant original features with new generated mixed features, we obtained an optimal clinical symptom subset of HCC. At last, the classification performance of this feature subset was further validated by least squares support vector machines (LSSVM) [33, 34].

Experimental Parameters. At first, we set a frequency threshold to identify the irrelevant symptoms. The NMF *R* package [35] was then employed as a computational framework for nonnegative matrix factorization algorithms in *R*. For this method, the optimal rank r should be determined firstly. Currently there are several approaches that had been proposed to determine the optimal value of r [36, 37]. In our study, two methods, that is, cophenetic coefficient [36] and RSS curve [37], had been adopted to determine the optimal rank r range from 2 to 7. After obtaining the results of NMF with optimal r, we calculated the similarity matrix SMX with all the basis rows and inferred the redundant symptoms with a threshold $\theta = 0.95$, which meet the following conditions: $\text{sim\_corr}(w_i, w_j) \geq 0.95$ and $\text{sim\_dist}(w_i, w_j) \geq 0.95$ in formulas (7)–(9). Finally, a LSSVM classifier had been implemented to validate the classification performance of inferred optimal symptom subset. In the LSSVM multiclass model, Gaussian RBF kernel was employed, and the kernel parameters σ^2 and γ were determined by grid search [38]. In our grid search, we set $\sigma^2 = 10^a$ and $\gamma = 10^b$. Variable a changes from −1 to 5 with step 0.25, and variable b changes from −1 to 4 with step 0.2. Therefore, we have the range of [0.1, 100000] for σ^2 and the range of [0.1, 10000] for γ. Totally, there are 24 levels for the value of σ^2 and 25 levels for γ. In other words, there are 600 pairs of σ^2, γ tested when training a LSSVM classifier. To find an optimal value of σ^2, γ, we used 5-fold cross-validation to evaluate the classification accuracy of LSSVM model.

4. Results and Discussion

Firstly, we calculated the frequencies of positive for all the original symptoms (57) at each clinical stage (see Supplementary Table S1 available online at http://dx.doi.org/10.1155/2015/846942). Eight irrelevant symptoms were judged as irrelevant features (threshold: 10%). From Table 2, we can clearly see that these symptoms appeared on few patients (less than 10% in each clinical stage) in the clinical observation and therefore they were considered as noisy features in the process of diagnosis. Because the total number of samples is large (407), we considered that the eight irrelevant symptoms identified with statistical analysis are very reliable. A part of symptoms shown in Table 2 was proved by previous studies. For example, Lai et al. concluded that no association is detected between "emotional depression" and the risk of hepatocellular carcinoma in older people in Taiwan [39, 40]. In addition, Peng et al. studied 169 Chinese patients with HCC; only three patients presented with hydrothorax, which also indicated that this symptom was not a key symptom in the process of liver cancer development [41, 42]. In addition, "edema in lower extremities" is undoubtedly a well-known symptom of HCC patients in clinic [43]; however, it was considered an irrelevant symptom in this study because it rarely appeared in all the three stages of our data. Increasing the observed samples or reducing the threshold will make it as a candidate symptom.

Secondly, the calculation of NMF was implemented after removing all the detected irrelevant symptoms. According to the description in "Simulation Design", NMF was applied on the representative matrix D_R with 120 HCC samples, which uniformly covered three clinical phases. Figure 2(a) represents the fact that D_R is a sparse matrix, in which large partition of elements is zero (no positive), such as symptom V_6 shown in Figure 2(b). However, there are also some symptoms that were positive on many patients, such as symptom V_{25} shown in Figure 2(c). Matrix D_R does not show obvious subtypes and patterns; hence, it is hard to compare the similarity directly between symptoms with the row vectors of D_R since the number of samples is still very large. In this study, we used NMF to compress

Table 2: Eight irrelevant symptoms were screened with threshold 10%. Each of them is rarely positive in each phase.

Symptoms	Phase I		Phase II		Phase III	
	Phase IA	Phase IB	Phase IIA	Phase IIB	Phase IIIA	Phase IIIB
Pale white lip [V_1]	0	5.41%	6.67%	5.19%	4.5%	0
Edema in lower extremities [V_{16}]	2.22%	8.1%	1.67%	5.19%	3.6%	0
Lack of urine output [V_{41}]	0	2.7%	0	0	5.41%	0
Emotional depression [V_{43}]	4.44%	0	5%	8.89%	6.31%	5.26%
Head body trapped heavy [V_{47}]	0	2.7%	3.33%	2.22%	2.7%	0
Hydrothorax [V_{51}]	6.67%	2.7%	1.67%	3.7%	2.7%	0
Rapid pulse [V_{55}]	4.44%	2.7%	1.67%	0.74%	5.41%	5.26%
Uneven pulse [V_{56}]	4.44%	5.41%	8.33%	3.7%	3.6%	0

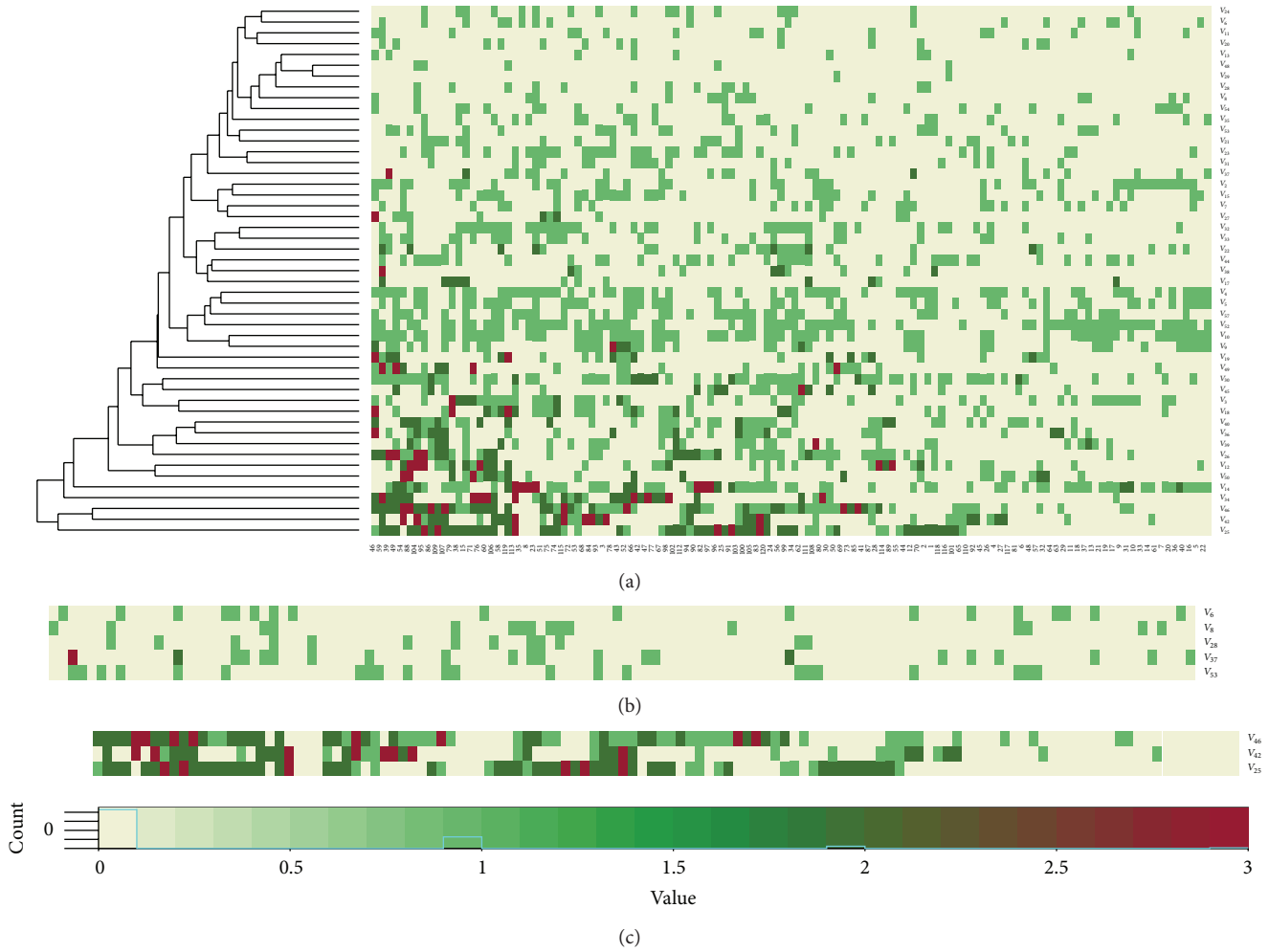


Figure 2: The heatmap of the representative clinical dataset D_R. (a) The heatmap of D_R with 49 symptoms and 120 samples. (b) The distribution patterns of symptoms V_6, V_8, V_{28}, V_{37}, and V_{53} indicate that the frequencies of positive are low. (c) The distribution patterns of symptoms V_{46}, V_{42}, and V_{25} indicate that the frequencies of positive are high.

the representative matrix D_R and to reveal the distribution patterns of features (symptoms) on fewer samples. Before the calculation of NMF, a critical parameter should be firstly determined: the value of factorization rank r. According to Brunet's method, the first value of r for which the cophenetic coefficient starts decreasing is the optimal one [36]. Frigyesi and Höglund suggested choosing the first value where the RSS curve presents an inflection point [37]. Based on these two methods, we determined that "3" is a reasonable value of rank r for the clinical data matrix D_R. The curves shown in Figure 3 also confirm this conclusion. Nonnegative matrix factorization was then implemented on the matrix D_R (49 ×

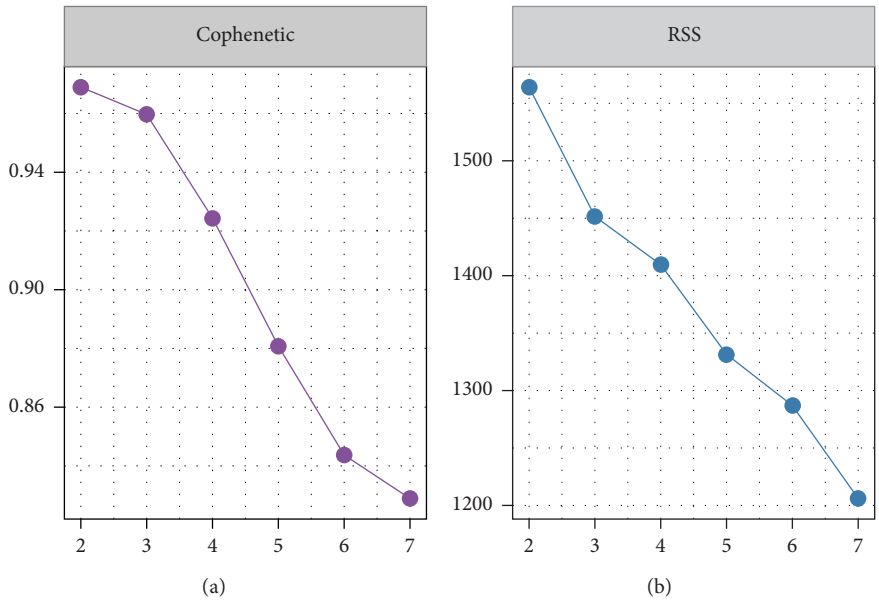


FIGURE 3: Estimation of the optimal rank r.

120) with rank 3. It also indicates that the number of metafeatures (basis) equals 3.

Figure 4 represents the final results of NMF which included the basis matrix W (49×3) and mixture coefficient H (3×120). Each row in matrix W uses a compressed pattern to approximatively represent the distribution of a symptom on all the original samples. Comparing with matrix D_R shown in Figure 2, the obvious difference in matrix W is that there are several groups of features revealing similar patterns in the compressed sample space, such as V_{40} and V_{36} in Figure 4. According to Figure 2(a), we can find that the distance between the vectors of symptoms V_{40} and V_{36} in D_R is also close; furthermore, the compressed patterns of V_{40} and V_{36} in matrix W (w_{40} and w_{36}) in Figure 4 facilitate easier identifying of redundant features which have very similar distribution patterns.

The matrix H has the same number of samples but much smaller number of metafeatures (basis) rather than original matrix X [36]. Therefore, the metafeature expression patterns in H usually provide a robust clustering of samples. Given the jth column in H as $H_j = [h_{j1}, h_{j2}, h_{j3}]^T$, we determined that jth clinical sample is placed into kth cluster if $\max(H_j) = H_j(k)$, where $k \in \{1, 2, 3\}$. Hence, we used matrix H to group all the samples into 3 clusters, which correspond to 3 bases (metafeature). Figure 5 shows that there are great overlaps between the clinical-staging markers (a priori knowledge of class labels) and indexes of basis components (metafeatures) on the 120 original clinical samples included in dataset D_R.

In matrix W, each column also corresponds to a metafeature or basis (see Figure 4). Entry w_{ij} in matrix W is the coefficient of original feature i in metafeature (basis) j [36]. Therefore, an original feature i relates to certain basis j if w_{ij} is the largest entry in row i of matrix W. From Figure 4, we can clearly see that the original symptom features participating in the same basis have similar expression patterns rather than that in other bases. Table 3 represents the symptoms which are related to all basis components. Combination of Figure 5 and Table 3 further indicates that the "basis 1" related symptoms are very related to the clinical samples of phase II, and "basis 2" and "basis 3" related symptoms are very related for phase I and phase III, respectively. This finding contributes to identifying *clinical phase-specific* important symptoms via NMF. Moreover, the partition of 49 clinical symptoms shown in Table 3 was well supported by some related studies. For example, *nausea* is observed as a common adverse effect in HCC patients in phase I [44]. The symptoms *ascites, anorexia, fever*, and *jaundice* often occurred in phase II [43, 45–48]. The symptoms "*yellow complexion*" and "*yellow skin and eye*" shown in Table 3 are obvious appearances of *jaundice*. For phase III, *pain* is the most obvious characteristic in HCC patients [49]. There are three pain-related symptoms presented in Table 3: "*pain in shoulder and back*," "*chest pain*," and "*distending pain in hypochondrium*." Moreover, *fatigue* and *weakness* were also common in HCC patients [43]. Together, these findings suggest that NMF with an optimal rank can reveal the latent associations between the potential symptom features and clinical phases.

Just as mentioned above in "Simulation Design," several groups of redundant features were then screened out according to a given threshold $\theta = 0.95$ (Table 4). We obtained two redundant symptom groups from each basis component, which indicates that the redundant symptoms included in the same group also might have similar patterns in the original sample space. Here, we take Figures 2(b)-2(c) as examples to collaborate the effectiveness of our method. Figure 2(b) represents the distribution of positive of five symptoms in the dataset D_R. These five symptoms (V_6, V_8, V_{28}, V_{37}, and V_{53}) were identified as basis 2 related features, and they are most

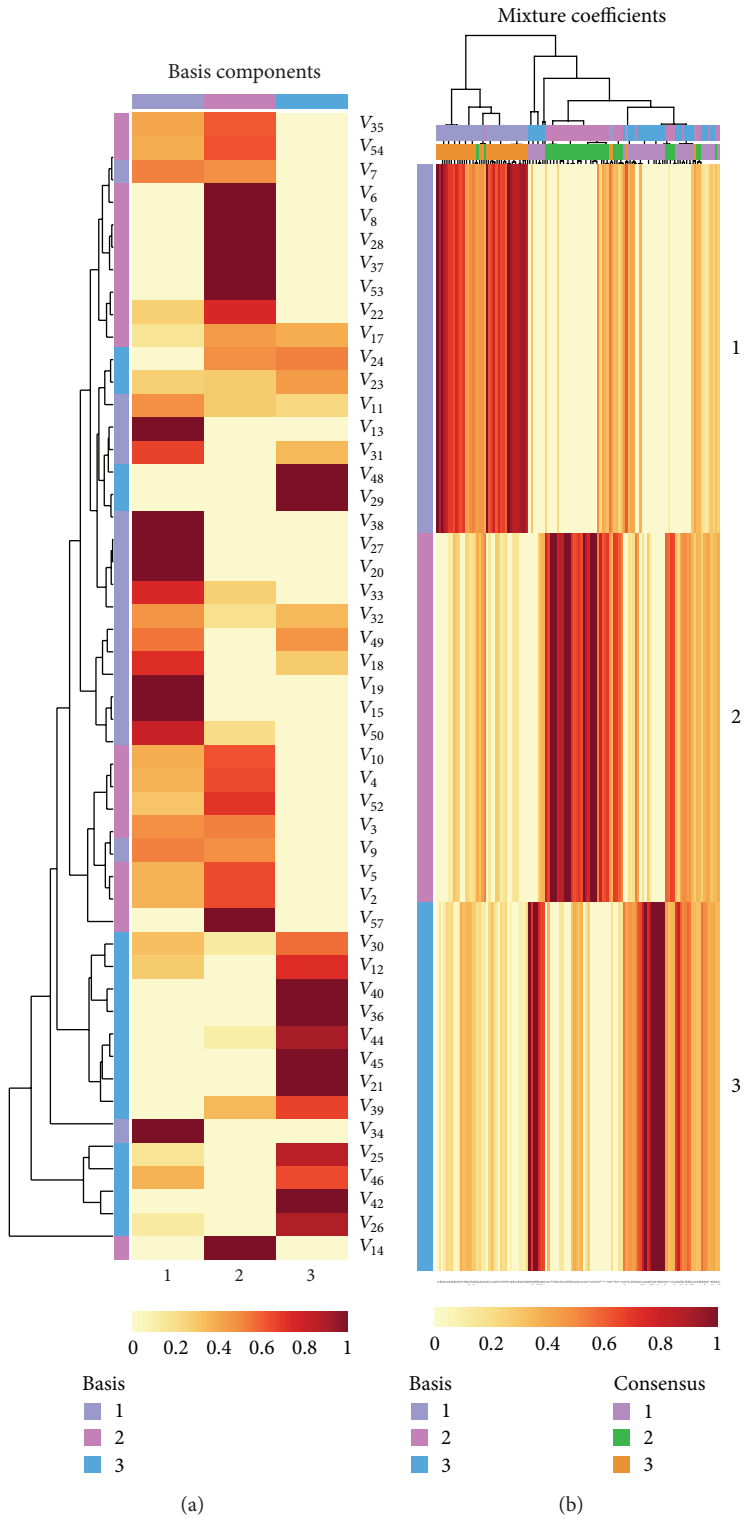


FIGURE 4: The result of NMF on the dataset D_R. The left side indicates the visualization of matrix W $(49 * 3)$, and right side denotes matrix H $(3 * 120)$.

Table 3: The NMF-derived participation of the symptoms to each corresponding basis component.

Basis components	Number of symptoms	The names of symptoms
Basis 1	16	Varicose veins [V_7]; yellow complexion [V_{11}]; yellow skin and eye [V_{13}]; stomach pain [V_{31}]; dry stool [V_{38}]; feeling thirsty [V_{27}]; hot flash [V_{20}]; doing belly full bilge [V_{33}]; fullness in stomach [V_{32}]; block under the rib [V_{49}]; chills [V_{18}]; fever [V_{19}]; spider telangiectasia in liver palm [V_{15}]; ascites [V_{50}]; yellow greasiness [V_9]; anorexia [V_{34}]
Basis 2	17	Nausea [V_{35}]; pulse slip [V_{54}]; petechial and ecchymosis tongue [V_6]; white slip [V_8]; chest distress [V_{28}]; semiliquid stool [V_{37}]; weak pulse [V_{53}]; night sweat [V_{22}]; dirty mouth [V_{17}]; red tongue [V_3]; thready pulse [V_{57}]; sticky greasy coating [V_{10}]; purple tongue [V_4]; stringy pulse [V_{52}]; pale white lip [V_2]; large and teeth-printed tongue [V_5]; gloomy complexion [V_{14}]
Basis 3	16	Tinnitus [V_{24}]; dizziness [V_{23}]; pain in shoulder and back [V_{48}]; chest pain [V_{29}]; distending pain in hypochondrium [V_{30}]; bitter taste [V_{26}]; insomnia [V_{42}]; appearance with stained yellow [V_{12}]; yellow urine [V_{40}]; hiccup [V_{36}]; soreness and weakness of waist and knees [V_{44}]; dry throat [V_{25}]; feverishness in palms and soles [V_{45}]; spontaneous perspiration [V_{21}]; night urination much [V_{39}]; physically and mentally fatigued [V_{46}]

Table 4: The mean similarity values about the pairs of redundant symptoms within the same groups.

Basis components	The screened redundant symptoms	Distance-based similarity sim_dist(w_i, w_j)	Correlation-based similarity sim_corr(w_i, w_j)
Basis 1	V_{38},V_{27}, V_{20}	0.9672	1.0
	V_{19},V_{15}	0.9507	1.0
Basis 2	V_{35},V_{54}	0.9685	0.9960
	V_6,V_8,V_{53},V_{37},V_{28}	0.9628	1.0
Basis 3	V_{48},V_{29}	0.9686	1.0
	V_{44},V_{45}	0.9520	0.9926

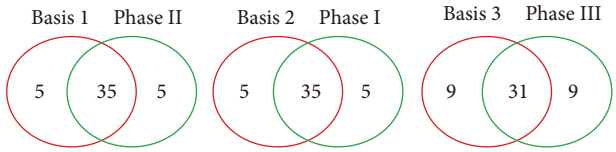


Figure 5: The relationships between NMF-derived basis components and clinical stages of samples.

possibly belonging to phase I (Table 4). Although each of the row vectors in Figure 2(b) is not completely equal, they all represent relative lower frequency of positive ($15.17 \pm 3.25\%$) and their local distribution patterns are similar in a way. Comparing the corresponding rows of these five symptoms in matrix W in Figure 4, we found that the compressed patterns of these symptoms are very similar. Similarly, the symptoms (V_{46}, V_{42}, and V_{25}) are potentially related to basis 3, the frequency of positive for each is over 50%, and the mean value of positive for these three symptoms is 1.77, which further indicate that they might be related to some patients whose conditions are very serious. Although the symptoms V_{46}, V_{42}, and V_{25} were not identified as redundant symptoms with given threshold (0.95), their compressed patterns in matrix W in Figure 4 also suggested that their patterns were very close. In summary, we considered a fact that the matrix W facilitates evaluating the difference among symptoms, and matrix H can validate the high degree of correlation between class labels of samples and basis indexes. After inferring the redundant symptoms with given threshold, we combined each symptoms' group together and converted it into a new feature (named mixed feature). Finally, we obtained 39 clinical features (FS_1) of HCC as the optimal feature subset, which consisted of two parts: 33 original symptom features (FS_2) and 6 new mixed features (FS_3) (Table 5). Based on the analysis of results of NMF, the feature space of original dataset was further reduced.

For evaluating the potential of NMFBFS-inferred optimal feature subset, we firstly tested the classification accuracy of three candidate feature subsets FS_1, FS_2, and OFS on the training set (120 representative samples). FS_1 and FS_2 were generated by feature selection with the threshold θ (0.95). OFS denoted 49 original symptom features in the dataset D_R. Table 6 indicates that the 39 optimal features, which covered 33 original symptom features and 6 new mixed features, result in the best classification accuracy on the training samples. The performance of FS_2 was much better than OFS; however, it was still worse than FS_1 because the new mixed features also have important contributions to classification.

We then compared the performance of our NMFBFS with three well-known feature selection methods (ReliefF [11], mRMR [12], and Elastic Net [13]). ReliefF was implemented

Table 5: The NMF-driven potential clinical features of HCC (threshold: 0.95).

Basis components	Original features	Mixed features	Description about mixed features
Basis 1	V_7; V_{11}; V_{13}; V_{31}; V_{33}; V_{32}; V_{49}; V_{18}; V_{50}; V_9; V_{34}	M_{11} M_{12}	Converted from $\{V_{38}, V_{27}, V_{20}\}$ and $\{V_{19}, V_{15}\}$, respectively.
Basis 2	V_{22}; V_{17}; V_3; V_{57}; V_2; V_{10}; V_4; V_{52}; V_5; V_{14}	M_{21} M_{22}	Converted from $\{V_{35}, V_{54}\}$ and $\{V_6, V_8, V_{53}, V_{37}, V_{28}\}$, respectively.
Basis 3	V_{24}; V_{23}; V_{30}; V_{26}; V_{42}; V_{12}; V_{40}; V_{36}; V_{25}; V_{21}; V_{39}; V_{46}	M_{31} M_{32}	Converted from $\{V_{48}, V_{29}\}$ and $\{V_{44}, V_{45}\}$, respectively.
Number of features	33	6	Total: 39 features

Table 6: Classification accuracy among three feature subsets on the training set (120 representative samples). FS_1 was obtained by the proposed approach with a given threshold ($\theta = 0.95$), in which 33 original symptom features and 6 new mixed features were included. FS_2 denotes the above-mentioned 33 original symptom features ($FS_2 \subset FS_1$). OFS indicates all the 49 symptoms before NMF calculation.

Feature subsets	Dimension	Classification accuracy in LSSVM (%)
FS_1	39	80.002 ± 9.95
FS_2	33	77.50 ± 12.36
OFS	49	72.50 ± 11.64

Table 7: Classification accuracy of inferred optimal feature subset via NMFBFS, ReliefF, mRMR, and Elastic Net on the training set.

Methods	Feature subset	Dimension	Classification accuracy in LSSVM (%)
NMFBFS	**FS_1**	**39**	**80.002 ± 9.95**
ReliefF	FS_{RF20}	20	65.00 ± 10.03
	FS_{RF40}	40	73.33 ± 15.76
mRMR	FS_{MR20}	20	70.83 ± 12.5
	FS_{MR40}	40	74.17 ± 9.03
Elastic Net	FS_{EN20}	20	70.00 ± 11.56
	FS_{EN40}	40	76.67 ± 10.46

Table 8: Classification accuracy of inferred optimal feature subset via NMFBFS, ReliefF, mRMR, and Elastic Net on the testing set.

Methods	Feature subset	Dimension	Classification accuracy in LSSVM (%)
NMFBFS	**FS_1**	**39**	**79.65 ± 6.48**
ReliefF	FS_{RF20}	20	50.71 ± 1.22
	FS_{RF40}	40	76.43 ± 8.27
mRMR	FS_{MR20}	20	63.79 ± 1.22
	FS_{MR40}	40	77.14 ± 9.18
Elastic Net	FS_{EN20}	20	67.57 ± 4.09
	FS_{EN40}	40	78.38 ± 9.62

using MATLAB function. "mRMRe" and "elasticnet" *R* packages were applied for mRMR and Elastic Net based feature selection, respectively. Supplementary Figure S1 represents the ReliefF-based feature ranking. Supplementary Figure S2 represents the Elastic Net ($\lambda = 0.5$) solution paths for feature selection. We selected Top 20 features and Top 40 features as two candidate feature subsets for each method to evaluate their classification performances: FS_{RF20} and FS_{RF40} generated from ReliefF; FS_{MR20} and FS_{MR40} inferred from mRMR; FS_{EN20} and FS_{EN40} inferred from Elastic Net. Table 7 represents the classification performance of the above six candidate feature subsets and the NMFBFS-derived optimal feature subset FS_1 on the training set (120 representative samples). The results indicate that NMFBFS-inferred feature subset has the best classification accuracy in training samples.

Except 120 representative training samples which were screened out to implement the NMF analysis, the remaining samples can be used to test the classification accuracy of optimal feature subset. We randomly selected 40 samples (10 : 20 : 10 for each clinical stage) from the rest of the samples and then evaluated the classification accuracy of inferred feature subset by each method (NMFBFS, ReliefF, mRMR, and Elastic Net). Table 8 shows the differences among all these methods, and it can be found that the optimal feature subset inferred by our proposed method has the best generalization performance.

Finally, the more important thing is that the selection of threshold θ determines how many groups of redundant symptoms will be screened out. Here, we further discussed the effects of threshold θ to the optimal feature subsets on the classification performance. Table 9 shows the differences among three optimal feature subsets inferred by the proposed approach with different values for threshold θ. From Table 9, we can obviously see that the bigger value of θ will screen redundant symptoms strictly, which leads to less similar symptoms that would be obtained. With a smaller value of θ, much more symptoms can be categorized into the same groups; hence, the original feature space will be sharply reduced by our approach. Table 9 denotes that, with the decrease of θ, the size of optimal feature subset was narrowed down but the classification accuracy was also decreased. These results suggest that a bigger value of θ will result in less redundant symptoms and therefore induce a larger size of optimal feature subset; oppositely, smaller θ can provide more redundant symptoms and sharply reduce the feature dimension. An extreme case is that θ equals "0," which means that we can get one mixed feature for each basis and the size of optimal feature subset is equal to the number of bases. In a word, how to determine the value of θ depends on the size of optimal feature subset and its corresponding classification performance.

Table 9: The performance of classification for the inferred optimal feature subsets with different threshold θ.

The values of θ	Original symptom features	New mixed features	Total number of features	Classification accuracy (%)
$\theta = 0.95$	33	6	39	80.002 ± 9.95
$\theta = 0.90$	21	9	30	70.83 ± 6.59
$\theta = 0.85$	10	8	18	70.00 ± 4.56

5. Conclusions

In this study, we developed the NMFBFS approach to efficiently extract the important clinical symptoms of HCC from clinical observation data. NMFBFS is a two-stage filter method for feature selection as follows. (1) In the first stage, preliminary screening is implemented to detect and remove the irrelevant features; (2) in the second stage, NMF was applied to identify the redundant features by groups which might represent similar distribution patterns. Each redundant symptom group was then transformed into a new mixed feature so that the dimension of dataset was further reduced.

The application of NMFBFS on a clinical dataset of HCC proved the effectiveness of this approach. The optimal clinical features derived from NMFBFS approach contained many well-recognized symptoms of HCC patients. Moreover, this study also provides a general computational framework of a novel feature selection approach to efficiently extract the optimal feature subset from a high-dimensional dataset.

Abbreviations

HCC: Hepatocellular carcinoma
TCM: Traditional Chinese Medicine
NMF: Nonnegative matrix factorization
LSSVM: Least squares support vector machines
KNN: K-nearest neighbor.

Conflict of Interests

The authors declare that they have no competing interests.

Authors' Contribution

Zhiwei Ji and Guanmin Meng contributed equally to this work.

Acknowledgments

This work was supported by the National Science Foundation of China (nos. 61472282 and 61133010). The data in this work was collected by the Changhai Hospital, Shanghai, China.

References

[1] F. X. Bosch, J. Ribes, R. Cléries, and M. Díaz, "Epidemiology of hepatocellular carcinoma," *Clinics in Liver Disease*, vol. 9, no. 2, pp. 191–211, 2005.

[2] M. M. Center, A. Jemal, R. A. Smith, and E. Ward, "Worldwide variations in colorectal cancer," *CA—Cancer Journal for Clinicians*, vol. 59, no. 6, pp. 366–378, 2009.

[3] H. B. El-Serag, "Hepatocellular carcinoma," *The New England Journal of Medicine*, vol. 365, no. 12, pp. 1118–1127, 2011.

[4] "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 and J. Bruix, "Diagnosis of hepatic nodules 20 mm or smaller in cirrhosis: prospective validation of the noninvasive diagnostic criteria for hepatocellular carcinoma—reply," *Hepatology*, vol. 47, no. 6, pp. 2146–2147, 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 and 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, Article ID bbr015, pp. 20–33, 2012.

[9] Z. Ji and B. Wang, "Identifying potential clinical syndromes of hepatocellular carcinoma using PSO-based hierarchical feature selection algorithm," *BioMed Research International*, vol. 2014, Article ID 127572, 12 pages, 2014.

[10] J.-X. Du, C.-M. Zhai, and Y.-Q. Ye, "Face aging simulation and recognition based on NMF algorithm with sparseness constraints," *Neurocomputing*, vol. 116, pp. 250–259, 2013.

[11] J. N. Liang, S. Yang, and A. Winstanley, "Invariant optimal feature selection: a distance discriminant and feature ranking based solution," *Pattern Recognition*, vol. 41, no. 5, pp. 1429–1439, 2008.

[12] H. C. Peng, F. H. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

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

[14] S. Wildi, B. C. Pestalozzi, L. McCormack, and P.-A. Clavien, "Critical evaluation of the different staging systems for hepatocellular carcinoma," *The British Journal of Surgery*, vol. 91, no. 4, pp. 400–408, 2004.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] R. Zdunek and A. Cichocki, "Nonnegative matrix factorization with constrained second-order optimization," *Signal Processing*, vol. 87, no. 8, pp. 1904–1916, 2007.

[20] Z. Chang, Z. Wang, C. Ashby et al., "eMBI: boosting gene expression-based clustering for cancer subtypes," *Cancer Informatics*, vol. 13, supplement 2, pp. 105–112, 2014.

[21] C.-H. Zheng, D.-S. Huang, L. Zhang, and X.-Z. Kong, "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 599–607, 2009.

[22] C.-H. Zheng, T.-Y. Ng, L. Zhang, C.-K. Shiu, and H.-Q. Wang, "Tumor classification based on non-negative matrix factorization using gene expression data," *IEEE Transactions on Nanobioscience*, vol. 10, no. 2, pp. 86–93, 2011.

[23] A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi, "Non-negative matrix factorization with α-divergence," *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1433–1440, 2008.

[24] R. Zdunek and A. Cichocki, "Nonnegative matrix factorization with quadratic programming," *Neurocomputing*, vol. 71, no. 10–12, pp. 2309–2320, 2008.

[25] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '01)*, 2001.

[26] C. Theys, H. Lantéri, and C. Richard, "SGM to solve NMF—application to hyperspectral data," in *New Concepts in Imaging: Optical and Statistical Models*, vol. 59 of *EAS Publications Series*, pp. 357–379, 2013.

[27] G. Casalino, N. del Buono, and C. Mencar, "Subtractive clustering for seeding non-negative matrix factorizations," *Information Sciences*, vol. 257, pp. 369–387, 2014.

[28] L. D. Vignolo, D. H. Milone, and J. Scharcanski, "Feature selection for face recognition based on multi-objective evolutionary wrappers," *Expert Systems with Applications*, vol. 40, no. 13, pp. 5077–5084, 2013.

[29] A. Anand, G. Pugalenthi, G. B. Fogel, and P. N. Suganthan, "An approach for classification of highly imbalanced data using weighting and undersampling," *Amino Acids*, vol. 39, no. 5, pp. 1385–1391, 2010.

[30] A. Bria, N. Karssemeijer, and F. Tortorella, "Learning from unbalanced data: a cascade-based approach for detecting clustered microcalcifications," *Medical Image Analysis*, vol. 18, no. 2, pp. 241–252, 2014.

[31] P. Cao, D. Z. Zhao, and O. Zaiane, "Hybrid probabilistic sampling with random subspace for imbalanced data learning," *Intelligent Data Analysis*, vol. 18, no. 6, pp. 1089–1108, 2014.

[32] A. Shubair, S. Ramadass, and A. A. Altyeb, "KENFIS: kNN-based evolving neuro-fuzzy inference system for computer worms detection," *Journal of Intelligent and Fuzzy Systems*, vol. 26, no. 4, pp. 1893–1908, 2014.

[33] H.-Q. Wang, F.-C. Sun, Y.-N. Cai, L.-G. Ding, and N. Chen, "An unbiased LSSVM model for classification and regression," *Soft Computing*, vol. 14, no. 2, pp. 171–180, 2010.

[34] Z. Mustaffa and Y. Yusof, "LSSVM parameters tuning with enhanced artificial bee colony," *International Arab Journal of Information Technology*, vol. 11, no. 3, pp. 236–242, 2014.

[35] Y. Li and A. Ngom, "The non-negative matrix factorization toolbox for biological data mining," *Source Code for Biology and Medicine*, vol. 8, no. 1, article 10, 2013.

[36] 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.

[37] A. Frigyesi and M. Höglund, "Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes," *Cancer Informatics*, vol. 6, pp. 275–292, 2008.

[38] L. F. Bo, L. Wang, and L. C. Jiao, "Multiple parameter selection for LS-SVM using smooth leave-one-out error," in *Advances in Neural Networks—ISNN 2005*, vol. 3496 of *Lecture Notes in Computer Science*, pp. 851–856, Springer, Berlin, Germany, 2005.

[39] S.-W. Lai, H.-J. Chen, C.-L. Lin, and K.-F. Liao, "No correlation between Alzheimer's disease and risk of hepatocellular carcinoma in older people: an observation in Taiwan," *Geriatrics & Gerontology International*, vol. 14, no. 1, pp. 231–232, 2014.

[40] S.-M. Ou, Y.-J. Lee, Y.-W. Hu et al., "Does Alzheimer's disease protect against cancers? A nationwide population-based study," *Neuroepidemiology*, vol. 40, no. 1, pp. 42–49, 2012.

[41] S.-Y. Peng, X.-D. Feng, Y.-B. Liu et al., "Surgical treatment of hepatocellular carcinoma originating from caudate lobe," *Zhonghua Wai Ke Za Zhi*, vol. 43, no. 1, pp. 49–52, 2005.

[42] S. Y. Peng, J. T. Li, Y. B. Liu et al., "Surgical treatment of hepatocellular carcinoma originating from caudate lobe—a report of 39 cases," *Journal of Gastrointestinal Surgery*, vol. 10, no. 3, pp. 371–378, 2006.

[43] M.-H. Lin, P.-Y. Wu, S.-T. Tsai, C.-L. Lin, T.-W. Chen, and S.-J. Hwang, "Hospice palliative care for patients with hepatocellular carcinoma in Taiwan," *Palliative Medicine*, vol. 18, no. 2, pp. 93–99, 2004.

[44] S. Fujiyama, J. Shibata, S. Maeda et al., "Phase I clinical study of a novel lipophilic platinum complex (SM-11355) in patients with hepatocellular carcinoma refractory to cisplatin/lipiodol," *British Journal of Cancer*, vol. 89, no. 9, pp. 1614–1619, 2003.

[45] X. Yu, H. Zhao, L. Liu et al., "A randomized phase II study of autologous cytokine-induced killer cells in treatment of hepatocelluar carcinoma," *Journal of Clinical Immunology*, vol. 34, no. 2, pp. 194–203, 2014.

[46] K. K. Ciombor, Y. Feng, A. B. Benson III et al., "Phase II trial of bortezomib plus doxorubicin in hepatocellular carcinoma (E6202): a trial of the Eastern Cooperative Oncology Group," *Investigational New Drugs*, vol. 32, no. 5, pp. 1017–1027, 2014.

[47] J. Wu, C. Henderson, L. Feun et al., "Phase II study of darinaparsin in patients with advanced hepatocellular carcinoma," *Investigational New Drugs*, vol. 28, no. 5, pp. 670–676, 2010.

[48] J.-J. Lin, C.-N. Jin, M.-L. Zheng, X.-N. Ouyang, J.-X. Zeng, and X.-H. Dai, "Clinical study on treatment of primary hepatocellular carcinoma by Shenqi mixture combined with microwave coagulation," *Chinese Journal of Integrative Medicine*, vol. 11, no. 2, pp. 104–110, 2005.

[49] M. Doffoël, F. Bonnetain, O. Bouché et al., "Multicentre randomised phase III trial comparing Tamoxifen alone or with Transarterial Lipiodol Chemoembolisation for unresectable hepatocellular carcinoma in cirrhotic patients (Federation Francophone de Cancerologie Digestive 9402)," *European Journal of Cancer*, vol. 44, no. 4, pp. 528–538, 2008.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 896176, 10 pages
http://dx.doi.org/10.1155/2015/896176

Review Article

Comparative Transcriptomes and EVO-DEVO Studies Depending on Next Generation Sequencing

Tiancheng Liu,[1] Lin Yu,[2] Lei Liu,[3] Hong Li,[1] and Yixue Li[1,3]

[1]*Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China*
[2]*Key Laboratory of Contraceptive Drugs and Devices of National Population and Family Planning Commission of China, Shanghai Institute of Planned Parenthood Research, 2140 Xietu Road, Shanghai 200032, China*
[3]*Shanghai Center for Bioinformation Technology, 1278 Keyuan Road, Shanghai 201203, China*

Correspondence should be addressed to Hong Li; honglibio@gmail.com and Yixue Li; yxli@sibs.ac.cn

Received 17 May 2015; Accepted 15 June 2015

Academic Editor: Lin Lu

High throughput technology has prompted the progressive omics studies, including genomics and transcriptomics. We have reviewed the improvement of comparative omic studies, which are attributed to the high throughput measurement of next generation sequencing technology. Comparative genomics have been successfully applied to evolution analysis while comparative transcriptomics are adopted in comparison of expression profile from two subjects by differential expression or differential coexpression, which enables their application in evolutionary developmental biology (EVO-DEVO) studies. EVO-DEVO studies focus on the evolutionary pressure affecting the morphogenesis of development and previous works have been conducted to illustrate the most conserved stages during embryonic development. Old measurements of these studies are based on the morphological similarity from macro view and new technology enables the micro detection of similarity in molecular mechanism. Evolutionary model of embryo development, which includes the "funnel-like" model and the "hourglass" model, has been evaluated by combination of these new comparative transcriptomic methods with prior comparative genomic information. Although the technology has promoted the EVO-DEVO studies into a new era, technological and material limitation still exist and further investigations require more subtle study design and procedure.

1. Introduction

Evolutionary developmental biology (EVO-DEVO) studies how the dynamics of development affects the phenotypic variation arising from genetic variation and its correlation with phenotypic evolution. In this subject there is a central issue, which is the most conserved period or the crucial section during the entire developmental process of an organism. To solve this issue, morphological studies, which are the major approach in developmental biology, have been conducted on different species in past years. However, these traditional observation methods are not sufficient for the requirement of precise quantification analysis. In such a demand, comparative transcriptomic studies have been utilized in these studies and generate some models about the evolutionary pressure of embryonic development.

Next generation sequencing technology has largely improved the scale of comparative genomics studies by the high throughput detection of gene sequences, which makes the assembly of new genome easy. Besides, not only have the comparative genomics studies with case-control studies design reached a new level, but also the evolution studies based on genome sequences of multiple species have been feasible. When comparative transcriptomic studies of embryo development are equipped with this powerful tool, it also has generated unprecedented revolution in EVO-DEVO field and improved the resolution from macro to micro view. Several strategies have been proposed to illustrate the existing models of selective pressure acting on embryonic development, which provide further understanding for the divergence of morphogenesis.

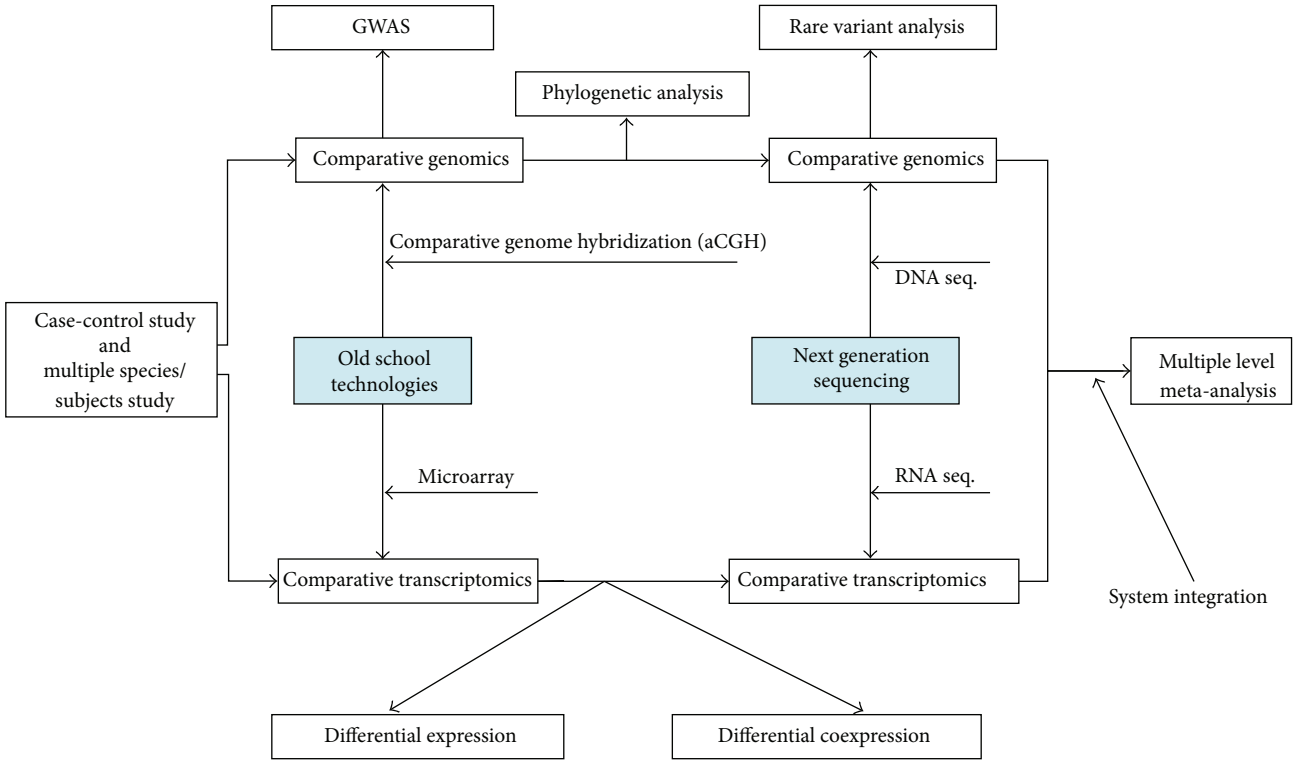


Figure 1: Illustration of the comparative genomics and comparative transcriptomics based on case/control study design are conducted with old school technologies and next generation sequencing technology. This figure shows the main concepts in the first part of this paper.

2. Comparative Genomic and Comparative Transcriptomic Study

2.1. Comparative Genomics: From Case-Control to Multiple Species. Case-control study design is widely adopted in epidemiology for investigating the relationship between disease and exposure and it is the initial principle of comparative studies. In genomic studies, this design works efficiently for the comparison of two objects and it aims to illustrate relationship between the phenotypic difference and the genetic difference. Phenotypic difference stands for disease while genetic difference stands for exposure in terms of epidemiology. From the genomic opinion, the genetic differences were variants between case and control samples. Many genome-wide association analysis studies (GWAS) also employ the case-control study design to examine the potential effects of genetic variants among populations [1–3], which has promoted the understanding of many kinds of diseases [4].

During the microarray era, there are many comparative genomic studies which adopted array comparative genome hybridization (aCGH) technology to determine copy number variations (CNVs) [5] or oligonucleotide array technology to investigate single nucleotide polymorphisms (SNPs) [6]. Along with the birth of next generation sequencing (NGS) technology, these microarray based technologies have been replaced as they are not convenient to acquire any interested genome sequences of organism as sequencing. Besides some de novo sequencing works [7, 8], most studies tend to conduct the resequencing procedure with case-control study design [9, 10]. It is meaningful to sequence comparable subjects and detect the underlying genetic difference, such as the fact that Atanur et al. have discovered the likely cellular basis of hypertension by comparing the genome of SHR strain rat with BN rat reference genomes [11]. The Bactrian Camels Genome Sequencing and Analysis Consortium have identified characters of domestication in camel by comparing the heterozygosity rate of wild and domestic Bactrian camels [12]. In the light of NGS technology, GWAS also have transformed from common variants to rare variants (Figure 1).

The case-control study design is narrow sense of comparative genomic as it is unnecessary to limit the comparison between two objects. Multiple objects comparison involves intraspecies comparison and interspecies comparison designed for different purpose. Intraspecies comparison intends to discover the strains diversity for specific species or the variation in population for certain species. The STAR Consortium has used SNP array to illustrate the diverse genetic background of different inbred laboratory rat strains [13], and the follow-on work has been conducted by Atanur et al. depending on next generation sequencing [14]. Similar study has been conducted to study the artificial selection during chicken domestication [15]. Navin et al. have applied single-nucleus sequencing to investigate tumor population structure and evolution in human breast cancer [16]. Actually the comparison between multiple objects is crucial especially in evolutionary analysis [17]. The interspecies comparison

focuses on evolutionary analysis which examines the selective constraints acting on sequence of genome. Zhang et al. have compared the expansion or contraction of gene families between two bats and other eight mammalian species to reveal the genetic and evolutionary background for the functional characters of bat [18]. Besides, many studies trace certain species in phylogeny based on similarity of ortholog sequences between the studied and several known species [12, 19–21].

2.2. Comparative Transcriptomic Methods: Differential Expression and Differential Coexpression. Sequencing technology gives great impetus to comparative genomic studies, while the sequencing object is far beyond the DNA sequence. Capturing transcripts in cell makes the RNAs also available to the sequencing platform, which is used for quantification of the expression or detection of alternative splicing events. Sequencing technology also has improved comparative transcriptomic studies as it has produced plenty RNA data for transcriptomic investigations (Figure 1).

The traditional comparative transcriptomics are also based on the case-control study design, in which the gene expressions of several samples for each group are measured and statistical tests are adopted to examine the differential gene expression between case and control subjects. The differentially expressed genes are considered to be associated with phenotypic divergence between compared objects and they have potential to be the candidate biomarker of case situation. Recently, in the light of the high throughput technology such as microarray or RNA sequencing, expressions of 10 thousand genes can be detected at the same time. The big advance expands the scale of expression detection but also leads to the problem of multiple comparisons. The problem reduces statistical power so that several genes with expression change are neglected. Beside the problem of multiple comparisons, differential expression analysis also is defectiveness in following network analysis. For instance, in order to study their functions, the differentially expressed genes are always aligned onto the interaction network which is built by prior knowledge of protein interactions, which will not discover the new connections of genes.

Considering these deficiencies, it is necessary to further mine the information hiding in the expression matrix, which prompts the birth of differential coexpression analysis focusing on the switch of the links between genes rather than the changes of expression values for a single gene between samples [22]. In the system of organism, genes are organized into networks rather than separated, and genes are always linked to regulators such as TF, which lead to two genes regulated by the same TF exhibiting correlation in their expression profiles. The correlation of gene pair varies in different condition because the regulation relationship between genes will switch when the organism is exposed to different situation. Based on this principle, with several samples measured in case and control group, respectively, we are able to measure the correlation coefficients of every gene pair in each group. By comparing these correlation coefficients between case and control groups, the differential coexpression gene pairs can be identified. The differential coexpression approach not only complements the result of differential expression analysis but also enables the identification of rewiring events in the gene regulation network (GRN).

2.3. Annotation of Regulatory Element: The Integration of Genomics and Transcriptomics. Next generation sequencing not only prompts the efficiency of genomic research [23] but also facilitates the construction of genomic libraries for populations [24]. However, for the accumulation of sequences we have found abstruse information associated with biological function underlining the genome sequences. In order to further understand the biological function, we need to analyze the regulatory mechanism of the genomic elements, which lead to the transformation from comparative genomics to comparative transcriptomics. In such kind of demand, the Encyclopedia of DNA Elements (ENCODE) project and Model Organism Encyclopedia of DNA Elements (modENCODE) project have born and focus on annotation of the regulatory elements in genomes including human, mouse, fly, and worm [25–28]. They have profiled several crucial features in transcriptome such as the binding sites of transcription factors (TFs), epigenomic modifications, and gene expression levels for these species, which provide plentiful datasets for transcriptomic analysis. Depending on the profiles of epigenomic modifications, Ernst et al. have classified the human chromatin into 15 kind states which represent the activated conditions [29]. Based on the binding of TFs, Yip et al. have used machine learning approach to discriminate genomic regions [30]. By correlating epigenomic modifications on the cis-regulation region and gene expression in each species, Cheng et al. have proved that gene expression is predictable by chromatin features in fly and worm [31]; at the same time Dong et al. also model gene expression levels by histone modification profiles in human cell lines [32]. Finally, a universal model has been proved for epigenomic modifications on cis-regulation region to predict gene expression in these three species [33].

Although not every organization is able to produce such diversiform datasets, the integration of genomic information with transcriptomic information has been adopted by many investigators. These studies have made difference in understanding the regulated elements in genomic sequences. The integration of multiple levels, which also represents the trend of omic study nowadays, is based on the hypothesis that switches in higher level will influence the lower level which is coordinated with the Central Dogma. In other words, it proposes that the genomic mutations in gene sequence will lead to the change of expression level of downstream genes. Applying this principle, Akavia et al. have developed an algorithm to identify the casual genetic aberrations in cancer through associating chromosomal copy number variation (CNV) and gene expression data [34]. Kim et al. have identified potential causal genes by combining the expression Quantitative Trait Loci (eQTL) analysis with pathway information [35]. The integration of multiple level data not only increases the utilization of datasets but also ensures the reliability of result. It is wildly adopted in biological investigation nowadays, especially in studies of cancer which are conducted by The Cancer Genome Atlas (TCGA) [36, 37].

In summary, as a branch of computational science, bioinformatics has been promoted by the coming of big data era. More and more datasets will be generated by consortium like ENCODE and TCGA, and the meta-analysis will still be the trend in future.

3. EVO-DEVO Studies for Understanding the Morphological Diversity of Species

3.1. From Macro to Micro: Morphological Study to Gene Study. The development process of animals has been proposed to be under stringent selective pressure in order to ensure the precision of the process. The evolutionary pressure constrains the phenotypic diversity of embryo for different organism at certain degree, which leads to the morphological similarities at some stages of embryo development for different species. And the extents of embryo similarities between species are diverse during development process, which enables development biologists to examine the fluctuations of evolutionary pressure acting on different embryo stages. Development biologists have used this embryo morphological comparison method to study organism development for many years. For instance, von Baer's third law has proposed that the earlier development stages are highly similar between different species and the embryos gradually present divergence from each other during ontogeny [38]. Ontogenic stages stand for developmental process in contrast with phylotypic period, in which the morphology of embryos from different species represents such a high similarity that these development stages are considered to recapture the phylogeny during evolution. As above mentioned, discriminating ontogenic stages from phylotypic stages are central issue in the EVO-DEVO studies. However, a defect of the morphological comparison method is that it is difficult to quantize the morphological features, which would cause problem using nonquantitative morphological characteristics to evaluate the quantitative degree of conservation. And the stages with certain morphological characters are various in multiple phylum, which limits the morphological comparison which only can be conducted in a certain phylum. Taken together, these will confuse the definite detection of selective constraints acting on stages in multiple species.

Along with the advance of the technologies in molecular biology, development biologists have been able to analyze the development stages from microcosmic view. For instance, Duboule has found that the expression of Hox genes is a feature of the phylotypic stages [39]. The information from molecular comparison provides more precise identification of patterns for ontogenic stages and phylotypic stages in embryo development as it can produce the quantitative information. Until recently, new high throughput technologies, which possess more accurate quantitative characteristics, have been applied to development studies. Depending on microarray, Vassena et al. have examined gene expression in human preimplantation development [40], and the expression profile of whole development time series for zebra fish has been inspected by Domazet-Lošo and Tautz [41]. RNA sequencing method also has been adopted to address the expression profile of development in multiple species including fly [28], worm [27], human, and mouse [42].

Advance in the technology enables the EVO-DEVO studies from macro to micro. From microcosmic view, development biologists would further decipher possible evolutionary mechanism underlying the hypothesis, which is more challenging and meaningful. These molecular level studies are thought to be superior compared with the morphological approaches, as the information of gene sequences is more close to the inherited entities compared with the morphology. However, it is still a controversial problem for the discrimination of ontogenic stages and phylotypic stages in embryo developmental process for multiple species. New high throughput technology has potential to distinguish these stages depending on comparative transcriptomic analysis, which would further contribute to understanding the underlying molecular and evolutionary mechanism of development.

3.2. Two Controversial Models about the Constraints on Development. Above we have mentioned the controversial partition about the ontogenic stages and phylotypic stages during development stages, which can be illustrated as problem of defining phylotypic stages in certain period of development. Phylotypic stages are supposed to be development stages with high similarity among different species, in which the features of nature selection such as gene expression or gene sequence should present to be conserved. In ontogenic stages, species specific differentiation happens and features in these stages should be less conserved. In particular, because of their conserved feature, evolutionists, who intend to label these stages in certain developmental period for understanding the evolution of development, which is the central issue of EVO-DEVO studies, are also interested in phylotypic stages.

Organism developmental process can be classified into three dominant periods: earlier stages marked by the prominent event, zygote genome activation (ZGA) [42], middle stages when Hox genes start expression [43], and late stages in which morphological formation starts [44]. The late stages are unanimous to be the most nonconserved because embryos of different species already present diversity in these stages, whether morphological divergence or variations of gene expression. Although Tian et al. have found that the late stages show the strongest conservation and weakest evolvability in the slime mold *Dictyostelium* [45], these stages are still considered to be less conserved among most organisms, especially vertebrate. Besides this rare case of slime mold, two canonical evolutionary models of development have been proposed: the "funnel-like" model, in which it is supposed that the earlier embryo stages are the most conserved, and the "hourglass" model, in which the middle stages of development are imposed with the strongest evolutionary constraints [46].

The "funnel-like" model, which describes the shape of selective constraints acting on development as a funnel (Figure 2), has been rooted in von Baer's third law. This law suggests that the selective constraints gradually decrease during the development and the earlier stages of development are under most stringent selective pressure. The development

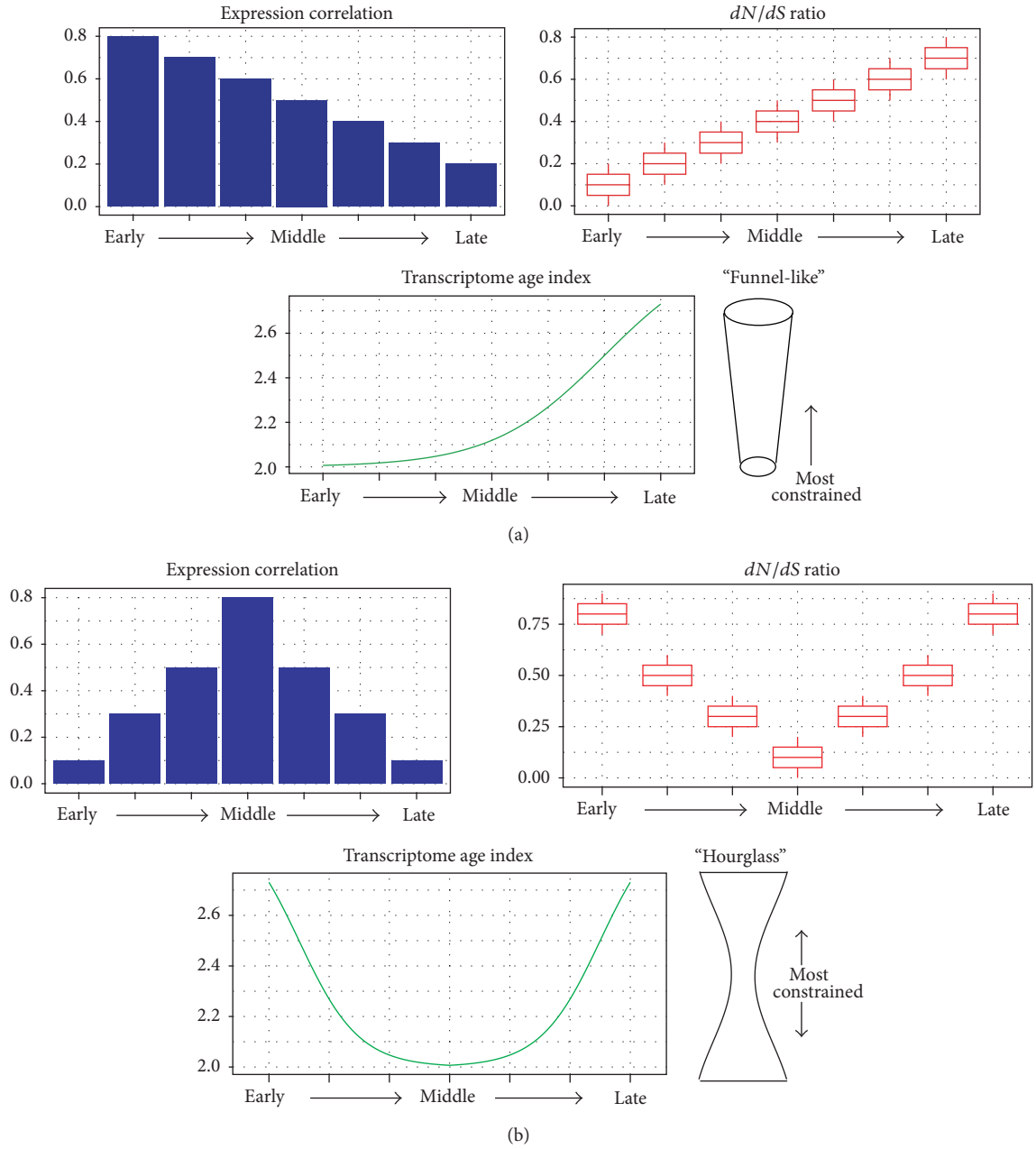


Figure 2: Illustration of the two major models about the selective pressure of embryonic development and their measurement. Pictorial charts in the right side stand for the "funnel-like" model and "hourglass" model. Histograms, boxplots, and lines from left to right stand for the three kinds of measurement of selective pressure based on transcriptomic data. In the figure of expression correlation, higher expression correlation means transcriptome similarity, which is the signature of conservation. In the figure of dN/dS ratio, lower dN/dS ratio means the conserved gene sequences. For the computation of transcriptome age index, ancient genes are labeled with small numbers while young genes are labeled with big numbers. Therefore, lower transcriptome age index of a stage means more ancient genes expressing in corresponding stages in the figure of transcriptome age index.

process starts from a single zygote cell, along with cell division occurring; it forms blastocyst which is composed of multiple cells with different fates. This process looks very similar to the evolution of creature, which starts from single cell to multiple cell. Therefore, phylotypic stages are thought to recapture the phylogeny in evolutional history and the development process is supposed to be an expand procedure from simple to complex. The earlier stages, which are considered to be simplex, should be exposed under strict selection so that the later developmental program can be subtly executed. This is in concordance with the developmental burden hypothesis [47], which has assumed that earlier elements in embryo are

responsible for downstream development infrastructure so that earlier stages tend to be evolutionarily conserved.

The "hourglass" model, which assumes the mid-embryonic stages, shows the most stringent constraints and the shape of constraints looks like an hourglass with two wide sides and a narrow middle (Figure 2). This model initially depends on the functional importance and complicated regulation network of middle stages, in which the Hox genes express and the embryo forms body plans [39]. This fundamental process is considered to be such a crucial infrastructure in embryo development that perturbation during these stages will cause tremendous influence on organogenesis. In old school embryo morphological time, although some alternative models have been proposed [48–51], the hourglass model has been validated by observations of morphological traits in multiple vertebrate embryos [52–55]. In recent years, the embryo development stages have been profiled by parallel sequencing so that the hourglass can be examined in gene level. By comparing the expression profile in embryo development for turtle and chicken, Wang et al. have validated the hourglass model in development of these two species [21].

There are two major approaches in comparative transcriptomic studies to illustrate the hourglass model or funnel-like model, which we will discuss in later part (Figure 2).

4. Comparative Transcriptomics for the Embryo Developmental Studies

4.1. Correlation of Gene Expression Methods. Based on the case-control study design, an intuitive measure for the conservation between two objects is to compare the similarity of their gene expression. The comparison of expression profile is conducted on one-to-one ortholog genes, which maintain single copy and usually are considered to possess the same biological functions in corresponding species. Therefore, the expression pattern of one-to-one ortholog gene pair should present certain degree of similarity. In definition of this method, conservation is measured by computing the correlation coefficients between all pairs of expressed one-to-one ortholog genes in each developmental stage. The levels of conservation are determined depending on the summary of all the correlation coefficients in each stage. The high correlation coefficients of a certain stage indicate conserved gene expression in this stage, which should be considered under strong selective pressure (Figure 2). As comparative genomics studies can be conducted between or within species, the comparative transcriptomics studies also can illustrate the diversity of gene expression within species or between species. Kalinka et al. have used these comparative transcriptomics studies to examine the correlation of gene expression within six sequenced *Drosophila* species [56], and Ninova et al. have detected the correlation of microRNA expression within two divergent fruit flies [57]. Both of these studies prove the broad existence of hourglass model for multiple kinds of transcripts within *Drosophila* species. As evidence for the hourglass model holding between species, the study of Wang et al. has been conducted on two different species [21]. The study of Irie and Kuratani proves the common existence of hourglass in vertebrate by comparing the expression profiles of four species [46]. These pioneering investigations have successfully applied expression correlation approach in EVO-DEVO studies, which proves comparative transcriptomic approach is powerful in evolutionary study.

The transcriptomic similarity method also has some defects, such as the fact that the correlations are examined only depending on a part of the whole transcriptome (one-to-one ortholog genes) and the computation must be conducted on two subjects/species. One-to-one ortholog genes only account for part of expression signature in each one of the compared objects. Particularly for studies conducted between distant species, the proportion of one-to-one ortholog genes becomes even smaller. It results in loss of expression information, which will further affect the conclusion. In particular, the evolutionary distance between two objective organisms is not in direct proportion to the loss of expression information, which will cause the difficulties in different studies using pairs of species with various evolutionary distances. Besides the problem of losing information, another difficulty is the choice of corresponding development time points in paired organisms. Only development stages of two species are aligned in corresponding development time points; the correlation coefficients can be computed in each of the aligned stages. However, the developmental time varies between species, which makes it difficult to find the precise alignments of stages. To solve this problem, investigators have adopted enumeration method which computes the correlation coefficients between paired stages in all-to-all manner [7, 56]. Enumeration method handles the problem of corresponding stage choice, but it will introduce artificial decision especially in the case in which one species has multiple corresponding stages in the other species, such as the dual alignment in fly and worm development stages found by Gerstein et al. [33].

4.2. Evolutionary Indices Based Methods. We have discussed the comparative transcriptomic method based on the correlation and its two major limitations above. This approach presents an oversimplified procedure. It not only neglects the information of nonortholog genes but also does not utilize the prior knowledge. Prior knowledge of conservation is contained in the sequence of expressed genes during each developmental stage. Such kind of knowledge has been evaluated by prior comparative genomic studies [58]. For instance, each gene has unique date of birth in the phylogeny which means a specific gene has been born in certain ancient time. Age information of gene should be applied to studies. In the evolutionary indices based approach, first the gene expression of a certain species during embryonic development has been profiled. Then a specifically activated set of genes have been identified for each developmental stage and the age indices of each gene set are used to measure the conservation of corresponding stage. Depending on the age index of genes, Domazet-Lošo et al. have developed a phylostratigraphy approach to specify different phylostratum for genes expressing in ectoderm, endoderm, or mesoderm of *D. melanogaster* embryo [59]. The principle of phylostratigraphy approach is labeling ancient genes with small numbers and

Table 1: Comparison of two detection approaches from different aspects.

	Sample	Prior knowledge	Advantages	Defects
Correlation of gene expression methods	Paired	No	Interspecies evaluation	Loss of information
Evolutionary indices based methods	One	Yes	Integration analysis	Single species evaluation

young genes with big numbers so that phylogenetic ages of genes are quantified. It also has been used to study the relationship between multicellularity and the origin of cancer [60]. Based on this approach, Domazet-Lošo et al. have further proposed a transcriptome age index (TAI), which combines phylostratigraphy and stage-specific gene expression information by multiplication, to evaluate the selective pressure on stages of zebra fish development [41]. Not only has this study proved the hourglass in zebra fish, but also another study of *Arabidopsis thaliana* embryogenesis, in which the conservation has also been measured by TAI, has showed the existence of hourglass in Plantage [61]. Depending on the transcriptomic information, TAI measures the relative proportion of ancient genes and young genes in a specific developmental stage (Figure 2). Such kind of approach represents the combination of prior comparative genomic knowledge with the gene expression information between different development stages within a species.

Compared with transcriptome similarity method, the evolutionary index based method has some significant advantages, such as the fact that it only requires expression profile of one species and makes full usage of prior knowledge. In particular, except for the gene age index, more evolutionary information can be retrieved from prior knowledge. For instance, the adaptive selections of genes can be traced by the nonsynonymous to synonymous substitution ratio (dN/dS) of sequences in specific phylogenetic clade, and genes with low dN/dS ratio are thought to be under selective pressure in certain species [62]. Besides, there are many expansions or contractions of gene families during the formation of each species, which lead to the copy number variations of homolog genes in different species [63]. Therefore, the states of gene duplication also imply diverse selective pressure on different genes for certain kind of species. Combining these two indices with the gene age index, Piasecka et al. have measured the transcriptomic conservation of embryonic development and evaluated the conservation of transcription regulation in zebra fish [64]. They completely reexamined conservation of development stages based on the expression profile of zebra fish embryogenesis, which is the same datasets adopted by Domazet-Lošo and Tautz [41]. Their result has showed the coexisting patterns of funnel model and hourglass model, as these evolutionary indices address different aspects of selective pressure and they are unable to make unanimous decision for either model. In addition, new method has been developed and tries to combine the evolutionary index with the gene expression for identification of conserved coexpression modules between species [65]. This method has been applied on the study carried out by Gerstein et al. [33], which has investigated the conservation of coexpression modules in development stages for worm and fly.

5. Discussion

As two major existing approaches of transcriptomic studies for EVO-DEVO, both of the correlation of gene expression method and the evolutionary indices based method show some advantages and defects (Table 1). Correlation of gene expression method can measure conservation inter-/intra-species/subjects while evolutionary indices based method combines age indices and evaluates conservation in a single subject. As the study of Piasecka et al. shows, these two approaches address different aspects of evolution so that combination of them would make a more comprehensive conclusion about the evolutionary model of embryonic development. We have summarized that the 3 major indices should be adopted to evaluate the model of development, for both hourglass and funnel model (Figure 2). These 3 measurements include gene expression correlation, dN/dS ratios, and transcriptome age index, which show different aspects of evolutionary selection. For instance, gene expression correlation stands for the similarity of paired transcriptome, dN/dS ratios show the selective pressure on gene sequences, and transcriptome age index combines the gene expression with phylogenetic age. These 3 measurements present significantly different patterns for each model. For instance, in hourglass model, the middle stages present the highest gene expression correlation and genes of these stages not only have conserved sequences but also are born in ancient time. In funnel-like model, these signatures present in the early stages of embryonic development (Figure 2).

Organism development is a cell expansion process which starts from single cell to multiple cells with different destinies. This procedure transforms from simple to complex in the view of the diversity of cell composition, which is in more concordance with von Baer's third law. However, nowadays more and more comparative transcriptomic researches support the hourglass model which proposes that the most conserved stages are in the middle period rather than the earlier period. The hourglass model is still not concluded as these comparative transcriptomic studies have technological limitation. For instance, except for the zygote, the rest of stages of embryo are composed of multiple cells, and the diversity of these embryonic cells increases along with the developmental time line. The RNA source for the comparative transcriptomic studies is extracted from embryo sample in multiple developmental time points and the RNA extractive is mixture of multiple cells. And along with development process, the RNA extractive includes more and more diverse RNAs from various cells. The different extent of RNA mixture at different development time points will affect the evolutionary conservation analysis results, as these comparative transcriptomic studies assume that every representative of

different development stages is considered to be single and equivalent. Based on the single cell RNA sequencing dataset of human preimplantation embryo [66], we have showed even in the early stages that there are up- and downfluctuation of selective pressure [67]. However, these single cell RNA sequencing datasets only cover the early stages of embryo for some species [42, 66, 68].

Besides the technological limitation, there are inherent problems in the experimental materials. For instance, many studies, which try to illustrate that the hourglass model universally exists in multiple species, have been conducted on the model organism such as mouse, worm, and fly. Compared with normal organisms, these model organisms share some common features such as short generation period and quick developmental time, which represent a specific mechanism of development and will potentially bias the result model of evolutionary studies [54]. Along with the decreasing of sequencing cost, more organisms, especially those that have long development procedure, should be profiled with multiple object sequencing. Depending on the single cell RNA sequencing technology, the whole embryonic development of more species would be profiled. Moreover, precise studies should be designed to illustrate this problem and construct sophisticated models about the evolution of development.

Conflict of Interests

The authors declare that they have no competing interests.

Authors' Contribution

Tiancheng Liu drafted the paper. Lin Yu provided the instruction of developmental biology. Hong Li revised the paper. Lei Liu and Yixue Li conceived the review. All authors read and approved the final paper. Tiancheng Liu and Lin Yu contributed equally.

Acknowledgments

This work was supported by National Basic Research Program of China (2011CB910204, 2011CB510102, and 2010CB529200), National Key Technology Support Program (2013BAI101B09), National Key Scientific Instrument and Equipment Development Project (2012YQ03026108), SIBS Knowledge Innovation Program (2014KIP215), and SA-SIBS Scholarship Program.

References

[1] M. Yeager, N. Orr, R. B. Hayes et al., "Genome-wide association study of prostate cancer identifies a second risk locus at 8q24," *Nature Genetics*, vol. 39, no. 5, pp. 645–649, 2007.

[2] A. Mahajan, M. J. Go, W. H. Zhang et al., "Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility," *Nature Genetics*, vol. 46, pp. 234–244, 2014.

[3] H. E. Speedy, M. C. Di Bernardo, G. P. Sava et al., "A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia," *Nature Genetics*, vol. 46, no. 1, pp. 56–60, 2014.

[4] D. Altshuler, M. J. Daly, and E. S. Lander, "Genetic mapping in human disease," *Science*, vol. 322, no. 5903, pp. 881–888, 2008.

[5] Z.-Q. Ye, S. Niu, Y. Yu et al., "Analyses of copy number variation of GK rat reveal new putative type 2 diabetes susceptibility loci," *PLoS ONE*, vol. 5, no. 11, Article ID e14077, 2010.

[6] T. Miyagawa, M. Kawashima, N. Nishida et al., "Variant between CPT1B and CHKB associated with susceptibility to narcolepsy," *Nature Genetics*, vol. 40, no. 11, pp. 1324–1328, 2008.

[7] R.-L. Ge, Q. L. Cai, Y.-Y. Shen et al., "Draft genome sequence of the Tibetan antelope," *Nature Communications*, vol. 4, article 1858, 2013.

[8] R. Q. Li, W. Fan, G. Tian et al., "The sequence and de novo assembly of the giant panda genome," *Nature*, vol. 463, pp. 311–317, 2010.

[9] A. Fujimoto, H. Nakagawa, N. Hosono et al., "Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing," *Nature Genetics*, vol. 42, no. 11, pp. 931–936, 2010.

[10] R. Bonasio, G. Zhang, C. Ye et al., "Genomic comparison of the ants camponotus floridanus and harpegnathos saltator," *Science*, vol. 329, no. 5995, pp. 1068–1071, 2010.

[11] S. S. Atanur, I. Birol, V. Guryev et al., "The genome sequence of the spontaneously hypertensive rat: analysis and functional significance," *Genome Research*, vol. 20, no. 6, pp. 791–803, 2010.

[12] W. Z. Jirimutu, G. Ding, G. Chen et al., "Genome sequences of wild and domestic bactrian camels," *Nature Communications*, vol. 3, article 1202, 2012.

[13] K. Saar, A. Beck, M.-T. Bihoreau et al., "SNP and haplotype mapping for genetic analysis in the rat," *Nature Genetics*, vol. 40, no. 5, pp. 560–566, 2008.

[14] S. S. Atanur, A. G. Diaz, K. Maratou et al., "Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat," *Cell*, vol. 154, no. 3, pp. 691–703, 2013.

[15] C.-J. Rubin, M. C. Zody, J. Eriksson et al., "Whole-genome resequencing reveals loci under selection during chicken domestication," *Nature*, vol. 464, no. 7288, pp. 587–591, 2010.

[16] N. Navin, J. Kendall, J. Troge et al., "Tumour evolution inferred by single-cell sequencing," *Nature*, vol. 472, no. 7341, pp. 90–95, 2011.

[17] J. Alföldi and K. Lindblad-Toh, "Comparative genomics as a tool to understand evolution and disease," *Genome Research*, vol. 23, no. 7, pp. 1063–1068, 2013.

[18] G. Zhang, C. Cowled, Z. Shi et al., "Comparative analysis of bat genomes provides insight into the evolution of flight and immunity," *Science*, vol. 339, no. 6118, pp. 456–460, 2013.

[19] E. B. Kim, X. Fang, A. A. Fushan et al., "Genome sequencing reveals insights into physiology and longevity of the naked mole rat," *Nature*, vol. 479, no. 7372, pp. 223–227, 2011.

[20] Y. Fan, Z. Y. Huang, C. C. Cao et al., "Genome of the Chinese tree shrew," *Nature Communications*, vol. 4, article 1426, 2013.

[21] Z. Wang, J. Pascual-Anaya, A. Zadissa et al., "The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan," *Nature Genetics*, vol. 45, pp. 701–706, 2013.

[22] A. de la Fuente, "From 'differential expression' to 'differential networking'—identification of dysfunctional regulatory networks in diseases," *Trends in Genetics*, vol. 26, no. 7, pp. 326–333, 2010.

[23] J. Zhang, R. Chiodini, A. Badr, and G. F. Zhang, "The impact of next-generation sequencing on genomics," *Journal of Genetics and Genomics*, vol. 38, no. 3, pp. 95–109, 2011.

[24] D. Altshuler, R. M. Durbin, and G. R. Abecasis, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, 2010.

[25] E. Birney, J. A. Stamatoyannopoulos, A. Dutta et al., "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," *Nature*, vol. 447, pp. 799–816, 2007.

[26] I. Dunham, A. Kundaje, S. F. Aldred et al., "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[27] M. B. Gerstein, Z. J. Lu, E. L. van Nostrand et al., "Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project," *Science*, vol. 330, pp. 1775–1787, 2010.

[28] S. Roy, J. Ernst, P. V. Kharchenko et al., "Identification of functional elements and regulatory circuits by *Drosophila* modENCODE," *Science*, vol. 330, no. 6012, pp. 1787–1797, 2010.

[29] J. Ernst, P. Kheradpour, T. S. Mikkelsen et al., "Mapping and analysis of chromatin state dynamics in nine human cell types," *Nature*, vol. 473, no. 7345, pp. 43–49, 2011.

[30] K. Y. Yip, C. Cheng, N. Bhardwaj et al., "Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors," *Genome Biology*, vol. 13, no. 9, article R48, 2012.

[31] C. Cheng, K.-K. Yan, K. Y. Yip et al., "A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets," *Genome Biology*, vol. 12, no. 2, article R15, 2011.

[32] X. J. Dong, M. C. Greven, A. Kundaje et al., "Modeling gene expression using chromatin features in various cellular contexts," *Genome Biology*, vol. 13, no. 9, article R53, 2012.

[33] M. B. Gerstein, J. Rozowsky, K. K. Yan et al., "Comparative analysis of the transcriptome across distant species," *Nature*, vol. 512, no. 7515, pp. 445–448, 2014.

[34] U. D. Akavia, O. Litvin, J. Kim et al., "An integrated approach to uncover drivers of cancer," *Cell*, vol. 143, no. 6, pp. 1005–1017, 2010.

[35] 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.

[36] The Cancer Genome Atlas Research Network, "Comprehensive molecular characterization of gastric adenocarcinoma," *Nature*, vol. 513, pp. 202–209, 2014.

[37] N. Agrawal, R. Akbani, and B. A. Aksoy, "Integrated genomic characterization of papillary thyroid carcinoma," *Cell*, vol. 159, pp. 676–690, 2014.

[38] K. E. von Baer, *Über Entwickelungsgeschichte der Thiere. Beobachtung und Reflexion*, Gebrüder Bornträger, Königsberg, Germany, 1828.

[39] D. Duboule, "Temporal colinearity and the phylotypic progression—a basis for the stability of a vertebrate bauplan and the evolution of morphologies through heterochrony," *Development*, pp. 135–142, 1994.

[40] R. Vassena, S. Boué, E. González-Roca et al., "Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development," *Development*, vol. 138, no. 17, pp. 3699–3709, 2011.

[41] T. Domazet-Lošo and D. Tautz, "A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns," *Nature*, vol. 468, no. 7325, pp. 815–819, 2010.

[42] Z. Xue, K. Huang, C. Cai et al., "Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing," *Nature*, vol. 500, no. 7464, pp. 593–597, 2013.

[43] D. Duboule, "Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony," *Development Supplement*, pp. 135–142, 1994.

[44] A. C. Burke, C. E. Nelson, B. A. Morgan, and C. Tabin, "Hox genes and the evolution of vertebrate axial morphology," *Development*, vol. 121, no. 2, pp. 333–346, 1995.

[45] X. J. Tian, J. E. Strassmann, and D. C. Queller, "Dictyostelium development shows a novel pattern of evolutionary conservation," *Molecular Biology and Evolution*, vol. 30, no. 4, pp. 977–984, 2013.

[46] N. Irie and S. Kuratani, "Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis," *Nature Communications*, vol. 2, no. 1, article 248, 2011.

[47] R. Riedl, *Order in Living Organisms*, Wiley-Interscience, West Sussex, UK, 1978.

[48] M. K. Richardson, J. Hanken, M. L. Gooneratne et al., "There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development," *Anatomy and Embryology*, vol. 196, no. 2, pp. 91–106, 1997.

[49] S. Poe and M. H. Wake, "Quantitative tests of general models for the evolution of development," *American Naturalist*, vol. 164, no. 3, pp. 415–422, 2004.

[50] O. R. P. Bininda-Emonds, J. E. Jeffery, and M. K. Richardson, "Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development," *Proceedings of the Royal Society B: Biological Sciences*, vol. 270, no. 1513, pp. 341–346, 2003.

[51] M. K. Richardson, S. P. Allen, G. M. Wright, A. Raynaud, and J. Hanken, "Somite number and vertebrate evolution," *Development*, vol. 125, no. 2, pp. 151–160, 1998.

[52] S. Poe, "Test of Von Baer's law of the conservation of early development," *Evolution*, vol. 60, no. 11, pp. 2239–2245, 2006.

[53] M. K. Richardson, A. Minelli, M. Coates, and J. Hanken, "Phylotypic stage theory," *Trends in Ecology & Evolution*, vol. 13, no. 4, p. 158, 1998.

[54] B. K. Hall, "Phylotypic stage or phantom: is there a highly conserved embryonic stage in vertebrates?" *Trends in Ecology & Evolution*, vol. 12, no. 12, pp. 461–463, 1997.

[55] T. A. Williams, "The nauplius larva of crustaceans—functional diversity and the phylotypic stage," *American Zoologist*, vol. 34, pp. 562–569, 1994.

[56] A. T. Kalinka, K. M. Varga, D. T. Gerrard et al., "Gene expression divergence recapitulates the developmental hourglass model," *Nature*, vol. 468, no. 7325, pp. 811–814, 2010.

[57] M. Ninova, M. Ronshaugen, and S. Griffiths-Jones, "Conserved temporal patterns of microrna expression in drosophila support a developmental hourglass model," *Genome Biology and Evolution*, vol. 6, no. 9, pp. 2459–2467, 2014.

[58] M. Long, E. Betrán, K. Thornton, and W. Wang, "The origin of new genes: glimpses from the young and old," *Nature Reviews Genetics*, vol. 4, no. 11, pp. 865–875, 2003.

[59] T. Domazet-Lošo, J. Brajković, and D. Tautz, "A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages," *Trends in Genetics*, vol. 23, no. 11, pp. 533–539, 2007.

[60] T. Domazet-Lošo and D. Tautz, "Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa," *BMC Biology*, vol. 8, article 66, 2010.

[61] M. Quint, H.-G. Drost, A. Gabel, K. K. Ullrich, M. Bönn, and I. Grosse, "A transcriptomic hourglass in plant embryogenesis," *Nature*, vol. 490, no. 7418, pp. 98–101, 2012.

[62] K. Nei, *Molecular Evolution and Phylogenetics*, Oxford University Press, New York, NY, USA, 2000.

[63] M. Lynch and J. S. Conery, "The evolutionary fate and consequences of duplicate genes," *Science*, vol. 290, no. 5494, pp. 1151–1155, 2000.

[64] B. Piasecka, P. Lichocki, S. Moretti, S. Bergmann, and M. Robinson-Rechavi, "The hourglass and the early conservation models—co-existing patterns of developmental constraints in vertebrates," *PLoS Genetics*, vol. 9, no. 4, Article ID e1003476, 2013.

[65] K. K. Yan, D. Wang, J. Rozowsky, H. Zheng, C. Cheng, and M. Gerstein, "OrthoClust: an orthology-based network framework for clustering data across multiple species," *Genome Biology*, vol. 15, no. 8, article R100, 2014.

[66] L. Yan, M. Yang, H. Guo et al., "Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells," *Nature Structural & Molecular Biology*, vol. 20, no. 9, pp. 1131–1139, 2013.

[67] T. Liu, L. Yu, G. Ding et al., "Gene coexpression and evolutionary conservation analysis of the human preimplantation embryos," *BioMed Research International*, In press.

[68] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg, "Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells," *Science*, vol. 343, no. 6167, pp. 193–196, 2014.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 362806, 8 pages
http://dx.doi.org/10.1155/2015/362806

Research Article

ROC-Boosting: A Feature Selection Method for Health Identification Using Tongue Image

Yan Cui,[1,2] Shizhong Liao,[1] and Hongwu Wang[3]

[1]*School of Computer Science and Technology, Tianjin University, 72 Weijin Road, Nankai District, Tianjin 300072, China*
[2]*Department of Common Required Courses, Tianjin University of Traditional Chinese Medicine, 312 Anshanxi Road, Nankai District, Tianjin 300193, China*
[3]*College of Traditional Chinese Medicine, Tianjin University of Traditional Chinese Medicine, 312 Anshanxi Road, Nankai District, Tianjin 300193, China*

Correspondence should be addressed to Shizhong Liao; szliao@tju.edu.cn and Hongwu Wang; tjwanghw55@163.com

Received 28 May 2015; Revised 5 August 2015; Accepted 5 August 2015

Academic Editor: Tao Huang

Objective. To select significant Haar-like features extracted from tongue images for health identification. *Materials and Methods*. 1,322 tongue cases were included in this study. Health information and tongue images of each case were collected. Cases were classified into the following groups: group containing 148 cases diagnosed as health; group containing 332 cases diagnosed as ill based on health information, even though tongue image is normal; and group containing 842 cases diagnosed as ill. Haar-like features were extracted from tongue images. Then, we proposed a new boosting method in the ROC space for selecting significant features from the features extracted from these images. *Results*. A total of 27 features were obtained from groups A, B, and C. Seven features were selected from groups A and B, while 25 features were selected from groups A and C. *Conclusions*. The selected features in this study were mainly obtained from the root, top, and side areas of the tongue. This is consistent with the tongue partitions employed in traditional Chinese medicine. These results provide scientific evidence to TCM tongue diagnosis for health identification.

1. Introduction

As society continues to develop, health status problems have become the main focus of studies in recent years, and health identification has been one of the most important problems. Health identification is a procedure of identifying the condition of a subject as healthy or ill. Western medicine diagnoses a person's health condition based on a series of laboratory examinations. However, these examinations are invasive and time-consuming and require a number of laboratory experiments. As an alternative diagnostic method, traditional Chinese medicine (TCM) proposes Su Wen (*Plain Questions*) as a concept for the preventive treatment of diseases. *Plain Questions* is part of a classical text written during the Zhanguo period of ancient China, which claims that TCM identifies the health status of a person before diagnosing the decease. Health identification is one of the most fundamental diagnostic methods applied in the preventive treatment of diseases in TCM [1]. Compared with Western medicine, TCM uses noninvasive, time-saving methods including tongue and pulse to identify the health status of an individual. In recent years, Western medicine has also begun to focus on establishing preventive treatments for diseases such as health identification, because these results can save medical time, effort, and cost [2].

However, tongue diagnosis in TCM has been criticized due to its subjective diagnostic criteria. Several studies have focused on tongue image diagnosis, and computer image processing has contributed to tongue criteria objectification. Color is the most common feature in tongue diagnosis due to its intuitiveness. The study of Pang et al. introduced lower order moments such as the mean value and standard deviation of color features to diagnose appendicitis [3]. The study of Zhao et al. found color differences between patients with and without chronic hepatitis B [4]. In the current study, tongue coating color features were extracted. For color

features, Wang et al. divided tongue colors into 12 categories through the statistical analysis of a large number of tongue images [5]. However, these color features are global and could not describe the local information of the images. The study of Kanawong et al. used color features to classify TCM ZHENG [6]. In the current study, tongue images were divided into several areas, and features in several colored spaces were extracted. The study of Jung et al. preformed a case-control study to investigate color distribution differences in tongue for sleep disorders [7]. The current study also used partition color features. Zhang et al. used their TCM partition knowledge in their AdaBoost algorithm for tongue recognition [8]. The common feature of the above-mentioned studies is that TCM knowledge is used in partitioning. However, from the perspective of TCM, results of these studies were not accepted, because these were based on TCM a priori knowledge. Moreover, these studies focused on disease diagnosis rather than health identification, which is not consistent with TCM diagnosis, in which TCM claims that health identification is more important than disease diagnosis [9].

Zhi et al. used hyperspectral features to classify tongue images [10], and this feature is also a global feature. Yang et al. used texture and curvature features to detect tongue cracks [11, 12]. The current study focused on special tongue images other than tongue diagnosis, and these features are also too global to represent any relationship among these tongue partitions.

Haar-like features are a class of image partition features. Viola and Jones used Haar-like features in face detection [13]. Face detection classifies partitions that contain and do not contain human faces. Currently, this approach has been proven to be successful for face detection. However, its performance significantly declines when applied in other areas [14]. Fu et al. used Haar-like features to segment tongue images from the background [15]. Wang et al. also used Haar-like features to detect and track the lips [16]. However, these two studies did not focus on diagnosing the disease or identifying the health status. Cui et al. used this approach in diagnosing hyperuricemia [17]. However, only a few similar studies can be found.

Feature selection is one of the most important steps in classification and diagnosis. There are three types of feature selection methods: the filter method, which selects an optimal feature set before classification; the wrapper method, which uses a fixed search strategy and interacts with the classifier; and the embedded method, which combines feature selection and the classifier together. Saeys et al. evaluated these methods by applying these in bioinformatics [18].

Boosting algorithm is an effective approach in feather selection and classification. It combines the results of many single classifiers, and the performance of combining these results is better than a single classifier. The reason why this boosting method has a better performance can be explained by the probably approximately correct theory of Valiant [19]. In this theory, the concepts of strong learnable and weak learnable were defined. Schapire was able to prove that these two concepts are equivalent [20]. This conclusion indicates that if a classification model with high predictive accuracy exists, an ensemble of a series of weak models is equivalent to it, even if their predictive results are only slightly better than a random guess.

Viola and Jones used a boosting algorithm in face detection [13] and demonstrated that this algorithm can be employed to cope with both feature selection and classification. However, this algorithm is only suitable for face detection, because the eyes and nose are naturally identifiable. Mamitsuka proposed a boosting algorithm based on the ROC curve for microarray classification [21]. Komori and Eguchi and Long and Servedio also proposed boosting algorithms for maximizing the area under the ROC (AUC) [22, 23]. These studies were able to partly solve the small observation problem but were not used in unbalanced sample problems. In our problem, the number of features is much larger than the number of examples. Fan and Lv proposed a theoretical guarantee for screening features from ultrahigh feature spaces [24].

In order to address this limitation, we propose a ROC-Boosting approach for TCM tongue diagnosis in health identification. This method first screens the features using t-test. Then, a Haar-like feature is selected using several different conditions and sends this feature to the ensemble classifier. Our method is generic compared to previous methods, because its conditions include the AUC value, sensitivity, specificity, and their combinations. It can also use positive-negative sample ratio conditions to deal with unbalanced sample problems. We name this method ROC-Boosting, because its feature selection conditions are all relative to the ROC space. Moreover, our method avoids the usage of TCM a priori, and its result is consistent with TCM tongue partitions.

2. Subjects and Methods

Tongue images and health information of 1,426 cases from 2011 to 2012 were collected. ROC-Boosting was employed to select Haar-like features extracted from tongue images. Natural partition tongue image features are selected. Then, partitions on the tongue image confirm the TCM diagnosis method. This procedure is illustrated in Figure 1.

2.1. Subjects. Tongue images and health information of 1,426 cases from 2011 to 2012 were obtained from the Teaching Hospital of Tianjin University of Traditional Chinese Medicine (TJUTCM). Among these 1,426 cases, 96 cases were excluded due to low quality or duplicate images and health records with missing values. Then, TCM students and experts were employed to discuss the tongue images and health information collected. During this discussion, eight additional duplicate images were found. An outpatient doctor confirmed that these duplicates resulted from the abuse of health insurance ID usage. Hence, these eight images were excluded. Finally, a total of 1,322 cases were included into this study. TCM diagnoses health/illness status before the specific disease, because TCM focuses on the preventive treatment of diseases [8]. For this reason, we focused on the health identification problem in this study, and all subjects were diagnosed as healthy or ill. The 1,322 cases were classified in the following groups: group A, diagnosed as healthy based on

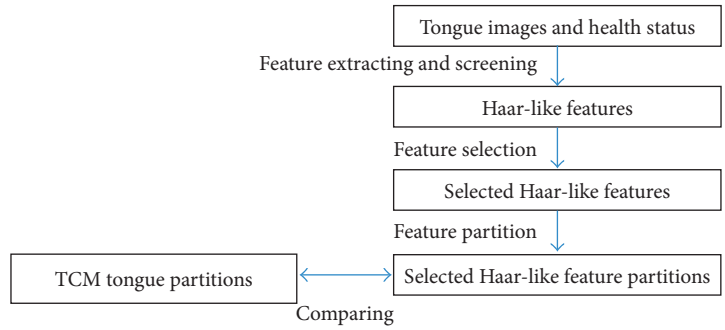


Figure 1: Study procedure. Haar-like features are extracted from tongue images and screened. Then, partitions on the tongue image confirm the TCM diagnosis method.

Table 1: Summary of the number of cases in each group.

Group	Status	Number of cases
A	Healthy, diagnosed based on tongue image and heath information	148
B	Ill, diagnosed based on health information, but tongue image is normal	332
C	Ill, diagnosed based on both tongue image and health information	842
Total		1,322

both tongue image and health information ($n = 148$ cases); group B, diagnosed as ill based on health information, even if tongue images are normal (TCM considers that tongue changes do not reflect all illnesses, $n = 332$ cases); and group C, diagnosed as ill based on both tongue image and health information ($n = 842$ cases). The number of cases in each group is summarized in Table 1. Before the features were extracted, all images were scaled to 120×100 and segmented from the background to exclude the impact of the background to the feature extraction and selection process, as shown in Figure 2.

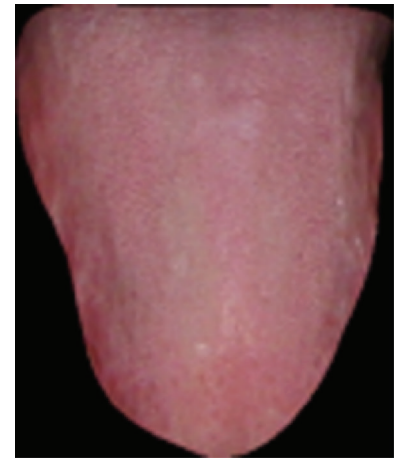


Figure 2: A segmented tongue image. A tongue image segmented from the background to exclude the impact of the background to the feature extraction and selection process.

2.2. Methods

2.2.1. Improved Haar-Like Feature Extraction. Usually, color features are extracted from the whole tongue and Haar-like features are extracted from partitions. We improved the Haar-like feature to have five partitions, considering that humans focus their view at the center of the target at first glance, as shown in Figure 3. In comparison, the original Haar-like feature was considered as the difference of the sum of the color values between two horizontal or vertical partitions. The center partition of this feature has two parameters: W and H. These parameters represent the width and height of this partition. The other four surrounding partitions have three parameters: T, X, and Y. T represents the width of these four partitions, while X and Y represent the position of the top-left corner of the Haar-feature. Considering that humans usually focus their view at the center partition and the other four partitions, the number of pixels at the center partition and the other four partitions in the improved Haar-like feature should be equal. To ensure that the number of pixels at the center partition is equal to the other four partitions, T is given by $[WH/(2W+2H)]$, where $[\,]$ represents the maximum integer number smaller than the calculated real number. Parameters X and Y represent the position of the left-top corner of this feature. The improved Haar-like feature uses the difference between the sum of the pixel color values at the center partition and the other four partitions.

Every improved Haar-like feature is composed of five partitions (1–5). This Haar-like feature has five parameters: X, Y, W, H, and T. W and H represent the width and height of the center partition. T represents the width of the other partitions. X and Y represent the position of the Haar-like feature. The feature value can be computed by the difference between the sum of the pixel values at center partition and the other four partitions.

Under this setup, the number of improved Haar-like features is very large. Considering that the number of significant features is very small in our previous study, we reduced the number of features similar to our pervious study [14].

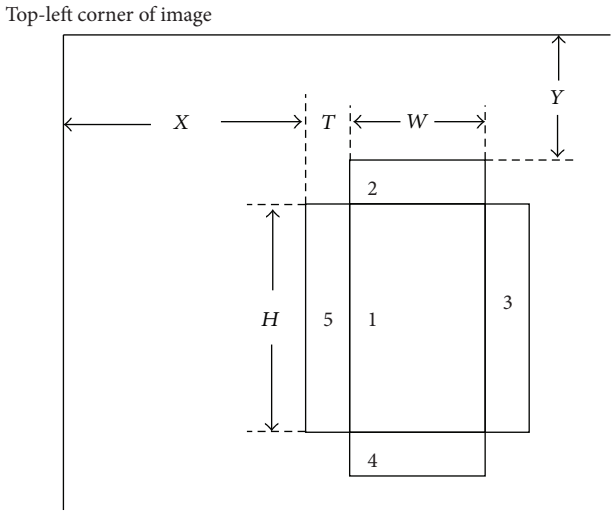


Figure 3: The improved Haar-like feature.

In this study, the density of the parameter grid is reduced to lower the number of improved Haar-like features. We set $W \in \{10, 12, 14, \ldots, 60\}$, $H \in \{10, 12, 14, \ldots, 72\}$, $X \in \{1, 6, 11, \ldots, \lfloor 100 - W - 2 * T + 1 \rfloor\}$, and $Y \in \{1, 7, 13, \ldots, \lfloor 120 - H - 2 * T + 1 \rfloor\}$ experimentally. After this simplification, the number of features is 98,592 in red, green, and blue color plains, respectively. These features are parts of the inputs of the ROC-Boosting algorithm.

2.2.2. ROC-Boosting. Concerning the difference between improved Haar-like feature values of the healthy and ill population, three tests were designed. The first test investigates the difference between the healthy group (group A) and the ill group diagnosed solely based on health information (group B). Cases in group B were diagnosed as healthy, because differences in tongue images cannot be observed by using the human eye. This test would prove whether a difference exists between these two groups. The second test is designed to verify the difference between the healthy and ill groups (groups A and C), because the difference between these two groups can be observed by using the human eye. The third test is designed to verify the difference between the healthy (A) and ill groups (B and C). As the number of improved Haar-like features becomes very large in comparison to the number of examples, t-tests were used to screen the improved Haar-like features in the first instance before applying our method. We reduced the number of improved Haar-like features to approximately 10^4 through the P value of the t-test. The P value and number of filtered features are listed in Table 2. These features are inputs of our method.

Our method, ROC-Boosting, is illustrated in Algorithm 1. This algorithm calculates the AUC value of all features in every loop. The AUC value would be set to its negative value when the ROC curve is concave; that is, the ROC curve is flipped around the random guess line. This flipping is designed to deal with the reversed prediction feature. Then, ROC-Boosting selects the feature through some conditions, which would be discussed later. After the correctly classified examples and selected features are removed, the loop is restarted. When conditions for selecting features no longer meet, the algorithm stops and presents all selected features. ROC-Boosting selects the most significant features on the tongue images to diagnose subjects from different groups in each step. These features can be used to build classifiers for identifying the health status of subjects from different groups. Verification of conformance between the positions of these features and TCM tongue partitions provides scientific evidence for TCM tongue diagnosis.

Table 2: P values and the number of features between groups.

Test	Group	P value	Number of filtered features
1	A/B	0.05	14,878
2	A/C	0.00005	12,280
3	A/B + C	0.0005	11,260

As described before, this algorithm is a generic version of the algorithm used by Yang et al. [12]. Viola's method only applies to situations when features with extremely high sensitivity exist such as features that describe the eyes and nose of a human face detection problem. When the condition in step 9 is changed to the highest sensitivity and specificity, ROC-Boosting would be similar to Viola's algorithm. The procedure for running Viola's algorithm has shown that such features do not exist in our problem. This is the reason why we generalized Viola's method.

In our problem, we use the next two conditions in step 9. The first is a negative/positive ratio condition. We compute $r = p/n$ and $r' = p'/n'$, where p is the number of positive examples, n is the number of negative examples, p' is the number of positive examples correctly classified by one feature, and n' is the number of negative examples corrected classified by one feature. This condition is $|r - r'|$. The second condition is the AUC $|a - 0.5|$ value, where a is the AUC of this feature. We used these two conditions, because we did not find any significant feature existing in our problem, and the positive/negative examples are not balanced.

2.3. Statistical Analysis Software. We extracted the ROC-Boosting features using a DELL PC (OptiPlex 7020, i5-4590; Quad-Core with 8 GB RAM). The R 2.15.2 64 bit version was the statistical software used [25]. AUC values were calculated using the ROCR package. The code for feature extraction and ROC-Boosting was programmed as a script in R.

3. Results and Discussions

3.1. Results. For Test 1, only eight features are selected. The ninth feature condition is $|a - 0.5| = 0.00943$. For Test 2, 25 features are selected. The 26th feature condition is $|a - 0.5| = 0.00687$. In these two tests, the algorithm comes to its end, because a feature that is better than the guess could no longer be found. For Test 3, 27 features are selected. The 28th feature exceedingly concerns the disease examples than the healthy

```
Input:   Example set with improved Haar-like feature filtered by t-test
Output: Feature set selected
(1)     Do
(2)       For each feature that is not selected
(3)         Compute area under ROC curve for this feature
(4)         If the feature is concave
(5)           Use negative feature value
(6)         End If
(7)         Exclude features whose ROC curve cross random guess line in ROC space
(8)       Next
(9)       Select one feature using some conditions in ROC space
(10)      Exclude examples correctly classified by this feature
(11)      If these conditions are not fulfilled
(12)        Exit Do
(13)      End if
(14)    Loop
```

Algorithm 1: ROC-Boosting algorithm.

examples (35/2). The algorithm comes to an end, because features concerning the positive and negative examples could no longer be found.

We overlaid the selected features and tongue image to investigate the relationship between these two, because we assume that tongue diagnosis is conducted by repetitive observations and each observation corresponds to one feature. Figure 4 shows the results of the superposition of all selected features. From (a) to (d) in Figure 4: (a) is our sample tongue image; (b) is the superposition of features for Test 1, and its AUC value is 0.662; (c) is for Test 2, and its AUC value is 0.740; and (d) is for Test 3, and its AUC value is 0.723.

Test 1 has a poor performance due to insignificant differences between groups A and B. Even the human eye cannot identify the health status of group B. Test 2 had the best result, because group A is composed of healthy subjects and group C is composed of ill subjects. The shape of the overlaid feature is distributed around the tongue. The difference between these two groups is the most significant among the three tests. The overlaid features for Test 2 are concentrated at the center of the tongue. The overlaid features in the last figure consist of three areas: root, center, and top of the tongue image. We marked these three areas in the last figure. Its performance is slightly worse than Test 2 due to the interference of group B.

3.2. Discussions. Our method is more generic than previous studies. Viola's method is only applicable to situations where high sensitivity features exist [13]. In this situation, the algorithm selects features with increasing specificity from high sensitivity features. A high performance classifier would be built when high sensitivity is maintained and specificity is increased. ROC-Boosting can also use specificity and sensitivity simultaneously, and the value of specificity and sensitivity can be relatively low. It can also use other conditions to select features. ROC-Boosting also works well even when high sensitivity features do not exist. One of the problems of health identification is the use of Haar-like features on tongue images. In a preexperiment, we tested all features in this study. No high sensitivity feature exists in our data, and we confirmed that Viola's method works. ROC-Boosting was able to select the features.

Learnability theory guarantees the effectiveness of the ROC-Boosting algorithm. If a high performance classifier exists in a health identification problem using Haar-like features on the tongue image, which is the basic hypothesis of this study, an ensemble of weak classifiers whose performance is better than the random guess would be equivalent to it. In ROC-Boosting, the condition $|a - 0.5| > 0$ keeps every weak classifier better than the random guess.

Furthermore, weak classifiers should focus on both negative and positive examples. As shown in Figure 5, even though these two features (181,520 in the left subfigure and 188,479 in the right subfigure) were $|a - 0.5| > 0$, 188,479 is excluded, because it focuses on the positive example only. In ROC-Boosting, $|r - r'|$ excludes these features.

We also compared the two different conditions of ROC-Boosting for Test 3. The first condition is the negative/positive ratio and the $|d - d'|$ condition, which was used in our study. The second condition is the solely AUC value condition, which was proposed in previous studies. Comparative results are shown in Figure 6. We also selected 27 features using the solely AUC value condition. However, after the 16th feature was selected, the next features selected by this condition could only correctly predict one to four subjects at two ends of the feature value. The number of correctly predicted subjects determines that the AUC value increases in this step. This indicates that the AUC value is unsustainable after the 16th feature is selected. To compare the results of two different conditions, we run the program continuously. Although the AUC value of the right image (0.727) is slightly larger than the left image (0.723), the solely AUC condition obtains this superiority, because it is inclined to predict all subjects as with disease. When the number of positive and negative subjects is unbalanced, prediction tending to the major class is the most common phenomenon.

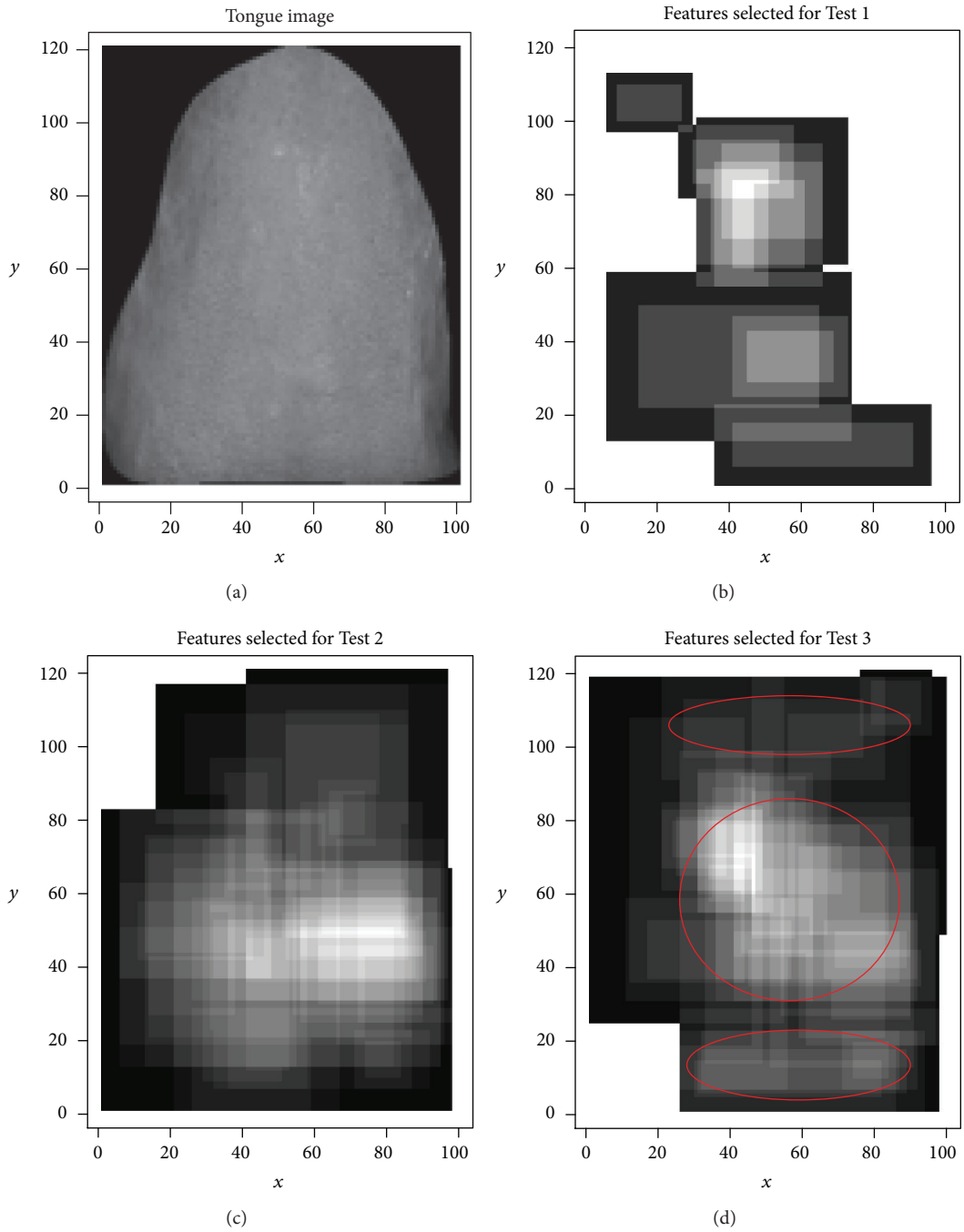


Figure 4: Superposition of all selected features is shown. From (a) to (d):(a) is in grayscale and is used as the background, (b) is the result of Test 1, (c) is the result of Test 2, and (d) is the result of Test 3. In (b, c, d), every rectangle represents one feature. The lighter partition in the rectangle is the center of the feature, and darker partitions surround this feature. Squares at the four corners of the feature were not removed for simplification. When repeatedly overlaid while keeping the color of the darker partitions fixed, the lighter partition would continue to be lighter. Finally, the lightest partition is observed intensively.

Relatively, the negative/positive ratio and $|d - d'|$ condition are associated with both positive and negative classes, while obtaining a similar AUC value. The correct prediction of both healthy and ill subjects is equally important in health identification.

4. Conclusion

We propose the application of the ROC-Boosting algorithm for health identification. This algorithm uses filtered Haar-like features and selects features from both positive and negative examples. The features selected for diagnosing

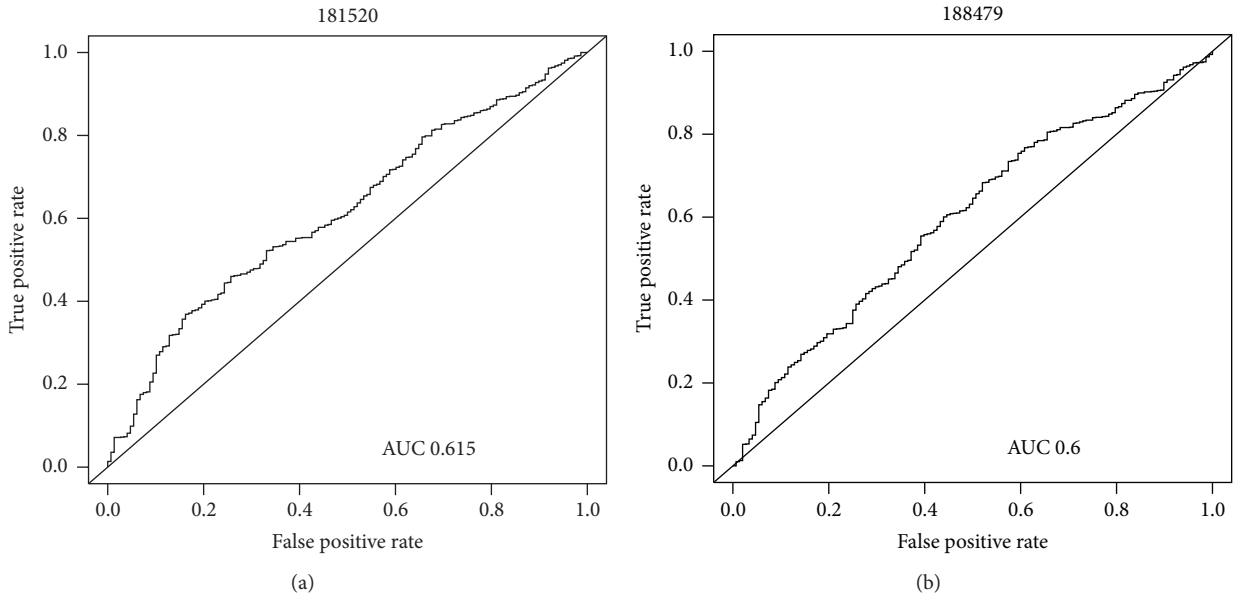


Figure 5: Two features of the AUC value larger than 0.5 are shown. Feature 181,520 in the left is selected, because the whole ROC curve is laid upward the random guess line. Feature 188,479 in the right is excluded, because its ROC curve crosses the random guess line at the corners of the ROC space.

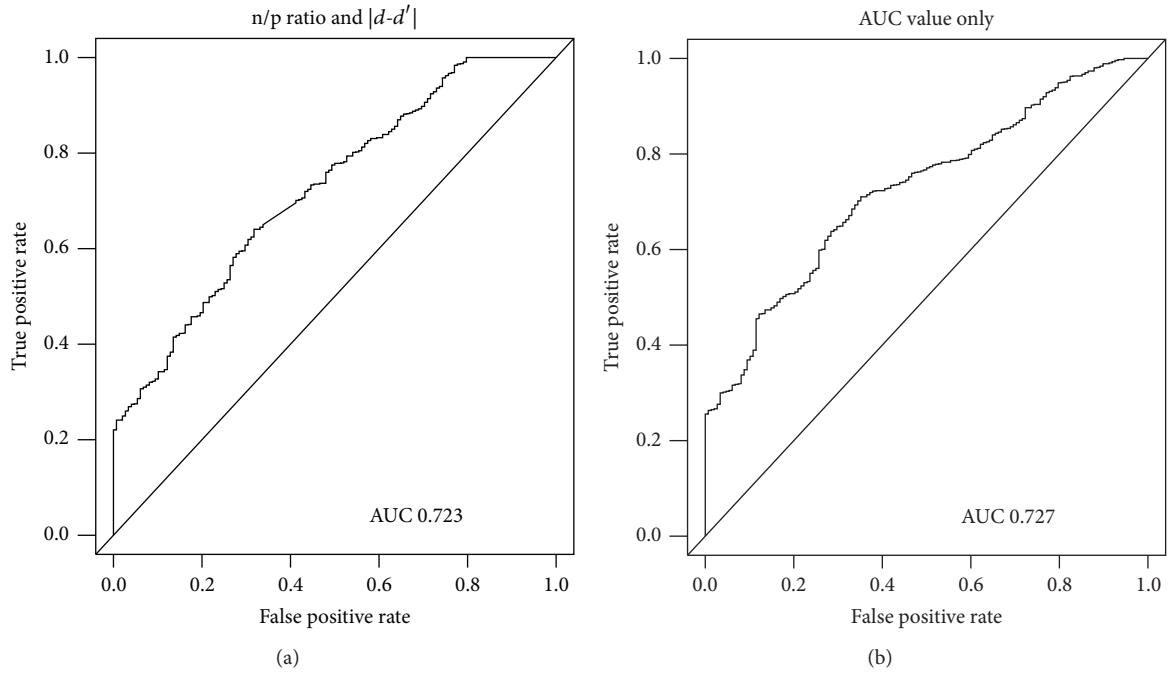


Figure 6: Comparative results of the negative/positive ratio and $|d - d'|$ conditions, as well as the solely AUC value condition. Although the AUC value on the right image (0.727) is slightly larger than the left image (0.723), the solely AUC value condition obtains this superiority, because it is inclined to predict all subjects as with disease. When the number of positive and negative subjects is unbalanced, prediction tending to the major class is the most common phenomenon.

health and ill subjects are concentrated in the root, center, and top partitions of the tongue images. Unlike previous studies, these partitions are not results of preexperience. A deterministic algorithm presents these partitions. These results provide scientific evidence to TCM tongue diagnosis for health identification.

Conflict of Interests

The authors declare no conflict of interests in this work.

Authors' Contribution

Yan Cui completed the algorithm and wrote the paper. Hongwu Wang and Shizhong Liao performed the mathematical models and methods of this study.

Acknowledgments

This work was supported by the National Basic Research Program of China (973 Program, Grant no. 2011CB505406) and the Tianjin Administration of Traditional Chinese Medicine, Chinese medicine, Integrative Medicine Research and Special Program (Grant no. 15269).

References

[1] L. D. Jiang, "Discussion on concepts of the health, subhealth, before sickness and prevention," *Chinese Journal of Traditional Chinese Medicine*, no. 25, pp. 167–170, 2010 (Chinese).

[2] J. P. Koplan, T. C. Bond, M. H. Merson et al., "Towards a common definition of global health," *The Lancet*, vol. 373, no. 9679, pp. 1993–1995, 2009.

[3] B. Pang, D. Zhang, and K. Wang, "Tongue image analysis for appendicitis diagnosis," *Information Sciences*, vol. 175, no. 3, pp. 160–176, 2005.

[4] Y. Zhao, X.-J. Gou, J.-Y. Dai et al., "Differences in metabolites of different tongue coatings in patients with chronic hepatitis B," *Evidence-Based Complementary and Alternative Medicine*, vol. 2013, Article ID 204908, 12 pages, 2013.

[5] X. Wang, B. Zhang, Z. Yang, H. Wang, and D. Zhang, "Statistical analysis of tongue images for feature extraction and diagnostics," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 5336–5347, 2013.

[6] R. Kanawong, T. Obafemi-Ajayi, T. Ma, D. Xu, S. Li, and Y. Duan, "Automated tongue feature extraction for ZHENG classification in Traditional Chinese Medicine," *Evidence-Based Complementary and Alternative Medicine*, vol. 2012, Article ID 912852, 14 pages, 2012.

[7] C. J. Jung, J. H. Nam, Y. J. Jeon, and K. H. Kim, "Color distribution differences in the tongue in sleep disorder," *Evidence-Based Complementary and Alternative Medicine*, vol. 2014, Article ID 323645, 8 pages, 2014.

[8] M. Zhang, X. Hu, Y. Wang et al., "Research of adaboost algorithm in tongue image of traditional Chinese medicine partitions recognition," *Journal of Chinese Computer Systems*, no. 6, pp. 1149–1153, 2008 (Chinese).

[9] L. Hong and H. Xi, "Theoretical study on preventive treatment theory in traditional Chinese medicine," *Chinese Journal of Basic Medicine in Traditional Chinese Medicine*, no. 2, pp. 92–94, 2007 (Chinese).

[10] L. Zhi, D. Zhang, J.-Q. Yan, Q.-L. Li, and Q.-L. Tang, "Classification of hyperspectral medical tongue images for tongue diagnosis," *Computerized Medical Imaging and Graphics*, vol. 31, no. 8, pp. 672–678, 2007.

[11] Z. Yang and N. Li, "Detection of tongue crack based on distant gradient and prior knowledge," *International Journal of Image and Graphics*, vol. 10, no. 2, pp. 273–288, 2010.

[12] Z. Yang, D. Zhang, and N. Li, "Kernel false-colour transformation and line extraction for fissured tongue image," *Journal of Computer-Aided Design & Computer Graphics*, vol. 22, no. 5, pp. 771–776, 2010.

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

[14] K. Tieu and P. Viola, "Boosting image retrieval," *International Journal of Computer Vision*, vol. 56, no. 1-2, pp. 17–36, 2004.

[15] Z. Fu, W. Li, X. Li, F. Li, and Y. Wang, "Automatic tongue location and segmentation," in *Proceedings of the International Conference on Audio, Language and Image Processing*, pp. 1050–1055, July 2008.

[16] L. Wang, X. Wang, and X. J. Xu, "Lip detection and tracking using variance based Haar-Like features and kalman filter," in *Proceedings of the 5th International Conference on Frontier of Computer Science and Technology (FCST '10)*, pp. 608–612, IEEE, Changchun, China, August 2010.

[17] Y. Cui, S. Liao, H. Wang, H. Liu, W. Wang, and L. Yin, "Relationship between hyperuricemia and haar-like features on tongue images," *BioMed Research International*, vol. 2015, Article ID 363216, 10 pages, 2015.

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

[19] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[20] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.

[21] H. Mamitsuka, "Selecting features in microarray classification using ROC curves," *Pattern Recognition*, vol. 39, no. 12, pp. 2393–2404, 2006.

[22] O. Komori and S. Eguchi, "A boosting method for maximizing the partial area under the ROC curve," *BMC Bioinformatics*, vol. 11, article 314, 2010.

[23] P. Long and R. Servedio, "Boosting the area under the roc curve," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '07)*, 2007.

[24] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 70, no. 5, pp. 849–911, 2008.

[25] R Core Team, *R: A Language and Environment for Statistical Computing*, R Core Team, 2012.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 842784, 7 pages
http://dx.doi.org/10.1155/2015/842784

Research Article

A Five-Gene Signature Predicts Prognosis in Patients with Kidney Renal Clear Cell Carcinoma

Yueping Zhan,[1] Wenna Guo,[1] Ying Zhang,[2] Qiang Wang,[3] Xin-jian Xu,[4] and Liucun Zhu[1]

[1] *School of Life Sciences, Shanghai University, Shanghai 200444, China*
[2] *Yangzhou Breeding Biological Agriculture Technology Co. Ltd., Yangzhou 225200, China*
[3] *State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing 210093, China*
[4] *Department of Mathematics, Shanghai University, Shanghai 200444, China*

Correspondence should be addressed to Xin-jian Xu; xinjxu@shu.edu.cn and Liucun Zhu; zhuliucun@shu.edu.cn

Received 28 June 2015; Revised 16 August 2015; Accepted 27 August 2015

Academic Editor: Tao Huang

Kidney renal clear cell carcinoma (KIRC) is one of the most common cancers with high mortality all over the world. Many studies have proposed that genes could be used to predict prognosis in KIRC. In this study, RNA expression data from next-generation sequencing and clinical information of 523 patients downloaded from The Cancer Genome Atlas (TCGA) dataset were analyzed in order to identify the relationship between gene expression level and the prognosis of KIRC patients. A set of five genes that significantly associated with overall survival time was identified and a model containing these five genes was constructed by Cox regression analysis. By Kaplan-Meier and Receiver Operating Characteristic (ROC) analysis, we confirmed that the model had good sensitivity and specificity. In summary, expression of the five-gene model is associated with the prognosis outcomes of KIRC patients, and it may have an important clinical significance.

1. Introduction

In recent years, the incidence and mortality of kidney cancer have been rising throughout the world [1]. In 2013, nearly 58,000 new cases occurred, and 130,001 patients died of kidney cancer in the United States [2]. Among them, kidney renal clear cell carcinoma (KIRC) is the most common histological subtype and accounts for 70%–80% of renal cancer cases [3]. KIRC tissue is resistant to traditional chemotherapeutic drugs [4], and patient outcomes varied a lot [5]. Although various researches have been done on KIRC, the clinical prognosis of KIRC patients still remains very poor; the survival time of 90% of patients with metastatic KIRC is less than 5 years [6]. Therefore, there is an urgent need to find potential molecular-based prognostic biomarkers in KIRC, and it is also one of the most important steps for prognostic prediction of patients.

Messenger RNA is one of the most common molecular markers. Many studies have suggested that genes were involved in the biological processes of many cancers and related to prognostic survival time of patients. For instance, *SIPL1* (Shank-Interacting Protein-Like 1) has reported to have overexpression during breast cancer tumorigenesis, and inhibiting the expression of *SIPL1* may contribute to inhibition of breast cancer [7]. *PLA2G16* has been proved as an important prognostic factor in primary osteosarcoma patients [8]. *Dicerl* has been found to be expressed at low level in nasopharyngeal carcinoma tissues no matter whether at the gene or at the protein levels, and it could also be a novel prognostic biomarker [9]. As for KIRC, several studies have been performed to detect gene expression signatures which may provide diagnostic and prognostic information [10–12]. Ge et al. have identified miRNA signature including 22 miRNAs as an independent novel predictor of patient outcomes [13]. Yu et al. have found that the expression of *CIDE* (cell death-inducing DFF45-like effector) is a novel predictor of prognosis [14]. However, detailed analyses of the associations between gene expression level and survival time of patients in KIRC remain limited.

The goal of this paper is identifying genes that are related to overall survival time of KIRC patients by analyzing high-throughput RNA sequencing data downloaded from TCGA [15]. In brief, the main goals are as follows: (1) identify genes that could predict the survival time of KIRC patient, and construct a model; (2) evaluate the prognostic value, sensitivity, and specificity of the model; and (3) investigate the independence and universality of the gene marker in different KIRC stages.

2. Materials and Methods

2.1. KIRC Gene Expression Data from TCGA. Up to January 2015, TCGA database (https://tcga-data.nci.nih.gov/tcga/) contained 533 KIRC patient samples [15]. The gene expression profiling was performed by using the Illumina HiSeq platforms (Illumina Inc., San Diego, CA, USA). After excluding patients without survival status information, UNC RNASeqV2 level 3 expression data for 523 patients including 20,531 human genes and corresponding clinical data were downloaded. Then the 523 KIRC samples were randomly divided into training set ($n = 262$) and testing set ($n = 261$). Specimen IDs in the two sets were shown in Supplemental Table S1 (in Supplementary Material available online at http://dx.doi.org/10.1155/2015/842784). Training set was used to identify gene expression signature, and the testing set was used for validation.

2.2. Statistical Analysis. Firstly, log 2 transformed was used for normalizing the RNA-seq expression values [16]. Subsequently, as previous reports [17, 18], genes that were significantly ($p < 0.001$) related to patient survival were identified by Cox regression analysis and random survival forests-variable hunting (RSFVH) algorithm [19]. Considering that a model with a smaller number of genes is generally accompanied with a practically better value, we performed Cox proportional-hazard regression analysis with two genes, three genes, and five genes, respectively, expecting to dig out a better model for predicting survival. Then, based on Cox regression analysis, a risk score formula was built to calculate the risk score for each patient. As reported by Margolin et al. [20] and Meng et al. [18], the survival differences between the low-risk and high-risk groups were evaluated, and the sensitivity and specificity of the model in the survival prediction were also compared.

3. Results

3.1. Patient Characteristics. All 523 patients used in this study were clinically and pathologically diagnosed with KIRC. Clinical stages of the tumor were classified into stages I to IV based on the Fuhrman nuclear grading system [21]. Here, there are 260 patients from stage 1, 57 patients from stage 2, 125 patients from stage 3, and 81 patients from stage 4, respectively. Additionally, the average age and average prognostic survival time of these 523 patients were 61 years and 902 days, respectively. All the statistical information was summarized in Table 1.

Table 1: Summary of patient demographics and clinical characteristics.

Characteristic	Patients			
	Training set	Testing set	Total	
Age				
Median	61	60	61	
Range	26–90	29–90	26–90	
Sex				
Male	164	174	338	64.63%
Female	98	87	185	35.37%
Vital status				
Living	173	184	357	68.26%
Dead	89	77	166	31.74%
Clinical stage				
Stage I	134	126	260	49.71%
Stage II	22	35	57	10.9%
Stage III	72	53	125	23.9%
Stage IV	34	47	81	15.49%

3.2. Detection of Genes Associated with Overall Survival Time of KIRC Patients in Training Set. To identify the gene which would be potentially associated with overall survival time of patients in KIRC, univariable Cox regression analysis (see Materials and Methods) for gene expression data was conducted in training set. With the significance level of 0.001, a total of 3,849 genes were identified (Table S2). Subsequently, 100 genes with the largest importance value in random survival forests analysis with default parameters [22, 23] were selected. Then, 1–5 genes were chosen from 100 genes as covariates by enumeration algorithm and 79,375,495 models were established in multivariate Cox regression analysis. After comparing with each other, the best model (indexed by AUROC) including 5 genes (*CKAP4, ISPD, MAN2A2, OTOF, and SLC40A1*) was determined, and the risk score formula for this model was (0.422 × expression value of *CKAP4*) + (−0.443 × expression value of *ISPD*) + (0.551 × expression value of *MAN2A2*) + (0.330 × expression value of *OTOF*) + (−0.369 × expression value of *SLC40A1*). The information of these five genes was shown in Table 2. And the functions of these genes were also summarized in Table 3. In addition, the error rate (27.27%) and variable importance values of these five genes were obtained with RSFVH (Figure 1). It can be seen from Figure 1 that the five genes have relatively large importance value; *CKAP4* has more importance than other predictors. Taking the median risk score as the cut-off, the 262 KIRC patients were separated into low-risk group ($n = 131$) and high-risk group ($n = 131$). Survival analysis was performed by using the Kaplan-Meier method with a log-rank statistical test. As shown in Figure 2(a), Kaplan-Meier curves indicated that patients in high-risk group have significantly ($p < 0.0001$) worse prognosis comparing with the low-risk group (Figure 2(a)).

3.3. Verification of Survival-Associated Genes in Testing Set. To determine the prognostic potential of the five-gene signature, Kaplan-Meier survival analysis was performed in testing

Table 2: Five genes significantly associated with the survival time of patients in the training set ($n = 262$).

Gene name	Parametric p value	Hazard ratio	Coefficient	Variable importance	Relative importance
CKAP4	$1.80E-09$	1.525	0.422	0.0365	1
SLC40A1	$9.30E-08$	0.691	−0.369	0.036	0.9862
OTOF	$4.60E-10$	1.391	0.33	0.28	0.7674
MAN2A2	0.00085	1.734	0.551	0.0192	0.5249
ISPD	$1.70E-05$	0.642	−0.443	0.0147	0.4012

Table 3: Five-gene functions' analysis.

Gene name	Chromosomal position	Start site	End site	Function	Study
CKAP4	chr12	106237881	106247935	Sequence specific DNA binding transcriptional activator or repressor	McHugh et al. [29] Li et al. [30] Zhang et al. [31]
ISPD	chr7	15916851	16530558	Mutations in ISPD cause Walker-Warburg syndrome	Willer et al. [36] Roscioli et al. [37]
MAN2A2	chr15	90902218	90922585	Catalyzes the committed step in the biosynthesis of complex N-glycans	Kroes et al. [38]
OTOF	chr2	26457203	26558698	Triggers membrane fusion and exocytosis	Padmanarayana et al. [33] Yildirim-Baylan et al. [22]
SLC40A1	chr2	189560590	189580811	Mediates cellular iron efflux	Moreno-Carralero et al. [34]

set. Just as it is in training set, based on the risk score of individual patient, patients in testing set were divided into low-risk and high-risk groups and Kaplan-Meier analysis was used to compare the patient survival differences. Statistically significant differences ($p < 0.0001$) between high-risk group and low-risk group were observed; in other words, higher risk score was related to shorter survival time (Figure 2(b)), which is in agreement with that in training set, revealing that five-gene signature may play an important role in predicting the survival of KIRC patients.

To further confirm the clinical performance of the five-gene model as a biomarker for predicting prognosis, the Receiver Operating Characteristic (ROC) analysis was performed for estimating the effect of the gene signature on patient survival. And the corresponding AUROC were calculated by hiring three years as the cut-off point. The AUROC was 0.783 (Figure 3), showing that the five-gene model has high sensitivity and specificity and could be used as a biomarker to predict the prognostic survival of patients.

3.4. The Independence and Universality of the Five-Gene Model. Studies have shown that age and clinical stage were also related to patient survival [5, 13, 21]. To examine whether the five-gene signature could distinguish the high-risk patients from low-risk patients when age of patients and stage were taken into account, multivariate Cox proportional hazard analyses were performed in both training and testing set. The results confirmed that risk score of five genes is independent of age and stage, as shown in Table 4. Besides, whether the five-gene signature was functional in different KIIRC stages was also investigated by using Kaplan-Meier and ROC analysis. Results showed that, in stage 3 and stage 4, the survival time of patients was dramatically different between high-risk group and low-risk group ($p < 0.001$, Figure S1). Moreover, the AUROC in stage 2, stage 3, and stage 4 were 0.761, 0.718, and 0.715, respectively (Figure S2), further revealing that the five-gene signature has predictive value in different clinical stages.

4. Discussion

KIRC is one of the most common primary renal malignancies with high morbidity and mortality [24]. However, the understanding of KIRC is not complete, and there are no clinical tools for predicting patient outcome apart from the traditional clinical parameters. Accurate data from the clinical examination of KIRC specimens could help doctors to decide appropriate treatment for patients [25]. Therefore, the identification and validation of novel biomarkers account for an important part of practical KIRC study [26]. In this study, we identified a five-gene signature that was significantly related to patient survival in KIRC based on genome-wide RNA profiling of 523 KIRC patients from TCGA database. In addition, we confirmed that the five-gene signature could be regarded as an independent predictor of prognostic survival after considering the various variables including age and stage, and it is also universal in different stages.

Many previous studies on genes in KIRC have mainly considered some known cancer-associated genes. For instance, Wei et al. have found that high expression of pituitary tumor-transforming gene-1 (*PTTG1*) in KIRC patients was associated with poor prognosis by using qRT-PCR and immunohistochemistry [27]. Peters et al. have proved that low gene expression levels of *GATA1* and *GATA2* were related to tumor aggressiveness and short survival time in KIRC [28]. With respect to the five genes

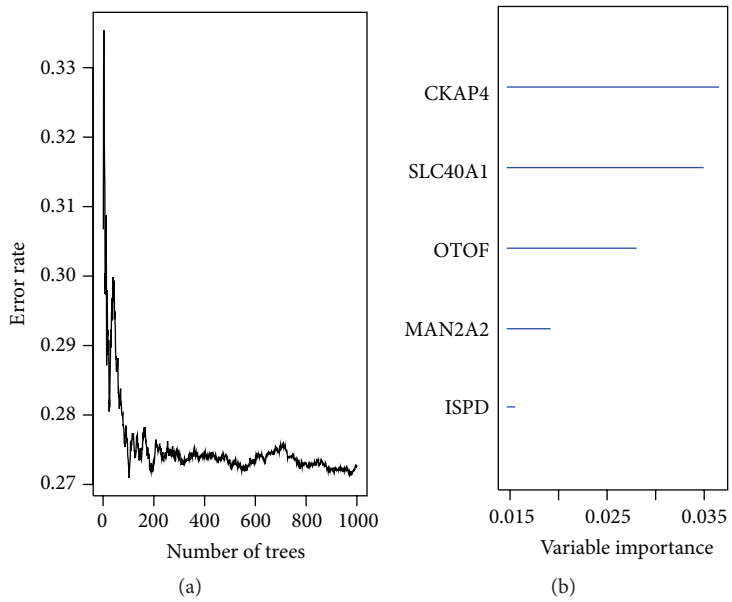


Figure 1: Random survival forests-variable hunting analysis reveals the error rate for the data as a function of trees (a) and the importance values for predictors (b). Importance values show the impact of genes on the model.

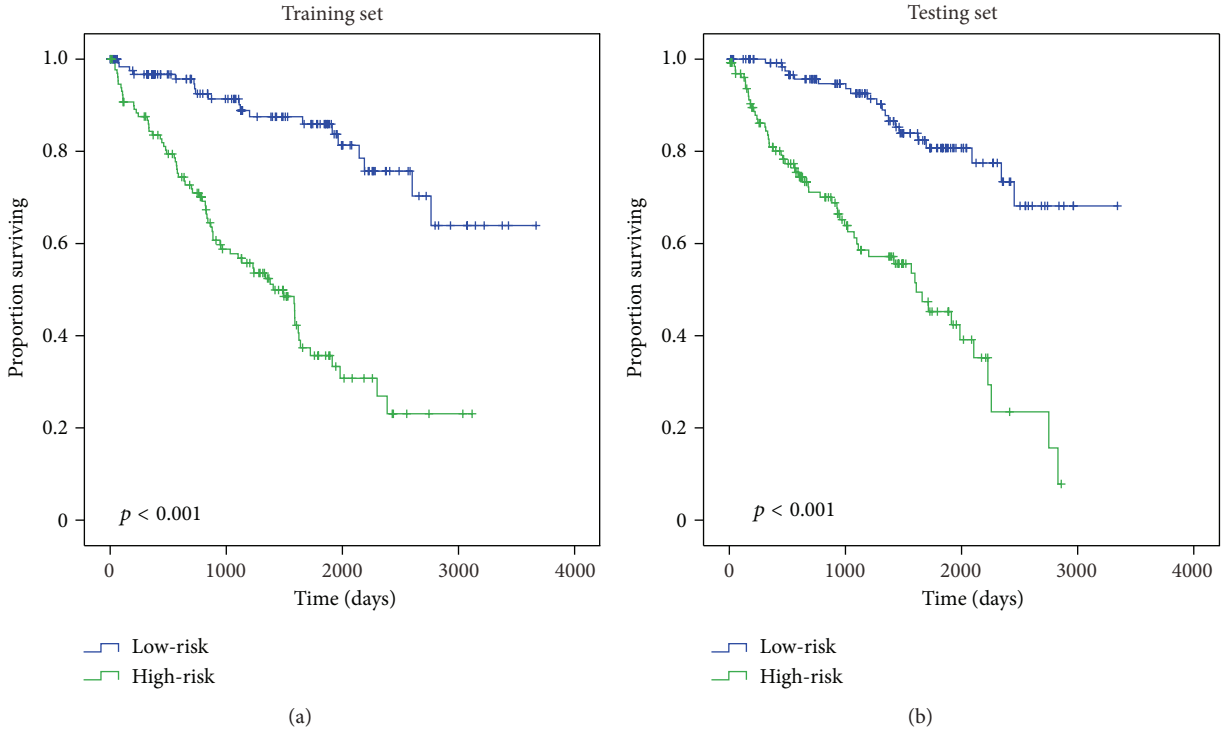


Figure 2: Kaplan-Meier curves with two-sided log-rank test show relationship between the risk score resulting from five genes and patients survival. Using the median risk score as a cut-off, patients were divided into the high-risk score and low-risk score. (a) Kaplan-Meier curves for training set patients ($n = 262$); (b) Kaplan-Meier curves for testing set patients ($n = 261$). The two-sided log-rank tests were used to determine the survival differences between the high-risk score and low-risk score.

we identified in this study, all of them have also been reported to be associated with cancer. It turned out that *CKAP4* could be used to distinguish primary salivary oncocytic lesions from metastatic RCC effectively in dubious cases with 100% accuracy [29] and related to lymphatic metastasis [30, 31]. Mutations in *OTOF*, which functionally triggers membrane fusion and exocytosis, may provide a link between calcium signaling and cancer [22, 32, 33].

Table 4: Univariable and multivariable Cox regression analyses in training and testing set.

Variables	Univariable model			Multivariable model		
	HR	95% CI of HR	p value	HR	95% CI of HR	p value
Training set ($N = 262$)						
Five-gene model	2.717	2.180–3.387	<0.001	2.752	2.193–3.454	<0.001
Age	1.031	1.014–1.050	0.001	1.032	1.009–1.048	0.003
Testing set ($N = 261$)						
Five-gene model	1.936	1.620–2.315	<0.001	1.875	1.560–2.253	<0.001
Age	1.022	1.004–1.041	0.019	1.011	0.993–1.031	0.234
Training set ($N = 262$)						
Five-gene model	2.717	2.180–3.387	<0.001	2.193	1.726–2.786	<0.001
Stage	2.097	1.734–2.537	<0.001	1.717	1.394–2.111	<0.001
Testing set ($N = 261$)						
Five-gene model	1.936	1.620–2.315	<0.001	1.700	1.390–2.078	<0.001
Stage	1.905	1.564–2.322	<0.001	1.679	1.367–2.062	<0.001

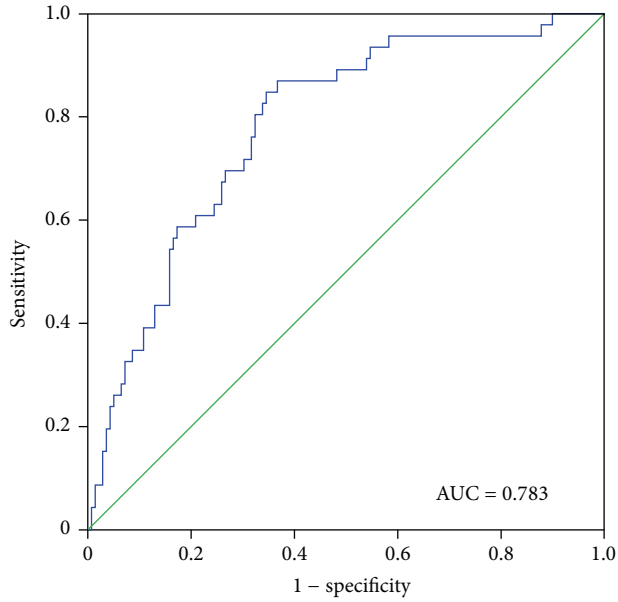


Figure 3: Receiver Operating Characteristic (ROC) analysis of the five-gene signature. The AUROC was 0.783 ($p < 0.001$), showing that the five-gene model has high sensitivity (true positive rate) and specificity (true negative rate) in predicting the survival time of KIRC patients.

SLC40A1 is a cell membrane protein that has been identified to mediate cellular iron efflux [23, 34] and contribute to the invasive phenotype [35]. Mutations in *ISPD* may cause Walker-Warburg syndrome [36, 37]. *MAN2A2* was downregulated in hepatocellular carcinoma [38]. However, up to now, such predictive markers were not analyzed in KIRC patients and the molecular study concerning these genes has not been reported in KIRC. Nevertheless, our research showed that the expressions of these genes were related to survival time of patients. ROC curve showed that the AUROC is approximately 0.8, considering that the larger AUROC usually implies a better model for prediction [6, 39], which further demonstrated that the five-gene signature in our study is a novel prognostic marker with high accuracy and has important clinical significance. Furthermore, the five-gene signature was an independent predictor, which was pervasive in different stages. In different stages, ROC analysis shows high sensitivity and specificity (AUROC >0.7) except stage 1, which is possibly because stage 1 is slow-growing tumor, cancer cells are not invasive and metastatic, and the number of patients that died of KIRC was smaller than that in other stages [40]. We found here that the average age of patients who died in stage 1 was more than 67, which is higher than in other stages, revealing that the age at diagnosis may have some influence on KIRC prognosis, and part of deaths was attributed to increased risk of disease mortality with increasing age. Therefore, these results suggested that the five-gene signature is significantly important in clinic. The functional mechanisms of these genes remain unclear. Moreover, the five-gene signature has not yet been tested in a clinical trial. The experimental studies on these genes and further well-designed studies should be conducted to verify our findings, thereby providing a better understanding of their roles in predicting KIRC prognosis.

5. Conclusions

In summary, a five-gene signature strongly associated with patients' survival was identified by performing Cox regression analysis and Kaplan-Meier analysis in training set. Subsequently, Kaplan-Meier and ROC analysis in testing set further indicated that the five-gene signature could be used as a novel biomarker to predict the treatment outcome of KIRC patient. Additionally, multivariate Cox regression analysis revealed that the five-gene signature was an independent predictor. These results suggested that the five-gene signature could help to predict the survival with significant clinical implications.

Abbreviations

KIRC: Kidney renal clear cell carcinoma
TCGA: The Cancer Genome Atlas
RSFVH: Random survival forests-variable hunting
ROC: Receiver Operating Characteristic.

Conflict of Interests

The authors declare that they have no competing interests.

Authors' Contribution

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

Acknowledgments

This work was supported by Shanghai Province Science Foundation for Youths (Grant no. 12ZR1444200), National Natural Science Foundation of China (Grant no. 31471200), Foundation for the Author of National Excellent Doctoral Dissertation of PR China (201134), Doctor Gathering Scheme of Jiangsu Province, and the High Performance Computing Platform of Shanghai University.

References

[1] R. Siegel, D. Naishadham, and A. Jemal, "Cancer statistics, 2012," *CA Cancer Journal for Clinicians*, vol. 62, no. 1, pp. 10–29, 2012.

[2] R. Siegel, D. Naishadham, and A. Jemal, "Cancer statistics, 2013," *CA: Cancer Journal for Clinicians*, vol. 63, no. 1, pp. 11–30, 2013.

[3] L. Zhang, B. Xu, S. Chen et al., "The complex roles of microRNAs in the metastasis of renal cell carcinoma," *Journal of Nanoscience and Nanotechnology*, vol. 13, no. 5, pp. 3195–3203, 2013.

[4] E. A. Singer, G. N. Gupta, D. Marchalik, and R. Srinivasan, "Evolving therapeutic targets in renal cell carcinoma," *Current Opinion in Oncology*, vol. 25, no. 3, pp. 273–280, 2013.

[5] J. R. Karamchandani, M. Y. Gabril, R. Ibrahim et al., "Profilin-1 expression is associated with high grade and stage and decreased disease-free survival in renal cell carcinoma," *Human Pathology*, vol. 46, no. 5, pp. 673–680, 2015.

[6] L. A. Tse, J. Dai, M. Chen et al., "Prediction models and risk assessment for silicosis using a retrospective cohort study among workers exposed to silica in China," *Scientific Reports*, vol. 5, Article ID 11059, 2015.

[7] J. De Melo and D. Tang, "Elevation of SIPL1 (SHARPIN) increases breast cancer risk," *PLoS ONE*, vol. 10, no. 5, Article ID e0127546, 2015.

[8] S. Liang, Z. Ren, X. Han et al., "PLA2G16 expression in human osteosarcoma is associated with pulmonary metastasis and poor prognosis," *PloS ONE*, vol. 10, no. 5, Article ID e0127236, 2015.

[9] S. Xu and X. Jiang, "Reduced expression of Diceri is associated with poor prognosis in patients with nasopharyngeal carcinoma," *Lin Chung Er Bi Yan Hou Tou Jing Wai Ke Za Zhi*, vol. 29, no. 2, pp. 126–131, 2015.

[10] A. R. Brannon, S. M. Haake, K. E. Hacker et al., "Meta-analysis of clear cell renal cell carcinoma gene expression defines a variant subgroup and identifies gender influences on tumor biology," *European Urology*, vol. 61, no. 2, pp. 258–268, 2012.

[11] A. R. Brannon, A. Reddy, M. Seiler et al., "Molecular stratification of clear cell renal cell carcinoma by consensus clustering reveals distinct subtypes and survival patterns," *Genes and Cancer*, vol. 1, no. 2, pp. 152–163, 2010.

[12] H. Zhao, B. Ljungberg, K. Grankvist, T. Rasmuson, R. Tibshirani, and J. D. Brooks, "Gene expression profiling predicts survival in conventional renal cell carcinoma," *PLoS Medicine*, vol. 3, no. 1, article e13, 2006.

[13] Y.-Z. Ge, R. Wu, H. Xin et al., "A tumor-specific microRNA signature predicts survival in clear cell renal cell carcinoma," *Journal of Cancer Research and Clinical Oncology*, vol. 141, no. 7, pp. 1291–1299, 2015.

[14] M. Yu, H. Wang, J. Zhao et al., "Expression of CIDE proteins in clear cell renal cell carcinoma and their prognostic significance," *Molecular and Cellular Biochemistry*, vol. 378, no. 1-2, pp. 145–151, 2013.

[15] The Cancer Genome Atlas Research Network, "Comprehensive molecular characterization of clear cell renal cell carcinoma," *Nature*, vol. 499, no. 7456, pp. 43–49, 2013.

[16] Y. Li, J. M. Krahn, G. P. Flake, D. M. Umbach, and L. Li, "Toward predicting metastatic progression of melanoma based on gene expression data," *Pigment Cell & Melanoma Research*, vol. 28, no. 4, pp. 453–463, 2015.

[17] A. E. Zou, J. Ku, T. K. Honda et al., "Transcriptome sequencing uncovers novel long noncoding and small nucleolar RNAs dysregulated in head and neck squamous cell carcinoma," *RNA*, vol. 21, no. 6, pp. 1122–1134, 2015.

[18] J. Meng, P. Li, Q. Zhang, Z. Yang, and S. Fu, "A four-long noncoding RNA signature in predicting breast cancer survival," *Journal of Experimental & Clinical Cancer Research*, vol. 33, article 84, 2014.

[19] H. Ishwaran and U. B. Kogalur, "Consistency of random survival forests," *Statistics & Probability Letters*, vol. 80, no. 13-14, pp. 1056–1064, 2010.

[20] A. A. Margolin, E. Bilal, E. Huang et al., "Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer," *Science Translational Medicine*, vol. 5, no. 181, Article ID 181re1, 2013.

[21] S. J. Nam, C. Lee, J. H. Park, and K. C. Moon, "Decreased PBRM1 expression predicts unfavorable prognosis in patients with clear cell renal cell carcinoma," *Urologic Oncology: Seminars and Original Investigations*, vol. 33, no. 8, pp. 340.e9–340.e16, 2015.

[22] M. Yildirim-Baylan, G. Bademci, D. Duman, H. Ozturkmen-Akay, S. Tokgoz-Yilmaz, and M. Tekin, "Evidence for genotype-phenotype correlation for OTOF mutations," *International Journal of Pediatric Otorhinolaryngology*, vol. 78, no. 6, pp. 950–953, 2014.

[23] N. Montalbetti, A. Simonin, G. Kovacs, and M. A. Hediger, "Mammalian iron transporters: families SLC11 and SLC40," *Molecular Aspects of Medicine*, vol. 34, no. 2-3, pp. 270–287, 2013.

[24] J. R. Srigley, B. Delahunt, J. N. Eble et al., "The International Society of Urological Pathology (ISUP) Vancouver classification of renal neoplasia," *The American Journal of Surgical Pathology*, vol. 37, no. 10, pp. 1469–1489, 2013.

[25] H. Moch, J. Srigley, B. Delahunt, R. Montironi, L. Egevad, and P. H. Tan, "Biomarkers in renal cancer," *Virchows Archiv*, vol. 464, no. 3, pp. 359–365, 2014.

[26] S. Gulati, P. Martinez, T. Joshi et al., "Systematic evaluation of the prognostic impact and intratumour heterogeneity of clear cell renal cell carcinoma biomarkers," *European Urology*, vol. 66, no. 5, pp. 936–948, 2014.

[27] C. Wei, X. Yang, J. Xi et al., "High expression of pituitary tumor-transforming gene-1 predicts poor prognosis in clear cell renal cell carcinoma," *Molecular and Clinical Oncology*, vol. 3, no. 2, pp. 387–391, 2015.

[28] I. Peters, N. Dubrowinskaja, H. Tezval et al., "Decreased mRNA expression of GATA1 and GATA2 is associated with tumor aggressiveness and poor outcome in clear cell renal cell carcinoma," *Targeted Oncology*, vol. 10, no. 2, pp. 267–275, 2015.

[29] J. B. McHugh, A. P. Hoschar, M. Dvorakova, A. V. Parwani, E. L. Barnes, and R. R. Seethala, "p63 immunohistochemistry differentiates salivary gland oncocytoma and oncocytic carcinoma from metastatic renal cell carcinoma," *Head and Neck Pathology*, vol. 1, no. 2, pp. 123–131, 2007.

[30] M.-H. Li, L.-W. Dong, S.-X. Li et al., "Expression of cytoskeleton-associated protein 4 is related to lymphatic metastasis and indicates prognosis of intrahepatic cholangiocarcinoma patients after surgery resection," *Cancer Letters*, vol. 337, no. 2, pp. 248–253, 2013.

[31] J. Zhang, S. L. Planey, C. Ceballos, S. M. Stevens Jr., S. K. Keay, and D. A. Zacharias, "Identification of CKAP4/p63 as a major substrate of the palmitoyl acyltransferase DHHC2, a putative tumor suppressor, using a novel proteomics method," *Molecular and Cellular Proteomics*, vol. 7, no. 7, pp. 1378–1388, 2008.

[32] X. Jiao, L. D. Wood, M. Lindman et al., "Somatic mutations in the notch, NF-KB, PIK3CA, and hedgehog pathways in human breast cancers," *Genes, Chromosomes and Cancer*, vol. 51, no. 5, pp. 480–489, 2012.

[33] M. Padmanarayana, N. Hams, L. C. Speight, E. J. Petersson, R. A. Mehl, and C. P. Johnson, "Characterization of the lipid binding properties of otoferlin reveals specific interactions between PI(4,5)P2 and the C2C and C2F Domains," *Biochemistry*, vol. 53, no. 30, pp. 5023–5033, 2014.

[34] M.-I. Moreno-Carralero, J.-A. Muñoz-Muñoz, N. Cuadrado-Grande et al., "A novel mutation in the SLC40A1 gene associated with reduced iron export in vitro," *American Journal of Hematology*, vol. 89, no. 7, pp. 689–694, 2014.

[35] S. Weissmueller, E. Manchado, M. Saborowski et al., "Mutant p53 drives pancreatic cancer metastasis through cell-autonomous PDGF receptor beta signaling," *Cell*, vol. 157, no. 2, pp. 382–394, 2014.

[36] T. Willer, H. Lee, M. Lommel et al., "*ISPD* loss-of-function mutations disrupt dystroglycan *O*-mannosylation and cause Walker-Warburg syndrome," *Nature Genetics*, vol. 44, no. 5, pp. 575–580, 2012.

[37] T. Roscioli, E.-J. Kamsteeg, K. Buysse et al., "Mutations in ISPD cause Walker-Warburg syndrome and defective glycosylation of α-dystroglycan," *Nature Genetics*, vol. 44, no. 5, pp. 581–585, 2012.

[38] R. A. Kroes, G. Dawson, and J. R. Moskal, "Focused microarray analysis of glyco-gene expression in human glioblastomas," *Journal of Neurochemistry*, vol. 103, supplement 1, pp. 14–24, 2007.

[39] P. J. Heagerty, T. Lumley, and M. S. Pepe, "Time-dependent ROC curves for censored survival data and a diagnostic marker," *Biometrics*, vol. 56, no. 2, pp. 337–344, 2000.

[40] Q. Liu, P. F. Su, S. Zhao, and Y. Shyr, "Transcriptome-wide signatures of tumor stage in kidney renal clear cell carcinoma: connecting copy number variation, methylation and transcription factor activity," *Genome Medicine*, vol. 6, no. 12, p. 117, 2014.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 674296, 10 pages
http://dx.doi.org/10.1155/2015/674296

Review Article

Survey of Natural Language Processing Techniques in Bioinformatics

Zhiqiang Zeng,[1] Hua Shi,[1] Yun Wu,[1] and Zhiling Hong[2]

[1]*College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China*
[2]*Software School, Xiamen University, Xiamen 361005, China*

Correspondence should be addressed to Zhiling Hong; hongzl@xmu.edu.cn

Received 10 May 2015; Revised 12 June 2015; Accepted 21 June 2015

Academic Editor: Tao Huang

Informatics methods, such as text mining and natural language processing, are always involved in bioinformatics research. In this study, we discuss text mining and natural language processing methods in bioinformatics from two perspectives. First, we aim to search for knowledge on biology, retrieve references using text mining methods, and reconstruct databases. For example, protein-protein interactions and gene-disease relationship can be mined from PubMed. Then, we analyze the applications of text mining and natural language processing techniques in bioinformatics, including predicting protein structure and function, detecting noncoding RNA. Finally, numerous methods and applications, as well as their contributions to bioinformatics, are discussed for future use by text mining and natural language processing researchers.

1. Introduction

Text mining and natural language processing refer to comprehending and analyzing natural language by using computer algorithms and programs. It is an important research direction in the application field of artificial intelligence. Research on natural language processing and text mining has been reported as early as the emergence of computers. With continuous and extensive research on machine learning and data mining algorithms, existing text mining technologies have achieved good results in automatic abstraction, automatic question answering, web relational network analysis, and anaphora resolution [1, 2].

Bioinformatics is an interdiscipline that emerged with the progress and accomplishment of the Human Genome Project. It predicts and solves live science problems related to genetics by using computer and statistical informatics. Data storage, retrieval, and analysis are the key processes in bioinformatics [3–7]. The National Center for Biotechnology Information established various databases for biological data, including sequence databases for storing DNA and protein data (e.g., dbEST and dbSNP) [8, 9], Online Mendelian Inheritance in Man database for storing disease data, Gene Expression Omnibus database for storing gene chip data, and PubMed database for storing biological and medical literature [10].

Text mining and natural language processing techniques are necessary to retrieve user preference knowledge from expanding databases. Therefore, researchers retrieve papers on certain topics of interest, such as determining protein-protein interactions, from PubMed using computer algorithms and programs. With the cracking of genetic codes, researchers have determined that biological sequences, particularly protein sequences, are similar to human language in terms of composition. In addition to using text mining to retrieve bioinformatics articles directly, an increasing number of researchers are regarding protein sequences as a special "text" and analyzing them based on existing text mining technologies. The relationship between bioinformatics and natural language processing is shown in Figure 1. Researchers have also predicted the structures and functions of proteins. Based on these two aspects, we summarize the text mining technologies used in bioinformatics research. We aim to present these technologies to more bioinformatics researchers and hope that the number of researchers who can use good text mining technologies in bioinformatics studies will increase.

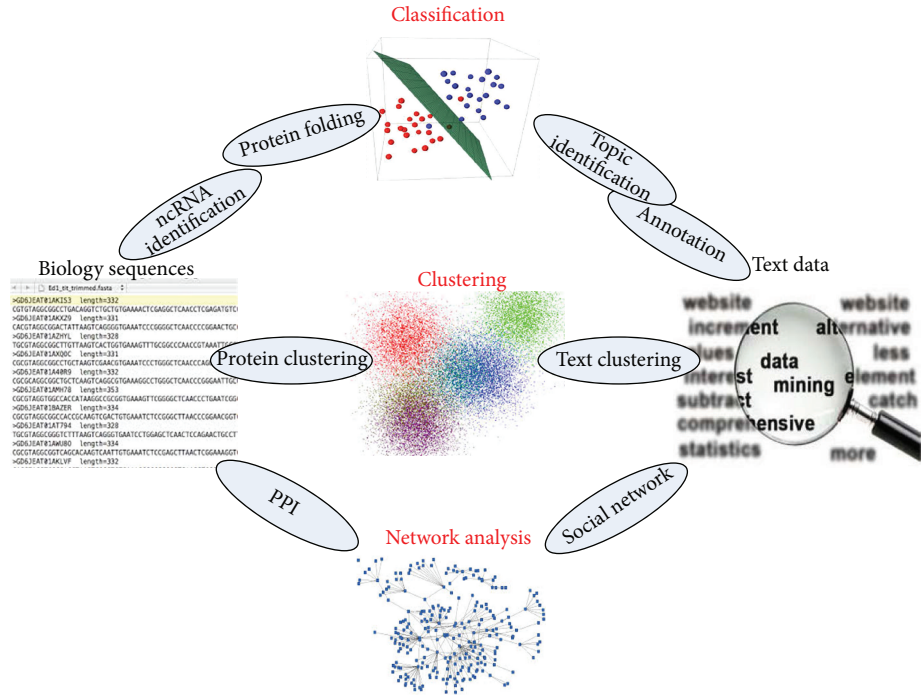


Figure 1: Problems and methodology relationship between NLP and bioinformatics.

2. Mining Bioinformatics Literature

The development of text mining technology plays an important role in retrieving biological literature, particularly in establishing biological information databases. A special workshop on biological literature retrieval problems was conducted during the Annual Meeting of the Association for Computational Linguistics and the Annual International Conference on Intelligent Systems for Molecular Biology in 2005 to discuss literature mining problems related to bioinformatics. Extracting protein-protein interactions and the relationship between gene functions and diseases are two leading application subjects.

2.1. Extracting Protein-Protein Interactions. Extracting the protein interaction network is an important research topic in bioinformatics and systems biology [11–14]. In previous studies, researchers searched for protein-protein interactions manually. However, with the exponential growth of biological literature, a program that can recognize protein-protein interactions automatically from PubMed abstracts is necessary. Nevertheless, no unified naming rule for proteins has been established yet. Many proteins and genes use the same name. Consequently, recognizing protein names from the literature abstracts and further determining their interactions are key problems in the application of text mining in searching for protein-protein interactions.

Initially, researchers extracted protein-protein interactions through statistical and counting methods. They manually created dictionaries of protein names and then searched abstracts that involve elements occurring at least twice. On this basis, researchers determined that associated proteins interact with one another [15]. Some researchers also used dynamic planning to extract and compare protein-protein interactions [16].

Extracting protein-protein interactions has been a research hot spot in bioinformatics for a long time and has attracted an increasing number of researchers in the fields of text mining and natural language processing. First, the grammar of literature abstracts is analyzed more carefully, rather than making a simple statistics of dictionary words. Kim et al. converted a complicated semantic structure analysis into calculating the shortest path in a graph by creating a nucleus [17]. Similar analysis methods of literature abstracts include grammatical analysis [18–21], context-free grammar analysis [22], ontology analysis [23], and other information retrieval methods. Protein-protein interactions are examined using these analysis methods. In addition, many machine learning methods, such as ensemble learning [24] and Bayesian network [25], are applied to recognize protein names and interactions.

2.2. Extracting the Relationship between Gene Functions and Diseases. Extracting protein-protein interactions involves searching for two proteins in the text and determining whether they interact with each other. Similarly, extracting the relationship between gene functions and diseases also involves searching for gene names and disease names simultaneously in the literature and then determining whether a particular gene is related to a certain disease [26].

In general, such extraction process can be divided into three steps. First, the abstracts of associated papers are searched through comparison with a dictionary. Second, the search scope has to be expanded forward and backward sometimes based on the location of the related word or clause

to ensure accuracy. Finally, facts are evaluated using grammar analysis methods or machine learning methods. Such extraction methods frequently yield good results for special genes and diseases. Bui et al. examined the relationship between drugs and HIV variation in PubMed [27]. Jiang et al. determined the relationship between approximately 3000 microRNAs and different diseases based on the naming rule of microRNA [28]. Cheng et al. developed a text mining system based on the relationship among human diseases, variations, and drug effects [29]. Iossifov et al. focused on investigating malformations of human and mouse encephalon [30]. Jensen et al. made a detailed summary of related document databases, literature mining software, and functions [31].

2.3. Retrieving References. A considerable amount of bioscience literature has been published. Searching for interacting proteins and examining the relationship between genes and diseases are only two application cases. Text mining technology is required to obtain answers to many other bioscience and bioinformatics problems in various databases, such as PubMed.

Biological literature mining and related problem solving have to cope with two major problems, namely, recognizing name entities and extracting relations. These problems are mainly solved by (1) methods based on linguistic analysis [32], (2) methods based on dictionaries [33], (3) machine learning methods [34, 35], and (4) statistical methods [36].

Several important databases are also selected with text mining. STRING [37] and BioGRID [38] are built for protein-protein interaction with literature mining. For predicting gene function, PubTator [39] and GeneCards [40] are important databases using text mining techniques. Related works were reviewed in detail in Huang and Lu's work [41] recently. As the development of crowdsource, artificial text searching and mining can also be helpful for biomedicine literature collection [42].

Moreover, converting PubMed database into an Extensible Markup Language relational database [43] and a fuzzy search of papers and author names through short-term matching are also current research hot spots [44].

3. Applying Text Mining Technologies to Protein Research

DNA and protein sequences are a meaningful genetic language and are regarded as the sealed book of life. Therefore, an increasing number of natural language processing and text mining algorithms are being applied to study bioinformatics. For example, latent semantic analysis was applied to protein remote homology detection [45, 46], and protein spectral analysis originates from word frequency statistics in natural language processing. Furthermore, some grammar rules of protein, DNA, and RNA sequences were discovered, and several web servers were constructed so as to extract these features and rules [47].

3.1. Predicting Protein Structure. Protein structure determines function [48]. Hence, it should be analyzed to determine protein function. The structural analysis of protein mainly focuses on certain protein sequences and classifies regions into the α-helix, β-lamella, and protein disordered regions. Predicting the α-helix and β-lamella regions is the same as predicting the secondary protein structure.

If a protein sequence is regarded as a natural language, then analyzing the type of protein in a region is similar to calibrating grammar in natural language processing. First, the secondary protein structure is predicted by combining rules and statistics [49–52]. However, faced with the bottleneck of statistical prediction, some researchers have proposed using machine learning prediction methods, including methods based on artificial neural network (ANN) [53], support vector machine (SVM) [54, 55], random forest [56–58], and maximum entropy [59].

Predicting the protein disordered region is also conducted. This region refers to the area without a stable or unique 3D structure in the protein space structure. Many text mining and machine learning methods, including ANN [60–62], SVM [63–65], conditional random field [66], and random forest [67], have been used to predict the protein disordered region. Common existing server addresses are listed in Table 1.

3.2. Predicting Protein Function. Predicting protein function is one of the most basic research topics in bioinformatics. It involves predicting protein-protein interactions and interaction sites [68, 69], localizing subcellular protein [70–78], predicting and classifying transmembrane protein [79–82], protein remote homology detection [83, 84], classifying protein functions [85–93], recognizing multifunctional enzymes [94–96], and DNA binding protein identification [97, 98].

The protein sequence is easy to determine. Similar to natural language, the protein sequence has many complicated rules. However, summarizing and understanding the rules of protein sequences are difficult. Therefore, analyzing and predicting the "protein language" expressed by amino acid sequences by using computational linguistics and machine learning methods are necessary. Through these procedures, we may be able to understand the functions of protein sequences.

Predicting protein-protein interactions is one of the most basic research topics in protein functions. Many researchers are committed to predicting whether two protein sequences exhibit interactions. To date, many machine learning methods have been applied, including SVM [99], kernel method [100, 101], decision-making tree [102, 103], random forest [104], Bayesian network [105], and the autoregressive model [106]. Several text processing methods, such as ontology annotation and sample weighting [107], are used to detect features and process training data. When predicting protein-protein interactions, researchers also aim to analyze the region of protein-protein interactions, which is used to predict protein-protein interaction sites. Information approaches commonly used in grammatical analyses, such as condition random fields [108] and a hidden Markov model (HMM) [109], have been used to analyze interaction sites and have achieved good results. Moreover, random forest [110], SVM [111], ANN [112], Bayesian network [113], linear regression [114], and other machine learning methods

Table 1: Web server for protein disorder prediction.

Problem	Name	Websites	Input format
Protein disorder prediction	DisProt	http://www.disprot.org/pondr-fit.php http://www.disprot.org/metapredictor.php http://www.dabi.temple.edu/disprot/predictor.php	Fasta or EMBL sequence format
	DisEMBL	http://dis.embl.de/	SwissProt ID
	DRIPPRED	http://www.sbc.su.se/~maccallr/disorder/cgi-bin/submit.cgi	Only plain sequence; one sequence once; slow
	FoldIndex	http://bip.weizmann.ac.il/fldbin/findex	Only plain sequence; one sequence once
	IUPred	http://iupred.enzim.hu/	SwissProt ID or plain sequence
	PONDR	http://www.pondr.com/cgi-bin/PONDR/pondr.cgi	Fasta
	PSIPRED	http://bioinf.cs.ucl.ac.uk/psipred/?disopred=1	Raw sequence or fasta format
	SCRATCH	http://scratch.proteomics.ics.uci.edu/	Only plain sequence; one sequence once; slow
	Spritz	http://distill.ucd.ie/spritz/	Raw sequence or fasta format
	RONN	http://www.strubi.ox.ac.uk/RONN/	Fasta, but only one sequence once

Table 2: Web server for protein-protein interaction and sites prediction.

Problem	Name	Websites	Input format
Protein interaction sites prediction	PPISP	http://pipe.scs.fsu.edu/ppisp.html http://pipe.scs.fsu.edu/meta-ppisp.html	PDB file
	Protemot	http://protemot.csbb.ntu.edu.tw/index.html	PDB ID
	SPPIDER	http://sppider.cchmc.org	PDB file or PDB ID
	Whiscy	http://nmr.chem.uu.nl/Software/whiscy/index.html	PDB file
Protein-protein interaction prediction	InterPreTS	http://www.russell.embl.de/cgi-bin/tools/interprets.pl	Fasta, 40 sequences at most
	PIE	http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/PIE/	Gene ID or name
	PPI	http://121.192.180.204:8080/PPI/Home.jsp	Fasta
	PredHS	http://www.predhs.org/	PDB files, 10 files at most
	Pred-PPI	http://cic.scu.edu.cn/bioinformatics/predict_ppi/default.html	Two fasta sequences
	Prism	http://cosbi.ku.edu.tr/prism/	Two PDB IDs or PDB files
	Struct2Net	http://groups.csail.mit.edu/cb/struct2net/webserver/	Gene names or keywords

are used to predict protein-protein interaction sites. Nevertheless, some researchers doubt that determining the protein sequence alone is inadequate to provide sufficient information for predicting interactions [115]. Text mining and machine learning researchers should develop new features and classification methods to solve this problem. The websites of existing common software used to predict protein-protein interactions and interaction sites are provided in Table 2.

4. Applying Natural Language Processing Techniques to Noncoding RNA Identification

4.1. Comparative RNA Prediction Methods. Alignment is also an important topic in natural language processing. DNA or RNA sequences can also be viewed as text. Sequence-based multiple sequence alignment methods can be used only at the sequence similarity level. The secondary structures of ncRNAs are usually more conserved than their sequences [116, 117]; for example, miRNA precursors share the common hairpin-like structure and tRNAs form cloverleaf structures [118, 119]. The functions of many ncRNAs are therefore determined by their secondary structure rather than by their sequences. As a result, structure-based multiple sequence alignment methods have been developed to align an input sequence to known ncRNA structures to determine the ncRNA class to which the input sequence belongs.

LocARNA [120] can produce fast and high-quality pairwise and multiple alignments of RNA sequences. It uses a complex RNA energy model for simultaneous folding and sequence/structure alignment of the RNAs. LocARNA performs global and local sequence alignments as well as local structural alignment of RNA molecules. An upgraded version of LocARNA, called LocARNA-P, has been developed recently [121]. The new version incorporates a probabilistic model that can compute accurate multiple alignments based on a probabilistic consistency transformation and reliability profiles for assessing local alignment quality and localizing RNA motifs. These features are based on computing sequence and structure match probabilities based on the LocARNA alignment model.

Table 3: Multiple sequence alignment tools.

Tool	Alignment method	URL
BLAT	Sequence-based	http://genome.ucsc.edu/
BLAST		http://www.ncbi.nlm.nih.gov/
BWA-SW		http://bio-bwa.sourceforge.net
Multilign	Structure-based	http://rna.urmc.rochester.edu/
FoldalignM		http://foldalign.ku.dk/
LocARNA/LocARNA-P		http://www.bioinf.uni-freiburg.de/Software/LocARNA/
MASTR		http://mastr.binf.ku.dk/
RAF		http://contra.stanford.edu/contrafold/
RNASampler		http://ural.wustl.edu/software.html
RNAshapes		http://bibiserv.techfak.uni-bielefeld.de/rnashapes/
RNAalifold		http://www.tbi.univie.ac.at/RNA/
StemLoc		N.A.
MAFFT		http://mafft.cbrc.jp/alignment/software/index.html
MiRAlign		http://bioinfo.au.tsinghua.edu.cn/miralign/

Table 4: miRNA identification methods.

Method	URL	Online service	Local service
MiPred	http://www.bioinf.seu.edu.cn/miRNA/	√	√
microPred	http://www.cs.ox.ac.uk/people/manohara.rukshan.batuwita/microPred.htm		√
TripletSVM	http://bioinfo.au.tsinghua.edu.cn/mirnasvm		√
PlantMiRNAPred	http://nclab.hit.edu.cn/PlantMiRNAPred/	√	√
miRNApre	http://121.192.180.205:8080/miRNApreWeb/	√	√
MIReNA	http://www.ihes.fr/~carbone/data8/		√
HuntMi	http://adaa.polsl.pl/agudys/huntmi/huntmi.htm		√
Mirident	http://www.regulatoryrna.org/pub/mirident		√
CSHMM	http://web.iitd.ac.in/~sumeet/mirna/		√
HeteroMirPred	http://ncrna-pred.com/premiRNA.html	√	√

Although comparative methods perform well in most cases, they have three intrinsic limitations: (1) they are highly dependent on the availability of homologous sequences or structures and cannot make predictions when no relevant sequence similarity or structure similarity is available; (2) they cannot correctly identify real ncRNAs that have low homology with known ncRNAs; and (3) they can identify only ncRNAs that are homologous with members of known ncRNA classes but cannot identify members of novel ncRNA classes. Most lncRNAs (long noncoding RNAs) cannot be predicted using comparative methods because they do not have specific structures or sequence similarity. These limitations mean that comparative methods display low specificity for identifying ncRNAs. The multiple sequence alignment tools that are currently available are listed in Table 3.

4.2. Noncomparative RNA Prediction Methods. The noncomparative methods are independent of homologous information and can, therefore, detect nonconserved ncRNAs. Most noncomparative methods employ machine learning techniques to make the predictions [122], which are similar to the text mining techniques.

Table 5: Secondary prediction tools.

Tool	URL
RNAfold	http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi
RNAstructure	http://rna.urmc.rochester.edu/rnastructure.html
mfold	http://www.bioinfo.rpi.edu/applications/mfold/
vsfold	http://www.rna.it-chiba.ac.jp/~vsfold/vsfold4/
evofold	http://users.soe.ucsc.edu/~jsp/EvoFold/
sfold	http://sfold.wadsworth.org/cgi-bin/index.pl

Because of the importance of RNA structure, several computational RNA folding tools have been developed, such as mfold, RNAfold, vsfold, evofold, and sfold. Generally, these algorithms determine the folded secondary structure from and input sequence by optimizing the intermolecular base pairing to minimize the free energy. Some miRNA identification methods are shown in Table 4 and existing RNA secondary prediction tools are listed in Table 5.

5. Conclusion and Future Research

As research on natural language and text mining methods develops, different application fields will be the key to future

studies. Interdisciplines represented by bioinformatics are becoming the focus of an increasing number of information science researchers. The application of text mining technologies and methods in bioinformatics study will become the focus of text mining researchers. Meanwhile, bioinformatics researchers have to learn text mining technologies intensively to solve specific bioinformatics problems.

In retrieving biological literature, apart from the aforementioned prediction of protein-protein interactions and gene-disease relationship, many problems, particularly those that require updating literature retrieval results, such as the relationships between adverse drug reaction and molecule composition as well as among single nucleotide polymorphism sites, diseases, and adverse drug effects, require the use of text mining to search for related knowledge in a literature database.

In bioinformatics, nearly all studies related to proteomics and predicting protein structure according to amino acid sequences can be conducted using text mining and natural language processing technology. Many mature texts mining technologies, such as word frequency statistics, condition random fields, HMM, and context-free grammar, have been successfully applied to predict secondary protein structures, irregular regions, interactions, and interaction sites. However, the latest research results in text mining and natural language processing should be verified by applying them in protein and DNA languages. No effective computation method is available yet for predicting third and fourth protein structures, protein homology remote detection, protein disordered region detection, interaction network establishment, and drug target prediction. Information science researchers should develop and provide more effective algorithms. In addition, new machine learning and text mining methods (e.g., semisupervised learning and active learning) have been proposed and will be applied in biological literature retrieval and bioinformatics. At present, recommending systems based on feedback has become a new hot spot problem in retrieving biological literature. And the Hadoop technique for big data is another hot spot for biology sequences [123].

The development of bioinformatics relies on information science. In particular, text mining and natural language processing researchers should provide a more extensive application space. Researchers of text mining algorithms should develop more effective intelligent algorithms based on the characteristics of biological data. This study does not only summarize text mining methods used in bioinformatics and corresponding problems, but it also provides related websites of successful prediction software. Recently, text mining researchers who are involved in bioinformatics can test and compare different types of software. The authors hope that the number of text mining researchers who can apply their own methods in bioinformatics will increase, which will facilitate the development of bioinformatics and even genetic studies.

Conflict of Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by Natural Science Foundation of China (Grant no. 31200769), the Natural Science Foundation of Fujian Province of China (Grants no. 2013J05103 and no. 2014J01253), Xiamen Science and Technology Planning Project (Grant no. 3502Z20143030), and Scientific Research Plan Project of Fujian Education Department (Grants nos. JB12184 and JB09203).

References

[1] C. Lin, Z. Huang, F. Yang, and Q. Zou, "Identify content quality in online social networks," *IET Communications*, vol. 6, no. 12, pp. 1618–1624, 2012.

[2] L. Chen, L. Chun, L. Ziyu, and Z. Quan, "Hybrid pseudo-relevance feedback for microblog retrieval," *Journal of Information Science*, vol. 39, no. 6, pp. 773–788, 2013.

[3] Y. Li, C. Wang, Z. Miao et al., "ViRBase: a resource for virus-host ncRNA-associated interactions," *Nucleic Acids Research*, vol. 43, no. 1, pp. D578–D582, 2015.

[4] L. Wang, K. Qian, Y. Huang et al., "SynBioLGDB: a resource for experimentally validated logic gates in synthetic biology," *Scientific Reports*, vol. 5, article 8090, 2015.

[5] Y. Wang, L. Chen, B. Chen et al., "Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network," *Cell Death & Disease*, vol. 4, no. 8, article e765, 2013.

[6] X. Zhang, D. Wu, L. Chen et al., "RAID: a comprehensive resource for human RNA-associated (RNA-RNA/RNA-protein) interaction," *RNA*, vol. 20, no. 7, pp. 989–993, 2014.

[7] Y. Li, L. Zhuang, Y. Wang et al., "Connect the dots: a systems level approach for analyzing the miRNA-mediated cell death network," *Autophagy*, vol. 9, no. 3, pp. 436–439, 2013.

[8] J. Wang, Q. Zou, and M. Z. Guo, "Mining SNPs from EST sequences using filters and ensemble classifiers," *Genetics and Molecular Research*, vol. 9, no. 2, pp. 820–834, 2010.

[9] J. Wang, L. Zhang, Q. Zou, J. Tan, X. Chen, and Y. Wu, "Association studies on mtDNA and Parkinson's disease population discrimination using the statistical classification," *Current Bioinformatics*, vol. 9, no. 5, pp. 481–489, 2014.

[10] Q. Zou, J. Li, Q. Hong et al., "Prediction of microRNA-disease associations based on social network analysis methods," *BioMed Research International*. In press.

[11] B. Liu, X. Wang, L. Lin, B. Tang, Q. Dong, and X. Wang, "Prediction of protein binding sites in protein structures using hidden Markov support vector machine," *BMC Bioinformatics*, vol. 10, article 381, 2009.

[12] F. Guo, S. C. Li, P. Du, and L. Wang, "Probabilistic models for capturing more physicochemical properties on protein-protein interface," *Journal of Chemical Information and Modeling*, vol. 54, no. 6, pp. 1798–1809, 2014.

[13] F. Guo, S. C. Li, L. Wang, and D. Zhu, "Protein-protein binding site identification by enumerating the configurations," *BMC Bioinformatics*, vol. 13, article 158, 2012.

[14] F. Guo, S. C. Li, and L. Wang, "Protein-protein binding sites prediction by 3D structural similarities," *Journal of Chemical Information and Modeling*, vol. 51, no. 12, pp. 3287–3294, 2011.

[15] M. Huang, X. Zhu, Y. Hao, D. G. Payan, K. Qu, and M. Li, "Discovering patterns to extract protein-protein interactions from full texts," *Bioinformatics*, vol. 20, no. 18, pp. 3604–3612, 2004.

[16] Y. Hao, X. Zhu, M. Huang, and M. Li, "Discovering patterns to extract protein-protein interactions from the literature: part II," *Bioinformatics*, vol. 21, no. 15, pp. 3294–3300, 2005.

[17] S. Kim, J. Yoon, and J. Yang, "Kernel approaches for genic interaction extraction," *Bioinformatics*, vol. 24, no. 1, pp. 118–126, 2008.

[18] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi, "Automated extraction of information on protein-protein interactions from the biological literature," *Bioinformatics*, vol. 17, no. 2, pp. 155–161, 2001.

[19] K. Fundel, R. Küffner, and R. Zimmer, "RelEx—relation extraction using dependency parse trees," *Bioinformatics*, vol. 23, no. 3, pp. 365–371, 2007.

[20] J. Šarić, L. J. Jensen, R. Ouzounova, I. Rojas, and P. Bork, "Extraction of regulatory gene/protein networks from Medline," *Bioinformatics*, vol. 22, no. 6, pp. 645–650, 2006.

[21] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles," *Bioinformatics*, vol. 17, no. 1, pp. S74–S82, 2001.

[22] J. M. Temkin and M. R. Gilder, "Extraction of protein interaction information from unstructured text using a context-free grammar," *Bioinformatics*, vol. 19, no. 16, pp. 2046–2053, 2003.

[23] A. Skusa, A. Rüegg, and J. Köhler, "Extraction of biological interaction networks from scientific literature," *Briefings in Bioinformatics*, vol. 6, no. 3, pp. 263–276, 2005.

[24] R. Malik, L. Franke, and A. Siebes, "Combination of text-mining algorithms increases the performance," *Bioinformatics*, vol. 22, no. 17, pp. 2151–2157, 2006.

[25] R. Chowdhary, J. Zhang, and J. S. Liu, "Bayesian inference of protein–protein interactions from biological literature," *Bioinformatics*, vol. 25, no. 12, pp. 1536–1542, 2009.

[26] Q. Zou, J. Li, C. Wang, and X. Zeng, "Approaches for recognizing disease genes based on network," *BioMed Research International*, vol. 2014, Article ID 416323, 10 pages, 2014.

[27] Q.-C. Bui, B. T. Nualláin, C. A. Boucher, and P. M. A. Sloot, "Extracting causal relations on HIV drug resistance from literature," *BMC Bioinformatics*, vol. 11, article 101, 2010.

[28] Q. Jiang, Y. Wang, Y. Hao et al., "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, no. 1, pp. D98–D104, 2009.

[29] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, and D. S. Wishart, "PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites," *Nucleic Acids Research*, vol. 36, pp. W399–W405, 2008.

[30] I. Iossifov, R. Rodriguez-Esteban, I. Mayzus, K. J. Millen, and A. Rzhetsky, "Looking at cerebellar malformations through text-mined interactomes of mice and humans," *PLoS Computational Biology*, vol. 5, no. 11, Article ID e1000559, 2009.

[31] L. J. Jensen, J. Saric, and P. Bork, "Literature mining for the biologist: from information retrieval to biological discovery," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 119–129, 2006.

[32] H.-M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: an ontology-based information retrieval and extraction system for biological literature," *PLoS Biology*, vol. 2, no. 11, article e309, 2004.

[33] N. Uramoto, H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi, and K. Takeda, "A text-mining system for knowledge discovery from biomedical documents," *IBM Systems Journal*, vol. 43, no. 3, pp. 516–533, 2004.

[34] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the web," in *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 51, pp. 68–74, New York, NY, USA, 2007.

[35] M. Banko and O. Etzioni, "The tradeoffs between open and traditional relation extraction," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 28–36, Columbus, Ohio, USA, June 2008.

[36] M. Abulaish and L. Dey, "Biological relation extraction and query answering from MEDLINE abstracts using ontology-based text mining," *Data and Knowledge Engineering*, vol. 61, no. 2, pp. 228–262, 2007.

[37] D. Szklarczyk, A. Franceschini, S. Wyder et al., "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, no. 1, pp. D447–D452, 2015.

[38] A. Chatr-Aryamontri, B.-J. Breitkreutz, R. Oughtred et al., "The BioGRID interaction database: 2015 update," *Nucleic Acids Research*, vol. 43, no. 1, pp. D470–D478, 2015.

[39] C.-H. Wei, H.-Y. Kao, and Z. Lu, "PubTator: a web-based text mining tool for assisting biocuration," *Nucleic Acids Research*, vol. 41, no. 1, pp. W518–W522, 2013.

[40] M. Safran, I. Dalah, J. Alexander et al., "GeneCards version 3: the human gene integrator," *Database*, vol. 2010, Article ID baq020, 16 pages, 2010.

[41] C. C. Huang and Z. Lu, "Community challenges in biomedical text mining over 10 years: success, failure and the future," *Briefings in Bioinformatics*, 2015.

[42] R. Khare, B. M. Good, R. Leaman, A. I. Su, and Z. Lu, "Crowdsourcing in biomedicine: challenges and opportunities," *Briefings in Bioinformatics*, 2015.

[43] D. E. Oliver, G. Bhalotia, A. S. Schwartz, R. B. Altman, and M. A. Hearst, "Tools for loading MEDLINE into a local relational database," *BMC Bioinformatics*, vol. 5, article 146, 2004.

[44] J. Wang, I. Cetindil, S. Ji et al., "Interactive and fuzzy search: a dynamic way to explore MEDLINE," *Bioinformatics*, vol. 26, no. 18, Article ID btq414, pp. 2321–2327, 2010.

[45] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, "A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis," *BMC Bioinformatics*, vol. 9, article 510, 2008.

[46] B. Liu, J. Xu, Q. Zou, R. Xu, X. Wang, and Q. Chen, "Using distances between Top-n-gram and residue pairs for protein remote homology detection," *BMC Bioinformatics*, vol. 15, supplement 2, article S3, 2014.

[47] B. Liu, F. Liu, L. Fang, X. Wang, and K. Chou, "repDNA: a python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects," *Bioinformatics*, vol. 31, no. 8, pp. 1307–1309, 2015.

[48] B. Liu, D. Zhang, R. Xu et al., "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[49] P. Y. Chou and G. D. Fasman, "Empirical predictions of protein conformation," *Annual Review of Biochemistry*, vol. 47, pp. 251–276, 1978.

[50] J. Garnier, D. J. Osguthorpe, and B. Robson, "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins," *Journal of Molecular Biology*, vol. 120, no. 1, pp. 97–120, 1978.

[51] Q. Dong, X. Wang, L. Lin, and Y. Wang, "Analysis and prediction of protein local structure based on structure alphabets," *Proteins: Structure, Function and Genetics*, vol. 72, no. 1, pp. 163–172, 2008.

[52] Q. Dong, X. Wang, and L. Lin, "Prediction of protein local structures and folding fragments based on building-block library," *Proteins: Structure, Function and Genetics*, vol. 72, no. 1, pp. 353–366, 2008.

[53] B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy," *Journal of Molecular Biology*, vol. 232, no. 2, pp. 584–599, 1993.

[54] H. Ding, H. Lin, W. Chen et al., "Prediction of protein structural classes based on feature selection technique," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 6, no. 3, pp. 235–240, 2014.

[55] H. Lin, C. Ding, Q. Song et al., "The prediction of protein structural class using averaged chemical shifts," *Journal of Biomolecular Structure & Dynamics*, vol. 29, no. 6, pp. 643–649, 2012.

[56] C. Lin, Y. Zou, J. Qin et al., "Hierarchical classification of protein folds using a novel ensemble classifier," *PLoS ONE*, vol. 8, no. 2, Article ID e56499, 2013.

[57] W. Chen, X. Liu, Y. Huang, Y. Jiang, Q. Zou, and C. Lin, "Improved method for predicting protein fold patterns with ensemble classifiers," *Genetics and Molecular Research*, vol. 11, no. 1, pp. 174–181, 2012.

[58] X. Zhao, Q. Zou, B. Liu, and X. Liu, "Exploratory predicting protein folding model with random forest and hybrid features," *Current Proteomics*, vol. 11, no. 4, pp. 289–299, 2014.

[59] Y. Liu, J. Carbonell, J. Klein-Seetharaman, and V. Gopalakrishnan, "Comparison of probabilistic combination methods for protein secondary structure prediction," *Bioinformatics*, vol. 20, no. 17, pp. 3099–3107, 2004.

[60] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker, "Sequence complexity of disordered protein," *Proteins: Structure, Function and Genetics*, vol. 42, no. 1, pp. 38–48, 2001.

[61] C.-T. Su, C.-Y. Chen, and Y.-Y. Ou, "Protein disorder prediction by condensed PSSM considering propensity for order or disorder," *BMC Bioinformatics*, vol. 7, article 319, 2006.

[62] C.-T. Su, C.-Y. Chen, and C.-M. Hsu, "IPDA: integrated protein disorder analyzer," *Nucleic Acids Research*, vol. 35, no. 2, pp. W465–W472, 2007.

[63] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *Journal of Molecular Biology*, vol. 337, no. 3, pp. 635–645, 2004.

[64] K. Shimizu, S. Hirose, and T. Noguchi, "POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix," *Bioinformatics*, vol. 23, no. 17, pp. 2337–2338, 2007.

[65] S. Hirose, K. Shimizu, S. Kanai, Y. Kuroda, and T. Noguchi, "POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions," *Bioinformatics*, vol. 23, no. 16, pp. 2046–2053, 2007.

[66] L. Wang and U. H. Sauer, "OnD-CRF: predicting order and disorder in proteins conditional random fields," *Bioinformatics*, vol. 24, no. 11, pp. 1401–1402, 2008.

[67] P. Han, X. Zhang, R. S. Norton, and Z.-P. Feng, "Large-scale prediction of long disordered regions in proteins using random forests," *BMC Bioinformatics*, vol. 10, article 8, 2009.

[68] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, "Exploiting three kinds of interface propensities to identify protein binding sites," *Computational Biology and Chemistry*, vol. 33, no. 4, pp. 303–311, 2009.

[69] B. Liu, B. Liu, F. Liu, and X. Wang, "Protein binding site prediction by combining hidden markov support vector machine and profile-based propensities," *The Scientific World Journal*, vol. 2014, Article ID 464093, 6 pages, 2014.

[70] Z. Wang, Q. Zou, Y. Jiang, Y. Ju, and X. Zeng, "Review of protein subcellular localization prediction," *Current Bioinformatics*, vol. 9, no. 3, pp. 331–342, 2014.

[71] H. Lin, H. Ding, F.-B. Guo, A.-Y. Zhang, and J. Huang, "Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition," *Protein & Peptide Letters*, vol. 15, no. 7, pp. 739–744, 2008.

[72] H. Lin, H. Ding, F.-B. Guo, and J. Huang, "Prediction of subcellular location of mycobacterial protein using feature selection techniques," *Molecular Diversity*, vol. 14, no. 4, pp. 667–671, 2010.

[73] H. Lin, H. Wang, H. Ding, Y.-L. Chen, and Q.-Z. Li, "Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition," *Acta Biotheoretica*, vol. 57, no. 3, pp. 321–330, 2009.

[74] H. Lin, W. Chen, L.-F. Yuan, Z.-Q. Li, and H. Ding, "Using over-represented tetrapeptides to predict protein submitochondria locations," *Acta Biotheoretica*, vol. 61, no. 2, pp. 259–268, 2013.

[75] H. Ding, S.-H. Guo, E.-Z. Deng et al., "Prediction of Golgi-resident protein types by using feature selection technique," *Chemometrics and Intelligent Laboratory Systems*, vol. 124, pp. 9–13, 2013.

[76] H. Lin, C. Ding, L.-F. Yuan et al., "Predicting subchloroplast locations of proteins based on the general form of Chou's pseudo amino acid composition: approached from optimal tripeptide composition," *International Journal of Biomathematics*, vol. 6, no. 2, Article ID 1350003, 2013.

[77] P.-P. Zhu, W.-C. Li, Z.-J. Zhong et al., "Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition," *Molecular BioSystems*, vol. 11, no. 2, pp. 558–563, 2015.

[78] H. Ding, L. Liu, F.-B. Guo, J. Huang, and H. Lin, "Identify golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition," *Protein & Peptide Letters*, vol. 18, no. 1, pp. 58–63, 2011.

[79] Q. Zou, X. Li, Y. Jiang, Y. Zhao, and G. Wang, "BinMemPredict: a web server and software for predicting membrane protein types," *Current Proteomics*, vol. 10, no. 1, pp. 2–9, 2013.

[80] H. Lin, "The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 252, no. 2, pp. 350–356, 2008.

[81] C. Ding, L.-F. Yuan, S.-H. Guo, H. Lin, and W. Chen, "Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions," *Journal of Proteomics*, vol. 77, pp. 321–328, 2012.

[82] H. Lin and H. Ding, "Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 269, pp. 64–69, 2011.

[83] B. Liu, X. Wang, Q. Zou, Q. Dong, and Q. Chen, "Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation," *Molecular Informatics*, vol. 32, no. 9-10, pp. 775–782, 2013.

[84] B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan, "Using amino acid physicochemical distance transformation for fast protein remote homology detection," *PLoS ONE*, vol. 7, no. 9, Article ID e46633, 2012.

[85] G. Yu, H. Rangwala, C. Domeniconi, G. Zhang, and Z. Yu, "Protein function prediction with incomplete annotations," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 3, pp. 579–591, 2014.

[86] Q. Zou, Z. Wang, X. Guan, B. Liu, Y. Wu, and Z. Lin, "An approach for identifying cytokines based on a novel ensemble classifier," *BioMed Research International*, vol. 2013, Article ID 686090, 11 pages, 2013.

[87] G. Yu, H. Rangwala, C. Domeniconi, G. Zhang, and Z. Yu, "Protein function prediction using multi-label ensemble classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 4, pp. 1045–1057, 2013.

[88] H. Ding, E.-Z. Deng, L.-F. Yuan et al., "iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels," *BioMed Research International*, vol. 2014, Article ID 286419, 10 pages, 2014.

[89] W.-X. Liu, E.-Z. Deng, W. Chen, and H. Lin, "Identifying the subfamilies of voltage-gated potassium channels using feature selection technique," *International Journal of Molecular Sciences*, vol. 15, no. 7, pp. 12940–12951, 2014.

[90] H. Ding and D. Li, "Identification of mitochondrial proteins of malaria parasite using analysis of variance," *Amino Acids*, vol. 47, no. 2, pp. 329–333, 2015.

[91] H. Ding, P.-M. Feng, W. Chen, and H. Lin, "Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis," *Molecular BioSystems*, vol. 10, no. 8, pp. 2229–2235, 2014.

[92] L.-F. Yuan, C. Ding, S.-H. Guo, H. Ding, W. Chen, and H. Lin, "Prediction of the types of ion channel-targeted conotoxins based on radial basis function network," *Toxicology in Vitro*, vol. 27, no. 2, pp. 852–856, 2013.

[93] H. Lin and W. Chen, "Prediction of thermophilic proteins using feature selection technique," *Journal of Microbiological Methods*, vol. 84, no. 1, pp. 67–70, 2011.

[94] X.-Y. Cheng, W.-J. Huang, S.-C. Hu et al., "A global characterization and identification of multifunctional enzymes," *PLoS ONE*, vol. 7, no. 6, Article ID e38979, 2012.

[95] H. Lin, W. Chen, and H. Ding, "AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes," *PLoS ONE*, vol. 8, no. 10, Article ID e75726, 2013.

[96] Q. Zou, W. Chen, Y. Huang, X. Liu, and Y. Jiang, "Identifying multi-functional enzyme by hierarchical multi-label classifier," *Journal of Computational and Theoretical Nanoscience*, vol. 10, no. 4, pp. 1038–1043, 2013.

[97] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, and X. Wang, "PseDNA-Pro: DNA-binding protein identification by combining chou's PseAAC and Physicochemical distance transformation," *Molecular Informatics*, vol. 34, no. 1, pp. 8–17, 2015.

[98] B. Liu, J. Xu, X. Lan et al., "IDNA-Prot—dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition," *PLoS ONE*, vol. 9, no. 9, Article ID e106691, 2014.

[99] J. R. Bock and D. A. Gough, "Predicting protein-protein interactions from primary structure," *Bioinformatics*, vol. 17, no. 5, pp. 455–460, 2001.

[100] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics*, vol. 21, supplement 1, pp. i38–i46, 2005.

[101] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction," *Proteins: Structure, Function and Genetics*, vol. 63, no. 3, pp. 490–500, 2006.

[102] L. V. Zhang, S. L. Wong, O. D. King, and F. P. Roth, "Predicting co-complexed protein pairs using genomic and proteomic data integration," *BMC Bioinformatics*, vol. 5, article 38, 2004.

[103] S. J. Darnell, D. Page, and J. C. Mitchell, "An automated decision-tree approach to predicting protein interaction hot spots," *Proteins: Structure, Function, and Bioinformatics*, vol. 68, no. 4, pp. 813–823, 2007.

[104] X.-W. Chen and M. Liu, "Prediction of protein-protein interactions using random decision forest framework," *Bioinformatics*, vol. 21, no. 24, pp. 4394–4400, 2005.

[105] R. Jansen, H. Yu, D. Greenbaum et al., "A bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, pp. 449–453, 2003.

[106] S. M. Gomez, W. S. Noble, and A. Rzhetsky, "Learning to predict protein-protein interactions from protein sequences," *Bioinformatics*, vol. 19, no. 15, pp. 1875–1881, 2003.

[107] M.-H. Li, X.-L. Wang, L. Lin, and T. Liu, "Effect of example weights on prediction of protein-protein interactions," *Computational Biology and Chemistry*, vol. 30, no. 5, pp. 386–392, 2006.

[108] M.-H. Li, L. Lin, X.-L. Wang, and T. Liu, "Protein-protein interaction site prediction based on conditional random fields," *Bioinformatics*, vol. 23, no. 5, pp. 597–604, 2007.

[109] T. Friedrich, B. Pils, T. Dandekar, J. Schultz, and T. Müller, "Modelling interaction sites in protein domains with interaction profile hidden Markov models," *Bioinformatics*, vol. 22, no. 23, pp. 2851–2857, 2006.

[110] M. Šikić, S. Tomić, and K. Vlahoviček, "Prediction of protein-protein interaction sites in sequences and 3D structures by random forests," *PLoS Computational Biology*, vol. 5, no. 1, Article ID e1000278, 2009.

[111] J. R. Bradford and D. R. Westhead, "Improved prediction of protein-protein binding sites using a support vector machines approach," *Bioinformatics*, vol. 21, no. 8, pp. 1487–1494, 2005.

[112] P. Fariselli, F. Pazos, A. Valencia, and R. Casadio, "Prediction of protein-protein interaction sites in heterocomplexes with neural networks," *European Journal of Biochemistry*, vol. 269, no. 5, pp. 1356–1361, 2002.

[113] J. R. Bradford, C. J. Needham, A. J. Bulpitt, and D. R. Westhead, "Insights into protein-protein interfaces using a bayesian network prediction method," *Journal of Molecular Biology*, vol. 362, no. 2, pp. 365–386, 2006.

[114] I. Kufareva, L. Budagyan, E. Raush, M. Totrov, and R. Abagyan, "PIER: protein interface recognition for structural proteomics," *Proteins*, vol. 67, no. 2, pp. 400–417, 2007.

[115] J. Yu, M. Guo, C. J. Needham, Y. Huang, L. Cai, and D. R. Westhead, "Simple sequence-based kernels do not predict protein-protein interactions," *Bioinformatics*, vol. 26, no. 20, pp. 2610–2614, 2010.

[116] Q. Zou, T. Zhao, Y. Liu, and M. Guo, "Predicting RNA secondary structure based on the class information and Hopfield network," *Computers in Biology and Medicine*, vol. 39, no. 3, pp. 206–214, 2009.

[117] Q. Zou, C. Lin, X.-Y. Liu, Y.-P. Han, W.-B. Li, and M.-Z. Guo, "Novel representation of RNA secondary structure used to improve prediction algorithms," *Genetics and Molecular Research*, vol. 10, no. 3, pp. 1986–1998, 2011.

[118] B. Liu, L. Fang, F. Liu et al., "Identification of real microRNA precursors with a pseudo structure status composition approach," *PLoS ONE*, vol. 10, no. 3, Article ID e0121501, 2015.

[119] B. Liu, L. Fang, J. Chen, F. Liu, and X. Wang, "miRNA-dis: microRNA precursor identification based on distance structure status pairs," *Molecular BioSystems*, vol. 11, no. 4, pp. 1194–1204, 2015.

[120] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen, "Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering," *PLoS Computational Biology*, vol. 3, no. 4, article e65, 2007.

[121] S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, and R. Backofen, "LocARNA-P: accurate boundary prediction and improved detection of structural RNAs," *RNA*, vol. 18, no. 5, pp. 900–914, 2012.

[122] C. Wang, L. Wei, M. Guo, and Q. Zou, "Computational approaches in detecting non-coding RNA," *Current Genomics*, vol. 14, no. 6, pp. 371–377, 2013.

[123] Q. Zou, X.-B. Li, W.-R. Jiang, Z.-Y. Lin, G.-L. Li, and K. Chen, "Survey of MapReduce frame operation in bioinformatics," *Briefings in Bioinformatics*, vol. 15, no. 4, Article ID bbs088, pp. 637–647, 2014.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 178572, 11 pages
http://dx.doi.org/10.1155/2015/178572

Research Article

A Systematic Evaluation of Feature Selection and Classification Algorithms Using Simulated and Real miRNA Sequencing Data

Sheng Yang, Li Guo, Fang Shao, Yang Zhao, and Feng Chen

Department of Biostatistics, School of Public Health, Nanjing Medical University, 101 Longmian Road, Nanjing, Jiangsu 211166, China

Correspondence should be addressed to Feng Chen; fengchen@njmu.edu.cn

Received 10 June 2015; Accepted 25 August 2015

Academic Editor: Lin Lu

Sequencing is widely used to discover associations between microRNAs (miRNAs) and diseases. However, the negative binomial distribution (NB) and high dimensionality of data obtained using sequencing can lead to low-power results and low reproducibility. Several statistical learning algorithms have been proposed to address sequencing data, and although evaluation of these methods is essential, such studies are relatively rare. The performance of seven feature selection (FS) algorithms, including baySeq, DESeq, edgeR, the rank sum test, lasso, particle swarm optimistic decision tree, and random forest (RF), was compared by simulation under different conditions based on the difference of the mean, the dispersion parameter of the NB, and the signal to noise ratio. Real data were used to evaluate the performance of RF, logistic regression, and support vector machine. Based on the simulation and real data, we discuss the behaviour of the FS and classification algorithms. The Apriori algorithm identified frequent item sets (mir-133a, mir-133b, mir-183, mir-937, and mir-96) from among the deregulated miRNAs of six datasets from The Cancer Genomics Atlas. Taking these findings altogether and considering computational memory requirements, we propose a strategy that combines edgeR and DESeq for large sample sizes.

1. Introduction

MicroRNAs (miRNAs) are small, endogenous, and noncoding RNAs that trigger messenger RNA (mRNA) deregulation and translational repression by binding the $3'$ untranslated region ($3'$UTR) of these targets [1]. Depending on their biological function and stability, miRNAs are also regarded as biomarkers to distinguish cases and controls [2, 3]. Therefore, emerging technologies, such as cDNA microarrays, high-density oligonucleotide chips, and next-generation sequencing (NGS), have been highly useful in the discovery of miRNAs that cause or prevent disease [4]. cDNA microarrays and high-density oligonucleotide chips are only capable of providing relative expression levels, whereas NGS can be used to count the exact number of reads and obtain sequence information (arm switching and isomiRs) [5].

To process high-dimensional NGS data and gain deep insight into biological processes, statistical learning methods are emerging with the goal of classifying labels by selecting a subset of features, minimizing the coefficients of features or reducing their dimension [6, 7]. Using a negative binomial distribution (NB) assumption, edgeR, DESeq, and baySeq are three important filter algorithms for selecting significant variables by intrinsic characteristics [8–10]. Wrapper algorithms based on classification apply a search strategy in the feature space, including sequential forward searching (SFS) and sequential forward floating searching (SFFS); however, the computational intensity of this approach is large [11]. Hybrids of feature selection and classification, known as embedded methods, such as random forest (RF), regard the classification model as an internal parameter and reduce the computational requirements [12]. Furthermore, independent of the distribution, shrinkage tricks, such as lasso, also play an important role in high-dimensional NGS [13].

Recently, an evaluation of statistical and machine learning algorithms for NGS data has become essential. This evaluation can be achieved from three perspectives: (i) comparing the performance of seven popular feature selection

Table 1: Parameter settings used for the simulation data.

Scenario	Parameter	Settings
A1–A5	Signal to noise (s2n)	0.01, 0.05, 0.1, 0.15, and 0.20
B1–B5	Mean of significant variables in the case	10, 15, 20, 25, and 30
C1–C5	Dispersion parameter of significant variables in the case	0.125, 0.5, 1, 2, and 8
	Sample size (+/−)	40 (20/20)
	Number of variables	500
	Mean of significant variables in the control	5
	Mean of insignificant variables	5
	Dispersion parameter of significant variables in the control	1
	Dispersion parameter of insignificant variables	1

algorithms in the context of simulation, using sensitivity and specificity; (ii) studying the properties of three classification algorithms, logistic regression, support vector machine (SVM), and RF, in the context of differentially expressed (DE) miRNAs from The Cancer Genomics Atlas (TCGA) data to gain deeper insight into the combination of FS and classification; and (iii) analysing the similarity of six cancers based on miRNAs and the corresponding pathways.

2. Methods

2.1. Simulations. First, we assumed that the distribution of NGS data was NB, corresponding to the parameters, mean, dispersion parameter (DP) of NB, and ratio of signal to noise (s2n) in the simulations. The inflating extent of the data is directly proportional to the DP, and s2n is the ratio of significant variables to insignificant variables. The second assumption was that all significant variables are causal, which indicates the means of case groups were larger than those of control groups.

Based on these two basic assumptions, three different settings were involved: s2n ranged from 0.01 to 0.2 (A1–A5), the means of the significant variables in the case group ranged from 10 to 30 by 5 (B1–B5), and the DP of the significant variables in the case group ranged from 0.125 to 8 (C1–C5). A total of 1,000 replications were produced to obtain a robust result. The parameter settings for the insignificant and significant variables were the same and fixed in all situations. When one parameter was studied, the others settings remained fixed. Details regarding the parameter settings are presented in Table 1.

2.2. Overviews of FS Algorithms and Their Evaluation Indexes. We compared seven different algorithms in the simulations, including three algorithms specific to NGS data (DESeq, edgeR, and baySeq), the Wilcoxon rank sum test, lasso, particle swarm optimal algorithm empowered by decision tree (PSODT), and RF. Each algorithm included different types of feature selection. The first five methods are filter methods because they select variables based on the order of the statistic or coefficient. PSODT, a wrapper algorithm, searches the subset of variables by PSO and evaluates the classification performance by DT. RF combines classification and feature selection. The Bioconductor packages *baySeq*, *DESeq2*, and *edgeR* were used, and lasso and RF were completed by the *glmnet* and *randomForest* packages in the R (version 3.0.3) framework, respectively.

DESeq and edgeR are two essential algorithms for feature selection in NGS data and are based on the NB distribution assumption. However, they use different methods for estimating the parameters. DESeq estimates the DP based on pooled data, which can normalize confounders from different library sizes. Local regression is then used to estimate the function of per-variable raw variance, a component of variance. edgeR algorithm defines the weighted conditional log-likelihood, which is a combination of common and individual likelihood, to estimate the parameter and uses α to weigh the importance of the common part. Exact testing is used by these two methods [14]. For baySeq, the difference between 1 and the posterior probability is considered as the P value. The *cv.glmnet* function estimates the penalty weight in lasso by cross-validation. We used the same parameter settings as Chen et al. for PSODT [11]. The score of each variable was identified as the time of *gbest* equal to *pbest*. For RF, we used the default setting, that is, the number of trees (*ntree*) = 500 and the number of random variables in each split (*mtry*) = $\sqrt{m}$, where m is the total number of variables.

In the simulations, type I errors and power were used to evaluate the performance of the four statistical algorithms (DESeq, edgeR, baySeq, and rank sum test) because they are based on hypothesis testing. Type I error and power correspond to the frequency of P values of noise and significant variables less than 0.05 or Bonferroni correction levels in 1,000 replications, respectively. As these procedures involved four machine learning methods, sensitivity and specificity were used to compare the entire techniques. These values were calculated according to

$$\begin{aligned} \text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FB}}, \end{aligned} \tag{1}$$

where TP, TN, FP, and FN are the means of the number of true cases, true controls, false cases, and false controls in 1,000 replications, respectively.

Table 2: Summary of the selected datasets.

Number	Cancer	Feature	Sample (+/−)	SDR[a]
1	BRCA	903	206 (103/103)	0.23
2	HNSC	906	162 (81/81)	0.18
3	KICH	796	82 (41/41)	0.10
4	LUAD	895	218 (109/109)	0.24
5	STAD	857	170 (85/85)	0.20
6	THCA	904	212 (106/106)	0.23

[a]SDR refers to the ratio between the number of samples and the number of features.

2.3. Real Data. For TCGA, six different cancer sequencing datasets (with features and samples) were involved, including breast invasive carcinoma (BRCA), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), lung adenocarcinoma (LUAD), stomach adenocarcinoma (STAD), and thyroid carcinoma (THCA). We only selected the matched samples. The low expression miRNAs whose sum expression levels in all samples were less than 10 were excluded (Table 2).

2.4. Landscape of Classification Algorithms and Indexes. Classification algorithms, including logistic regression, RF, and SVM, were regarded as another essential point because they indicate the predictive performance of the selected biomarkers. Logistic regression, a type of generalized linear model (GLM), was widely applied in case-control study, as its exponential coefficient, odds ratio (OR), directly elucidated the risk of variables. Based on the theory of Lagrange duality and kernel function, SVM solved dual problems rather than the minimum primary problem and mapped the variables to a higher dimension. Therefore, the nonlinear classified samples were discriminated using hyperplane. The following equation shows the standard form of this method:

$$y = h_{w,b}(x) = w^T x + b. \quad (2)$$

We chose the default settings of the *svm* function, which was a Gaussian kernel, and set the hyperparameter to $\gamma = 3$ and error term to $\varepsilon = 0.2$.

Random fivefold cross-validation was applied to real data to estimate the performance of the classification algorithms. This cross-validation meant that four-fifths of samples were used to construct the model and select the features, and the residual was used to test the validation; this process was replicated 100 times. The area under the ROC curve (AUC), positive predictive value (PPV), and negative predictive value (NPV) evaluated the classification performance of the featured subsets.

2.5. Apriori for Detecting the Frequent Item Set of miRNAs from Different Datasets. Apriori defines the frequency of item sets based on three indexes, including *support*, *confidence*, and *lift*. The *support* of an item set is defined as the percentage of the dataset that contained it. The *confidence* represents the association of the rule like $\{A\} \rightarrow \{B\}$, which is calculated by the conditional probability of $P(B \mid A)$. The *lift*, the ratio of $P(B \mid A)/P(B)$, is the quotient of the posterior and the prior confidence of an association rule. The first two standards can select the frequent item set.

The frequent miRNA sets were defined from the DE miRNAs in the six datasets by the following criteria: (a) the miRNAs satisfied the Bonferroni correction; (b) the miRNAs were selected more than or equal to 80 times in one algorithm; and (c) the miRNAs were defined by at least 3 algorithms. The frequent DE miRNA was then identified as having *support* and *confidence* values larger than or equal to 0.5. Finally, their targets were predicted twice from three datasets (TargetScan, miRanda, and miRTarBase), and enrichment analysis defined the deregulated pathways by Gene Ontology (GO) [15–17].

3. Results

3.1. The Evaluation of FS Algorithms Using a Simulation

3.1.1. Empirical Type I Error and Power of Four Statistical Algorithms. The type I error and power results are shown in Figures 1 and 2. baySeq, DESeq, and the rank sum test appeared to control type I error at a significance level of 0.05, although the rank sum test failed after Bonferroni correction. The type I error of edgeR was slightly inflated. s2n appeared to have no relationship to the power, whereas the mean and DP influenced the power. Based on the difference between the increasing mean or decreasing DP, the power of all of the algorithms increased. In particular, a decreasing trend in the rank sum test was observed with increasing DP because it included little consideration of the dispersion of the variables. However, the power of the three sequencing methods was high, especially for baySeq.

3.1.2. Sensitivity and Specificity with Different Settings of Three Parameters. The results from the simulation using scenarios A1–A5, B1–B5, and C1–C5, including the variable frequency, sensitivity, and specificity in different situations, are presented in Table 3 and Figure 3. First, DP influenced the two indexes of the machine learning algorithms and rank sum, although it had only a small influence on the performance of three sequencing methods. The sensitivities of edgeR and DESeq were larger than that of baySeq, although the extent of the increase and decrease of their sensitivity was larger. With increasing dispersion, the sensitivities of the rank sum and lasso methods were approximately zero. Second, when the difference between the means of the case and the control samples increased from 5 to 25, the sensitivity increased to different extents. For the three sequencing methods and the rank sum test, the index showed an obvious increase, and the frequency of selected significant variables was higher. The sensitivity and specificity of PSODT and RF showed little change with the change of mean. Third, greater s2n values led to increased sensitivity for the frequency of significant variables of baySeq and RF but appeared to have no relationship with the residuals.

For the seven algorithms, we obtained the following outcomes. The sensitivity of baySeq appeared to be lower than the other sequencing methods. The variations of the power of DESeq and edgeR were relatively similar, although the latter was not control type I error. Lasso also strictly controlled

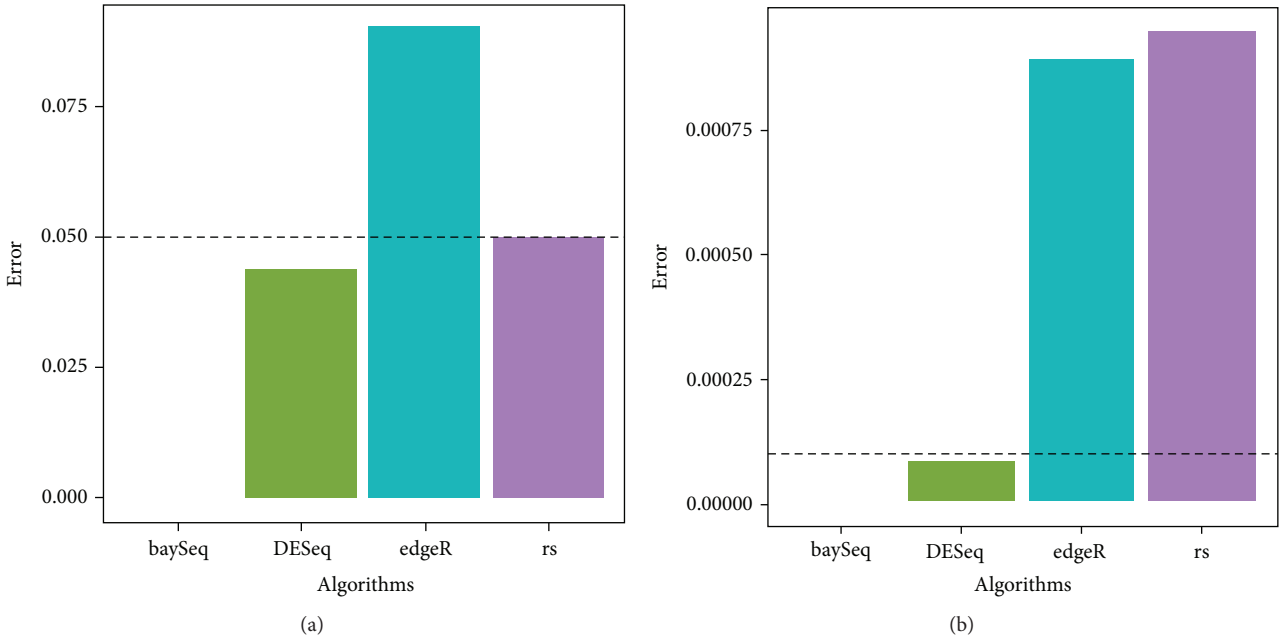


Figure 1: Type I error of four statistical algorithms. (a) $\alpha = 0.05$ condition. (b) Bonferroni correction.

type I errors, although its power was lower than that of the other methods in multiple situations. The rank sum test, a nonparametric method, was also influenced by the three parameters and is perhaps not suitable for sequencing data. In particular, when DP increased from 0.125 to 8, its sensitivity decreased from 1.00 to 0.23. The sensitivity of PSODT was highly stable when the parameters changed. The sensitivity of RF was only related to the s2n factor.

3.2. The FS and Classification Methods in Real Data. The number of significant miRNAs identified by different FS algorithms and the relationships between them are shown in Figure 4 and Additional File 1 available online at http://dx.doi.org/10.1155/2015/178572. Based on the frequency bar plots and Venn diagrams of each dataset, these results are clear. First, baySeq, edgeR, and the rank sum test selected the highest number of miRNAs in different datasets. For example, in KICH, the rank sum test selected 87 significant miRNAs, which was the greatest number of significant miRNAs identified by the six algorithms. Second, the three sequencing methods and the rank sum test had more intersections. However, PSODT rarely identified the same significant miRNAs during cross-validation, and intersections were also rare.

As shown in Table 4, the results of classification algorithms were as follows. First, RF and SVM performed better than logistic regression. For example, based on the results from edgeR in the KICH, the ROC of logistic regression was 0.39, which was lower than that of RF and SVM. Interestingly, logistic regression performed best using the variables selected by lasso, perhaps because the ratio between the number of variables and the number of samples was unsuitable for logistic regression, with the exception of lasso. Second, although the power of PSODT was lowest among the seven FS algorithms, the classification performance was not the worst. For example, in BRCA, the classification of the variables selected form PSODT was better than that of the rank sum test.

3.3. Run Time. The run time of the seven algorithms is shown in Additional File 2. In the simulations, baySeq required approximately 2 hours, which is longer than the other methods. However, different results were observed using real data. The time of DESeq sharply increased with larger sample sizes; however, the variations of other methods were not obvious with increasing sample sizes. Thus, baySeq consumed the greatest computational resources, and the resource consumption of DESeq in particular was largely determined by the sample size.

3.4. The Frequency miRNA Sets in Six Cancers and Enrichment Analysis. For the DE miRNAs in each cancer set, Apriori selected the frequency item sets that might be co-DE miRNAs in cancers (Additional File 3). mir-133a-1, mir-133b, mir-183, mir-937, and mir-96 were frequently identified DE miRNAs in six cancers. Some miRNAs were deregulated at the same time; for example, the *confidence* of mir-96 to mir-133a-1 was 1, and the *lift* was equal to 2. Furthermore, the enrichment pathways of their cotargets were also identified using GO (Additional Files 4 and 5).

4. Discussion

Using simulations and real data, we compared the performance of seven feature selection algorithms and three classification algorithms. Simulations identified the differing performances of the seven FS methods: baySeq, DESeq,

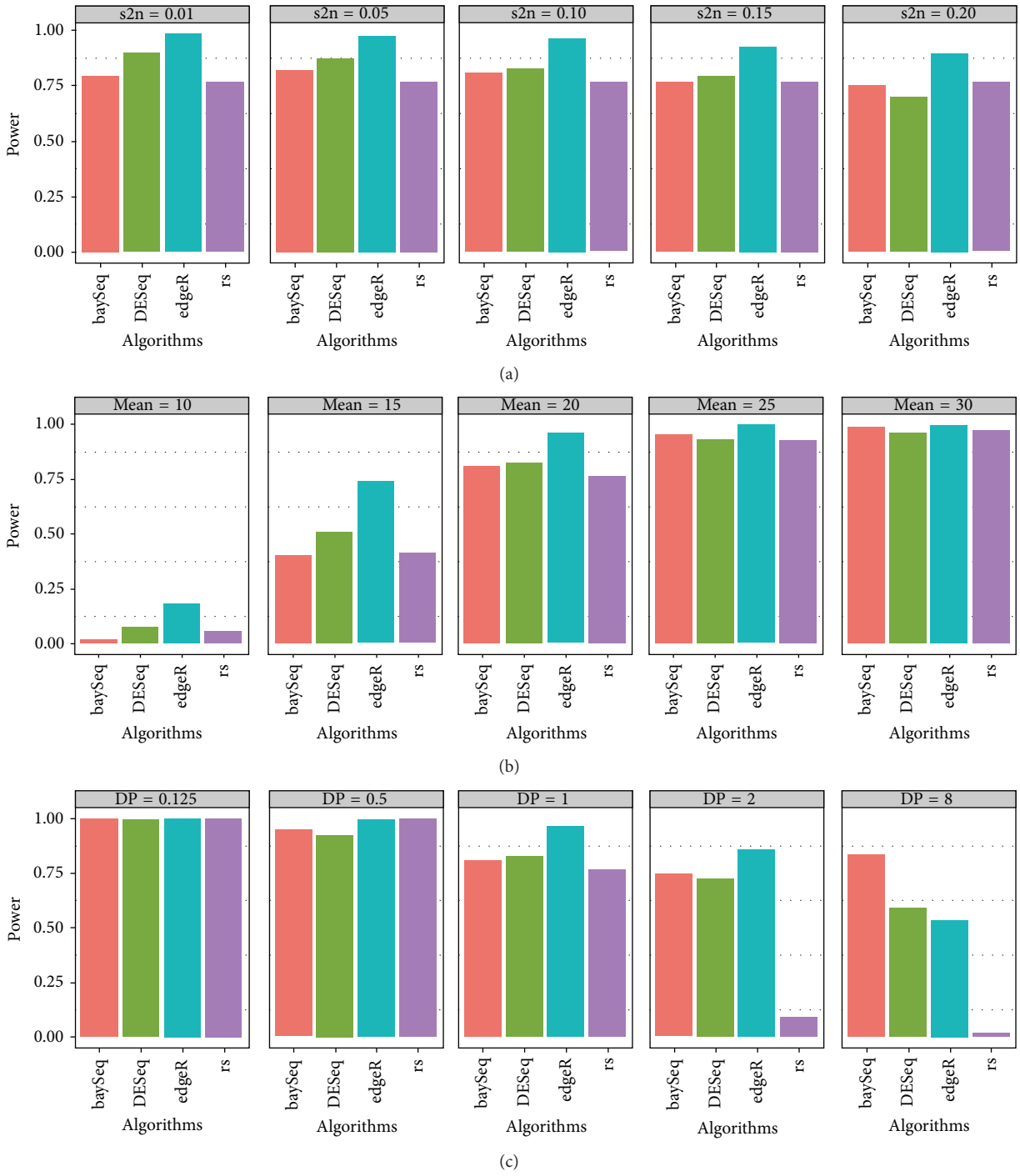
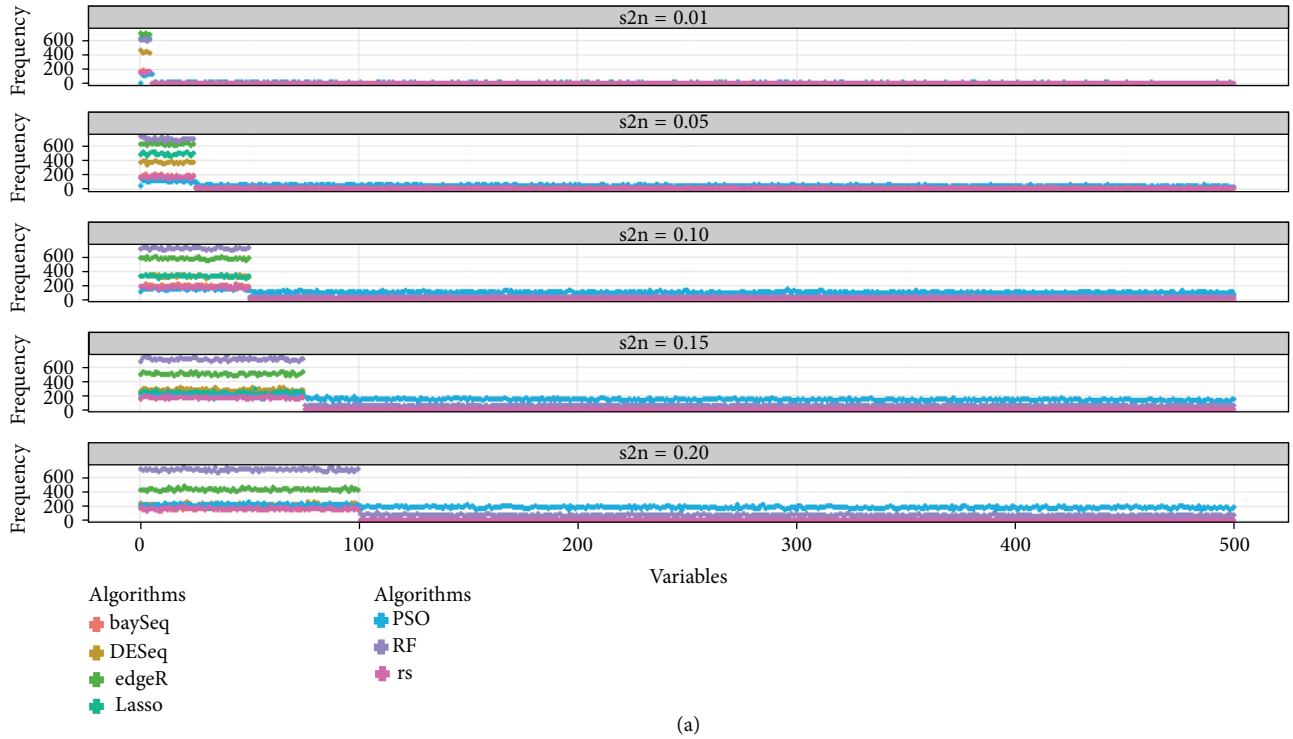
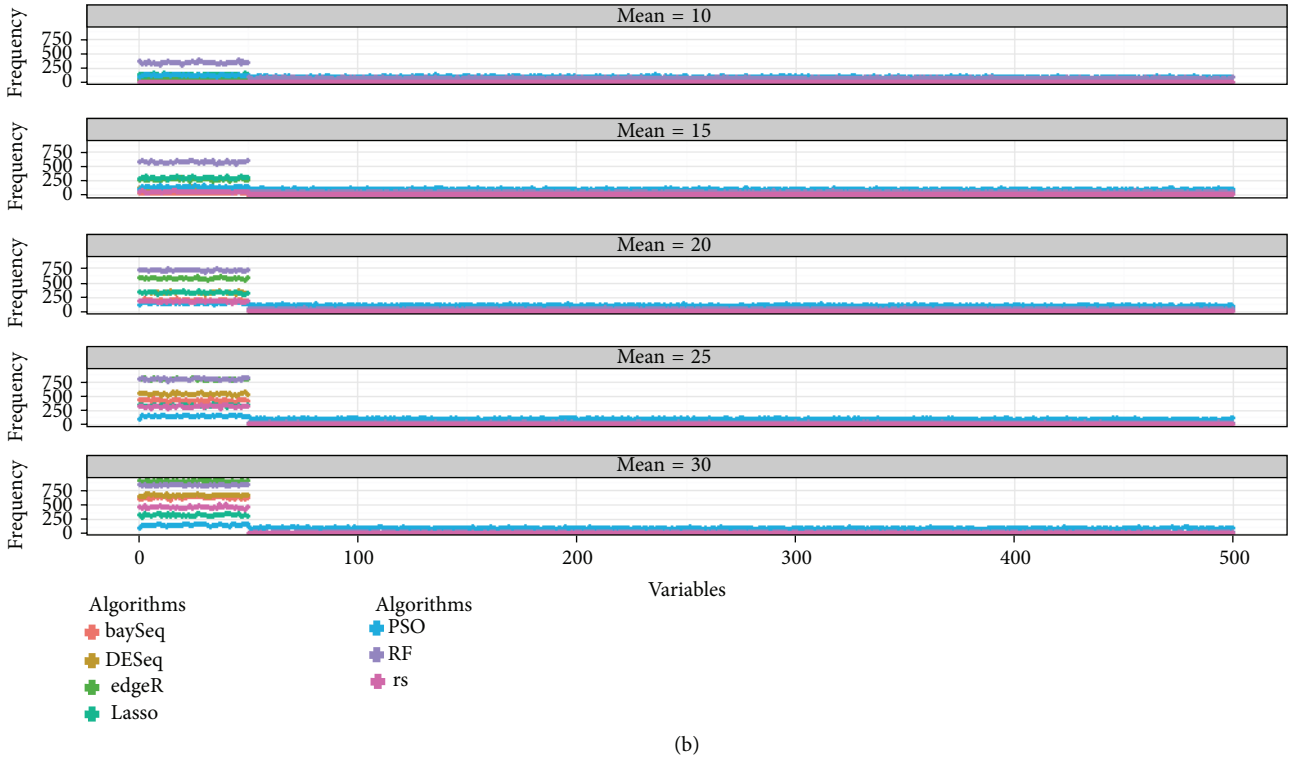


Figure 2: Power of four statistical algorithms with different settings of three parameters. (a) Different settings of the s2n of the variables. (b) Different settings of the mean. (c) Different settings of DP in the case group.

edgeR, the rank sum test, lasso, PSODT, and RF. In the comparisons of four statistical methods, we observed the following: (a) a larger DP may lead to a low power in the rank sum test due to a failure to estimate DP; (b) when the difference of the mean is greater than 15, the power of the sequencing methods is robust; (c) with increasing DP, there is a small decrease in the power of the sequencing methods, especially for baySeq. Regarding the sensitivity and specificity, the following conclusions were reached: (a) s2n influences the performance of baySeq and RF; (b) an increase in the difference of means causes increased sensitivity; and (c) increasing DP has little effect on the three sequencing algorithms but decreases the sensitivity of the others. Furthermore, real data showed that (a) logistic regression is unsuitable for the high dimension and small sample data and (b) the performance of RF is better than that of SVM.



(a)



(b)

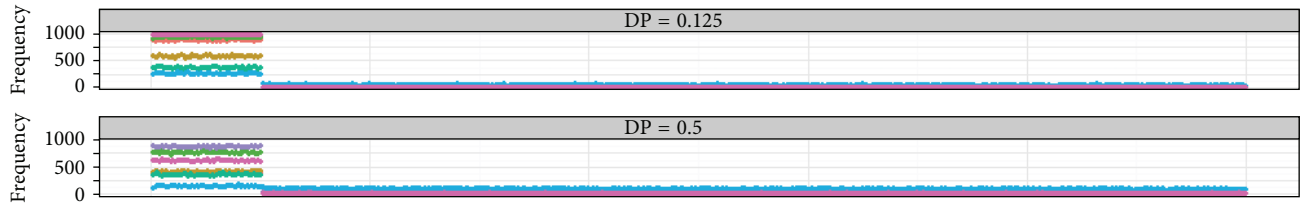
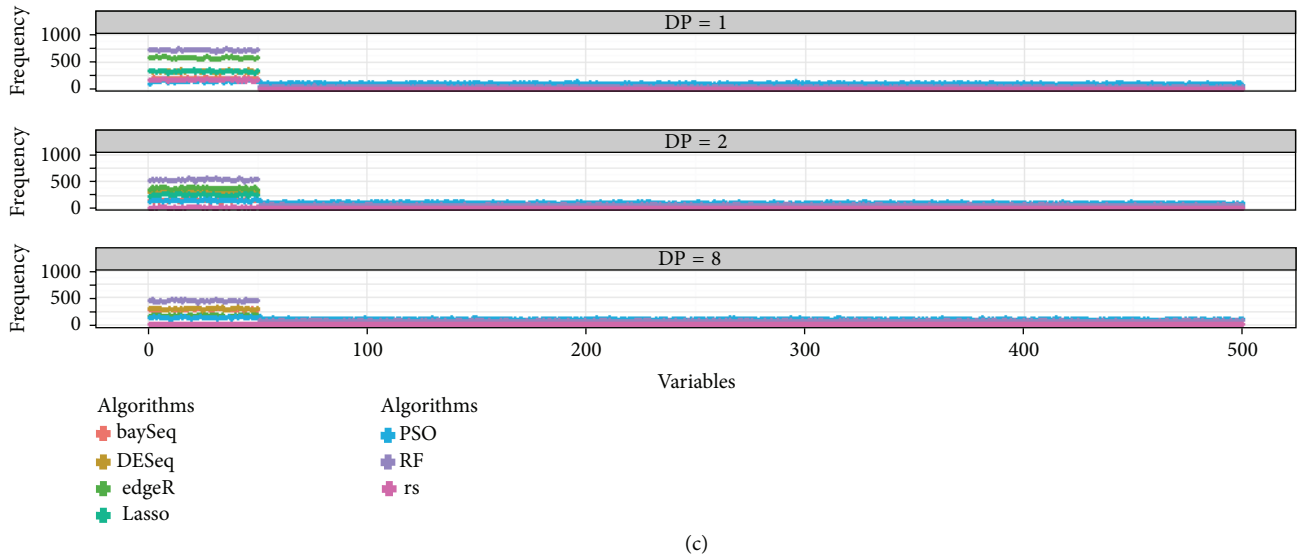


Figure 3: Continued.



(c)

Figure 3: The frequency of selected variables of seven FS methods in the simulation. (a) Different settings of the s2n of the variables. (b) Different settings of the mean. (c) Different settings of DP.

Table 3: Sensitivity and specificity of the seven algorithms in different settings[a].

Scenario	baySeq		DESeq		edgeR		Lasso	Rank sum		PSODT	RF
	$P = 0.05$	Bon[b]	$P = 0.05$	Bon[b]	$P = 0.05$	Bon[b]		$P = 0.05$	Bon[b]		
						s2n					
A1	0.68/1.00	0.16/1.00	0.97/0.96	0.43/1.00	0.99/0.93	0.68/1.00	0.62/0.99	0.92/0.95	0.16/1.00	0.10/0.99	0.60/1.00
A2	0.74/1.00	0.18/1.00	0.97/0.96	0.37/1.00	0.98/0.92	0.63/1.00	0.49/0.99	0.92/0.95	0.16/1.00	0.11/0.95	0.70/0.98
A3	0.77/1.00	0.19/1.00	0.95/0.94	0.32/1.00	0.98/0.91	0.57/1.00	0.32/1.00	0.92/0.95	0.16/1.00	0.14/0.90	0.71/0.97
A4	0.77/1.00	0.17/1.00	0.93/0.93	0.27/1.00	0.96/0.89	0.50/1.00	0.23/1.00	0.92/0.95	0.16/1.00	0.18/0.86	0.70/0.95
A5	0.76/1.00	0.16/1.00	0.90/0.90	0.22/1.00	0.95/0.86	0.43/1.00	0.18/1.00	0.18/1.00	0.92/0.95	0.17/1.00	0.70/0.93
						Mean of significant variables					
B1	0.05/1.00	0.00/1.00	0.41/0.95	0.01/1.00	0.52/0.92	0.04/1.00	0.13/0.99	0.40/0.95	0.01/1.00	0.12/0.90	0.33/0.93
B2	0.45/1.00	0.03/1.00	0.81/0.95	0.12/1.00	0.88/0.92	0.27/1.00	0.30/1.00	0.77/0.95	0.05/1.00	0.13/0.90	0.57/0.95
B4	0.91/1.00	0.41/1.00	0.99/0.94	0.53/1.00	0.99/0.91	0.78/1.00	0.32/1.00	0.97/0.95	0.31/1.00	0.14/0.91	0.79/0.98
B5	0.96/1.00	0.62/1.00	1.00/0.94	0.66/1.00	1.00/0.90	0.90/1.00	0.32/1.00	0.99/0.95	0.45/1.00	0.14/0.91	0.83/0.98
						Dispersion parameter of significant variables					
C1	1.00/1.00	0.87/1.00	1.00/0.92	0.57/1.00	1.00/0.89	0.92/1.00	0.37/1.00	1.00/0.95	0.97/1.00	0.26/0.95	0.97/1.00
C2	0.89/0.99	0.38/0.94	0.98/1.00	0.40/0.94	0.99/1.00	0.75/0.97	0.35/0.93	1.00/1.00	0.61/0.96	0.14/0.90	0.86/0.90
C4	0.71/1.00	0.14/1.00	0.90/0.95	0.29/1.00	0.92/0.92	0.36/1.00	0.24/0.99	0.46/0.95	0.01/1.00	0.14/0.90	0.52/0.95
C5	0.73/1.00	0.28/1.00	0.73/0.96	0.29/1.00	0.71/0.93	0.16/1.00	0.00/1.00	0.23/0.95	0.00/1.00	0.13/0.90	0.44/0.94

[a]The conditions where the mean = 20, dispersion parameter = 1, and s2n = 0.1 are the same. Each cell includes the sensitivity and specificity.
[b]Bon indicates a result using the Bonferroni correction.

Moreover, seven algorithms were evaluated using different conditions. edgeR was found to be suitable for large sample sizes because of low calculation time, although its type I error increases slightly. The type I error and power indicate that the performance of baySeq is perhaps best for selecting significant genes, although a large sample size may require a long computation time [18]. Similar to baySeq, DESeq requires more time with increasing sample size, although its advantage is that it can analyse data using only one replicate in each treatment group (Figure 1 and Additional File 2) [10]. The selection of the three algorithms is determined by the experimental design [18]. The rank sum test can be fit to any distribution assumption, but it fails to select the variables in NB, especially with increasing DP. The penalty of lasso is possibly too large because few significant variables are selected. PSODT rarely chooses the significant variables and has no association with the three factors because it defines a combination of variables having the best performance of DT. Considering the power, type I error, and calculation cost, an FS selection process can consist of two or more processes: (a) primary selection, which requires fast and high-power algorithms, and (b) further selection, which

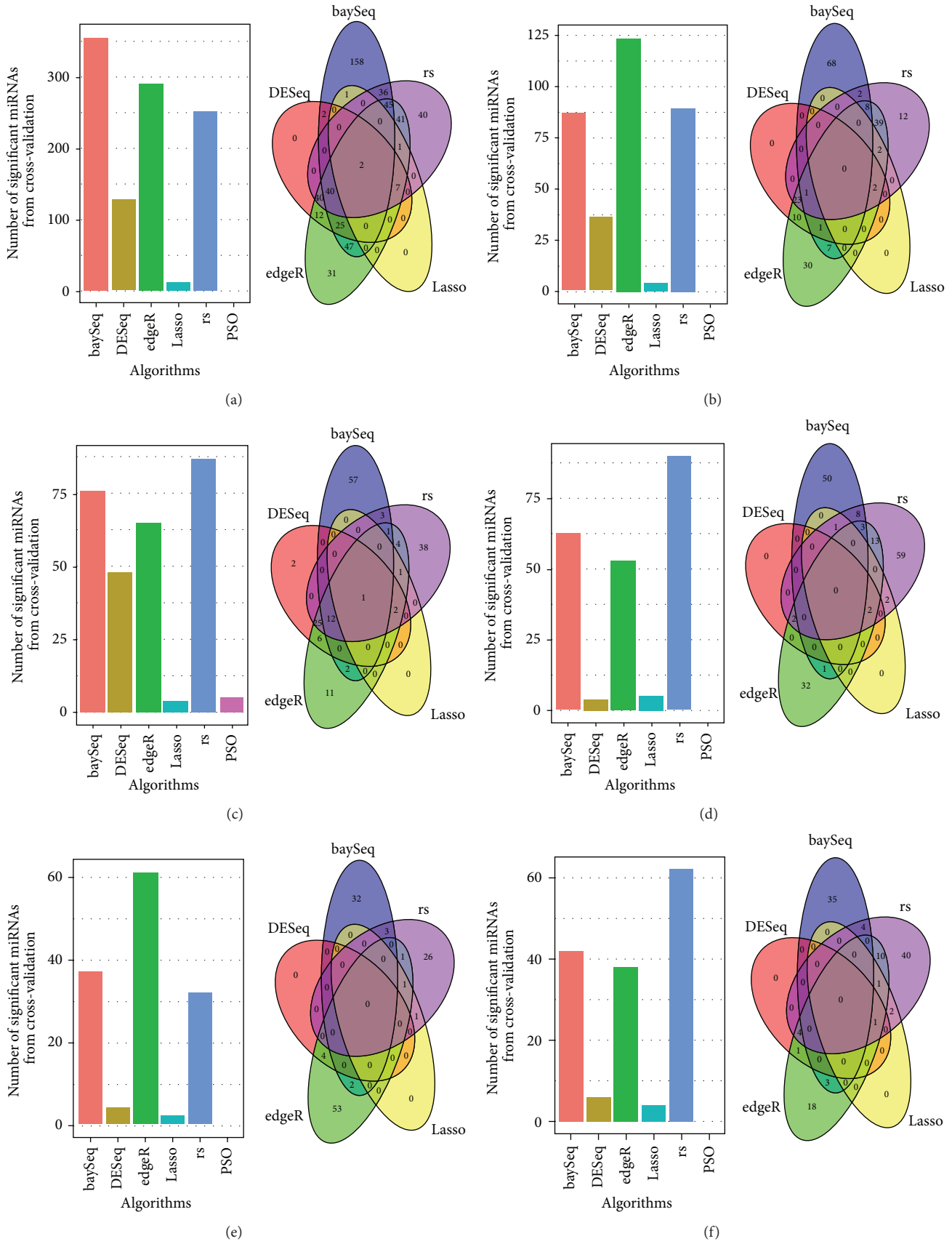


FIGURE 4: The bar plots and Venn diagrams of a number of significant miRNAs identified by different FS algorithms in six cancers. The bar plot indicates the number of significant variables. The Venn diagram illustrates the relationships of the significant variables among the six methods. (a) BRCA; (b) HNSC; (c) KICH; (d) LUAD; (e) STAD; and (f) THCA.

Table 4: Summary of three classification methods using real data.

Datasets	FS	Logistic regression			RF			SVM		
		PPV	NPV	AUC	PPV	NPV	AUC	PPV	NPV	AUC
BRCA	baySeq	0.53	0.53	0.53	1.00	0.99	0.99	0.95	0.96	0.96
	DESeq	0.70	0.72	0.70	1.00	0.99	1.00	1.00	0.94	0.97
	edgeR	0.54	0.55	0.55	1.00	0.99	0.99	0.54	0.55	0.55
	Lasso	0.97	0.98	0.98	1.00	0.99	0.99	0.97	0.98	0.98
	Rank sum	0.55	0.55	0.55	1.00	0.99	0.99	0.55	0.55	0.55
	PSODT	0.85	0.86	0.86	0.99	0.98	0.98	0.85	0.86	0.86
HNSC	baySeq	0.35	0.38	0.37	0.54	0.56	0.55	0.63	0.52	0.58
	DESeq	0.52	0.57	0.55	0.53	0.52	0.52	0.91	0.47	0.69
	edgeR	0.32	0.35	0.33	0.54	0.54	0.54	0.32	0.35	0.33
	Lasso	0.52	0.76	0.64	0.55	0.55	0.55	0.52	0.76	0.64
	Rank sum	0.35	0.31	0.33	0.54	0.54	0.54	0.35	0.31	0.33
	PSODT	0.43	0.44	0.43	0.55	0.54	0.54	0.43	0.44	0.43
KICH	baySeq	0.36	0.38	0.37	0.65	0.66	0.66	0.68	0.70	0.69
	DESeq	0.37	0.39	0.38	0.66	0.65	0.66	0.68	0.84	0.76
	edgeR	0.40	0.38	0.39	0.66	0.65	0.66	0.40	0.38	0.39
	Lasso	0.64	0.82	0.73	0.65	0.66	0.65	0.64	0.82	0.73
	Rank sum	0.39	0.38	0.39	0.66	0.66	0.66	0.39	0.38	0.39
	PSODT	0.37	0.38	0.37	0.66	0.65	0.66	0.37	0.38	0.37
LUAD	baySeq	0.40	0.47	0.43	0.46	0.45	0.46	0.45	0.69	0.57
	DESeq	0.30	0.78	0.54	0.46	0.41	0.44	0.95	0.36	0.65
	edgeR	0.44	0.47	0.46	0.47	0.45	0.46	0.44	0.47	0.46
	Lasso	0.47	0.74	0.61	0.47	0.45	0.46	0.47	0.74	0.61
	Rank sum	0.30	0.36	0.33	0.47	0.45	0.46	0.30	0.36	0.33
	PSODT	0.36	0.50	0.43	0.47	0.45	0.46	0.36	0.50	0.43
STAD	baySeq	0.42	0.56	0.49	0.44	0.45	0.44	0.44	0.63	0.54
	DESeq	0.14	0.85	0.49	0.41	0.38	0.40	0.91	0.25	0.58
	edgeR	0.37	0.42	0.40	0.49	0.46	0.47	0.37	0.42	0.40
	Lasso	0.43	0.77	0.60	0.46	0.46	0.46	0.43	0.77	0.60
	Rank sum	0.40	0.48	0.44	0.44	0.46	0.45	0.40	0.48	0.44
	PSODT	0.36	0.44	0.44	0.44	0.46	0.45	0.36	0.44	0.40
THCA	baySeq	0.49	0.63	0.56	0.56	0.57	0.57	0.77	0.50	0.63
	DESeq	0.49	0.85	0.67	0.54	0.58	0.56	0.54	0.82	0.68
	edgeR	0.53	0.59	0.56	0.56	0.60	0.58	0.53	0.59	0.56
	Lasso	0.54	0.88	0.71	0.56	0.59	0.57	0.54	0.88	0.71
	Rank sum	0.44	0.44	0.44	0.57	0.58	0.58	0.44	0.44	0.44
	PSODT	0.48	0.56	0.52	0.56	0.56	0.56	0.48	0.56	0.52

requires an algorithm that controls type I error. In our study, we present the combination of edgeR and DESeq as a strategy for selecting the significant variables for large sample sizes.

This study has some advantages over previous studies [18, 19]. First, the simulations not only assumed that the NGS data had a NB distribution but also compared the FS or classification algorithms in different settings of the mean, DP, and s2n. Lacking a gold standard, the real data failed to compare the FS methods. To guarantee the effectiveness, the parameter settings are obtained from the real data. Second, this study involves not only three sequencing algorithms but also machine learning methods.

However, this study also has many drawbacks. First, the three involved classifiers perhaps neglect the interactions between different variables; however, the interactions play important roles in explaining the association between molecules and diseases. With the network successfully used in biology, the classifiers based on network are perhaps more suitable to explain the association [20]. Second, some new bioinformatics classifiers are not included, such as LibD3C, HPFP, and miRClassify [21–23]. Particularly, LibD3C, classifying the cytokines from the protein sequence, applies ensemble classifiers in each layer to improve the prediction accuracy and uses SMOTE to overcome the imbalance of

samples. It also selects 120 features as the eight physicochemical properties of protein and can be used in analyzing the sequencing data [21].

When studying real data, we found that mir-133a-1, mir-133b, mir-183, mir-937, and mir-96 were frequent miRNAs sets in six cancers, and some combination of these can increase the probability of finding others. By regulating the expression of MCL-1 and BCL2L2, mir-133b is associated with lung cancer, which was also observed in our results [24]. As one of the frequent item sets, mir-133b is also related to oesophageal squamous cell carcinoma by *FSCN1* [25]. mir-96 and mir-183 both contribute to the stage and grade of urothelial carcinoma [26].

In conclusion, we propose the use of a combination of edgeR and DESeq to analyse miRNA sequencing data with a large sample size. Apriori detects the frequent item sets that might contribute to other tumours.

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 Natural Science Foundation of China (nos. 81530088, 61301251, 81473070, 81502888, and 81373102), the Research and Innovation Project for College Graduates of Jiangsu Province (no. KYLX_0944), Jiangsu Natural Science Foundation (no. BK20140907), the Jiangsu Shuangchuang Plan, the Science and Technology Development Fund Key Project of Nanjing Medical University (no. 2014NJMUZD003), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

[1] E. Huntzinger and E. Izaurralde, "Gene silencing by microRNAs: contributions of translational repression and mRNA decay," *Nature Reviews Genetics*, vol. 12, no. 2, pp. 99–110, 2011.

[2] Z. Williams, I. Z. Ben-Dov, R. Elias et al., "Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 11, pp. 4255–4260, 2013.

[3] S. Zadran, F. Remacle, and R. D. Levine, "MiRNA and mRNA cancer signatures determined by analysis of expression levels in large cohorts of patients," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 47, pp. 19160–19165, 2013.

[4] A. Git, H. Dvinge, M. Salmon-Divon et al., "Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression," *RNA*, vol. 16, no. 5, pp. 991–1006, 2010.

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

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, Springer, 2013.

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

[8] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[9] T. J. Hardcastle and K. A. Kelly, "BaySeq: empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics*, vol. 11, article 422, 2010.

[10] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, article R106, 2010.

[11] K.-H. Chen, K.-J. Wang, M.-L. Tsai et al., "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm," *BMC Bioinformatics*, vol. 15, article 49, 2014.

[12] R. Díaz-Uriarte and S. A. de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, article 3, 2006.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society—Series B: Statistical Methodology*, vol. 73, no. 3, pp. 273–282, 2011.

[14] M. D. Robinson and G. K. Smyth, "Small-sample estimation of negative binomial dispersion, with applications to SAGE data," *Biostatistics*, vol. 9, no. 2, pp. 321–332, 2008.

[15] S.-D. Hsu, Y.-T. Tseng, S. Shrestha et al., "MiRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions," *Nucleic Acids Research*, vol. 42, no. 1, pp. D78–D85, 2014.

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

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

[18] V. M. Kvam, P. Liu, and Y. Si, "A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data," *American Journal of Botany*, vol. 99, no. 2, pp. 248–256, 2012.

[19] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinformatics*, vol. 11, article 94, 2010.

[20] Q. Zou, J. Li, C. Wang, and X. Zeng, "Approaches for recognizing disease genes based on network," *BioMed Research International*, vol. 2014, Article ID 416323, 10 pages, 2014.

[21] Q. Zou, Z. Wang, X. Guan, B. Liu, Y. Wu, and Z. Lin, "An approach for identifying cytokines based on a novel ensemble classifier," *BioMed Research International*, vol. 2013, Article ID 686090, 11 pages, 2013.

[22] C. Lin, Y. Zou, J. Qin et al., "Hierarchical classification of protein folds using a novel ensemble classifier," *PLoS ONE*, vol. 8, no. 2, Article ID e56499, 2013.

[23] Q. Zou, Y. Mao, L. Hu, Y. Wu, and Z. Ji, "miRClassify: an advanced web server for miRNA family classification and annotation," *Computers in Biology and Medicine*, vol. 45, no. 1, pp. 157–160, 2014.

[24] M. Crawford, K. Batte, L. Yu et al., "MicroRNA 133B targets pro-survival molecules MCL-1 and BCL2L2 in lung cancer," *Biochemical and Biophysical Research Communications*, vol. 388, no. 3, pp. 483–489, 2009.

[25] M. Kano, N. Seki, N. Kikkawa et al., "miR-145, miR-133a and miR-133b: tumor-suppressive miRNAs target FSCN1 in esophageal squamous cell carcinoma," *International Journal of Cancer*, vol. 127, no. 12, pp. 2804–2814, 2010.

[26] Y. Yamada, H. Enokida, S. Kojima et al., "MiR-96 and miR-183 detection in urine serve as potential tumor markers of urothelial carcinoma: correlation with stage and grade, and comparison with urinary cytology," *Cancer Science*, vol. 102, no. 3, pp. 522–529, 2011.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 246374, 5 pages
http://dx.doi.org/10.1155/2015/246374

Research Article

Identification of Chemical Toxicity Using Ontology Information of Chemicals

Zhanpeng Jiang, Rui Xu, and Changchun Dong

School of Software, Harbin University of Science and Technology, Harbin 150080, China

Correspondence should be addressed to Zhanpeng Jiang; jzpp@vip.qq.com

Received 26 January 2015; Revised 20 March 2015; Accepted 22 March 2015

Academic Editor: Tao Huang

With the advance of the combinatorial chemistry, a large number of synthetic compounds have surged. However, we have limited knowledge about them. On the other hand, the speed of designing new drugs is very slow. One of the key causes is the unacceptable toxicities of chemicals. If one can correctly identify the toxicity of chemicals, the unsuitable chemicals can be discarded in early stage, thereby accelerating the study of new drugs and reducing the R&D costs. In this study, a new prediction method was built for identification of chemical toxicities, which was based on ontology information of chemicals. By comparing to a previous method, our method is quite effective. We hope that the proposed method may give new insights to study chemical toxicity and other attributes of chemicals.

1. Introduction

In drug discovery, detecting the toxicity of candidate drugs is a very important procedure. Some approved drugs such as phenacetin [1] and troglitazone [2], which have passed Phase III clinical trials, have to be withdrawn from the market, because their unexpected toxicities were detected. Pharmaceutical companies thus lost millions of dollars. In view of this, it is necessary to detect the toxicity of chemicals before they are selected as candidate drugs. However, evaluating the toxicity of a certain chemical requires comprehensive experimental testing, which costs millions of dollars and takes many years. On the other hand, with the advance of the combinatorial chemistry, a large number of synthetic compounds have surged, inducing that detecting chemical toxicities through traditional methods is an impossible task. Thus, quick, effective, and non-animal-involved prediction methods are urgently necessary.

In recent years, some prediction methods have been built for detecting chemical toxicities. Most of them can only deal with a single toxicity at the same time [3, 4], that is, predict a certain chemical to be toxic or nontoxic for a single toxicity. To detect all toxicities of a chemical, these methods have to be executed many times. Recently, Chen et al. built a multi-class prediction method using chemical-chemical interaction information [5], which can provide a candidate toxicity sequence ranging from the most likely toxicity to the least likely one. Their method was applied to detect the toxicities of chemicals listed in Accelrys Toxicity Database [6], in which six types of toxicity are reported: (1) acute toxicity; (2) mutagenicity; (3) tumorigenicity; (4) skin and eye irritation; (5) reproductive effects; (6) multiple dose effects. In this study, we employed the data in Chen et al.'s study [5] and adopted a new kind of information of chemicals to identify chemical toxicities. ChEBI ontology, integrated in a well-known database ChEBI (Chemical Entities of Biological Interest) [7], reports the ontology information of chemicals and is composed of the following subontologies: (1) molecular structure; (2) biological role; (3) application; (4) subatomic particle. Since gene ontology [8], the ontology information for proteins has been deemed to be a useful tool to investigate protein-related problems [9–12]. It is believed that ChEBI ontology is also a useful tool for studying chemicals and building effective prediction methods to identify chemical attributes. Here, we established a prediction method based on this information and compared to the method reported in [5]. The results indicate that this information is suitable to identify chemical toxicity. And we hope that the proposed method may stimulate extensive investigation based on this information, thereby promoting the study of chemicals and drug discovery.

2. Materials and Methods

2.1. Dataset. The toxicity information of chemicals was retrieved from a previous study [5], which was collected from the Accelrys Toxicity Database [6]. Six types of toxicity are reported in this database; there are (1) acute toxicity; (2) mutagenicity; (3) tumorigenicity; (4) skin and eye irritation; (5) reproductive effects; (6) multiple dose effects. Thus, the toxic chemicals in Accelrys Toxicity Database can be assigned to six classes. To investigate the problem of predicting chemical toxicity more throughout, we also employed the nontoxic chemicals, which were also retrieved from Chen et al.'s study [5]. These chemicals were collected from DrugBank (http://www.drugbank.ca/) [13] and Human Metabolome database (HMDB) (http://www.hmdb.ca/) [14]. Totally, 174,137 chemicals were collected and each of them was nontoxic or had at least one type of toxicity.

To obtain a well-defined dataset, the chemicals with no ontology information were excluded, resulting in 4,177 chemicals. Thus, we obtained a dataset $\mathbf{S}$ consisting of 4,177 chemicals, in which 3,769 chemicals were toxic and 408 chemicals were nontoxic. As mentioned in the above paragraph, each toxic chemical has at least one type of toxicity. For convenience, let us tag the six types of toxicity using $t_1, t_2, \ldots, t_6$ and nontoxicity using t_7. Accordingly, the dataset $\mathbf{S}$ can be separated into seven subsets formulated by

$$\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_2 \cup \mathbf{S}_3 \cup \mathbf{S}_4 \cup \mathbf{S}_5 \cup \mathbf{S}_6 \cup \mathbf{S}_7, \quad (1)$$

where $\mathbf{S}_i$ consisted of chemicals having toxicity t_i. The number of chemicals in each subset (i.e., number of chemicals having each type of toxicity) is listed in Table 1, column 3, from which we can see that the acute toxicity was a greatest type of toxicity containing most chemicals, followed by mutagenicity, multiple dose effects, and so forth, while the number of nontoxic chemicals was least. Since some chemicals may have more than one type of toxicity, that is, they may occur in more than one set of $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_6$, the sum of numbers in seven subsets was larger than the total number of chemicals in $\mathbf{S}$. Thus, it is a multilabel classification problem. Figure 1 gives the number of chemicals having 1–7 types of toxicity. Like many previous studies dealing with multilabel classification problem [5, 15, 16], the proposed method would give a series of candidate toxicities for each query chemical with the sequence from most likely toxicity to the least likely one.

2.2. Construction of a Graph by Ontology Information of Compound. The ontology information of compound was retrieved from ChEBI (http://www.ebi.ac.uk/chebi/init.do) [7]. We downloaded a file named as "chebi.obo" (accessed November 2014) from its ftp website: ftp://ftp.ebi.ac.uk/pub/databases/chebi/ontology/, which contains larger number of ontology terms and their descriptions. Since the ontology terms can be conceived as graph-theoretical structures, a graph can be constructed according to the information of all ontology terms, in which nodes represent ontology terms and edges denote the relationship between two terms. By using the entries "is a" and "relationship" in the obtained file to indicate the relationship between two terms, we constructed a large graph G with 45,206 nodes and 113,549 edges.

Table 1: Distribution of chemicals in $\mathbf{S}$ and $\mathbf{S}_c$.

Tag of toxicity	Type of toxicity	Number of chemicals in $\mathbf{S}$[a]	Number of chemicals in $\mathbf{S}_c$[b]
t_1	Acute toxicity	3144	2993
t_2	Mutagenicity	1850	1814
t_3	Tumorigenicity	881	871
t_4	Skin and eye irritation	954	935
t_5	Reproductive effects	1099	1080
t_6	Multiple dose effects	1600	1570
t_7	Nontoxic	408	374
Total	—	9936	9637

[a] $\mathbf{S}$ is a chemical set consisting of 4,177 chemicals, which was used to examine our method.
[b] $\mathbf{S}_c$ is another chemical set consisting of 3,955 chemicals, which was used to compare our method with a previous method.

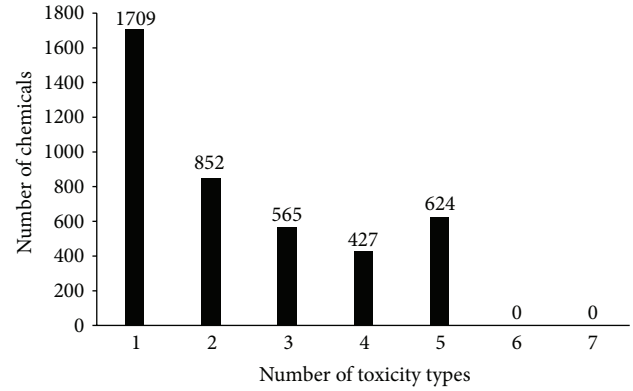


Figure 1: A histogram illustrating the number of chemicals having 1–7 types of toxicity.

2.3. Prediction Method. As mentioned in Section 2.2, a graph was constructed according to the ontology information of compounds. It can be observed that the corresponding ontology terms of two adjacent nodes in G have some special relationship. And it can be further inferred that if two nodes are with small distance in G, the corresponding ontology terms have close linkage. In view of this, using the distance in G to quantitatively measure the relationship between two ontology terms is reasonable. For two terms a_1 and a_2, let us denote the distance of the corresponding nodes in G by $d(a_1, a_2)$.

For two chemicals c_1 and c_2, let $a_{11}, a_{12}, \ldots, a_{1k}$ be the ontology terms of c_1 and let $a_{21}, a_{22}, \ldots, a_{2l}$ be the ontology terms of c_2. It is obvious that if $d(a_{1i}, a_{2j})$ $(1 \le i \le k, 1 \le j \le l)$ is small, c_1 and c_2 are highly related and have high probability to share same structures, functions, and so on. Thus, we gave the following formulation to measure the common features of chemicals c_1 and c_2:

$$S(c_1, c_2) = \min\left\{d\left(a_{1i}, a_{2j}\right) : 1 \le i \le k,\ 1 \le j \le l\right\}, \quad (2)$$

where $d(a_{1i}, a_{2j})$ denote the distance of terms a_{1i} and a_{2j} in the graph constructed in Section 2.2, which can be obtained

by Dijkstra's algorithm [17]. The smaller the $S(c_1, c_2)$ is, the closer the relationship c_1 and c_2 have.

The proposed prediction method highly relied on the result of (2). To introduce the method clearly, it is necessary to employ some notations. Let $\mathbf{S}'$ be a training set consisting of n chemicals, say $c_1, c_2, \ldots, c_n$; that is, $\mathbf{S}' = \{c_1, c_2, \ldots, c_n\}$. The toxicity information of each c_i $(1 \le i \le n)$ can be represented by

$$T(c_i) = [b_{i1}, b_{i2}, \ldots, b_{i7}]^T \quad (1 \le i \le n), \tag{3}$$

where b_{ij} $(1 \le j \le 7)$ was defined by

$$b_{ij} = \begin{cases} 1 & \text{if } c_i \text{ has toxicity } t_j \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

For a query chemical c, its score of having toxicity t_j was calculated as follows.

(1) For each chemical c_i in the training set $\mathbf{S}'$, calculate $S(c, c_i)$ according to (2). Then, find all nearest neighbors, say $c_1, c_2, \ldots, c_k$, without generalization, such that $S(c, c_i) = \min\{S(c, c') : c' \in \mathbf{S}'\}$ $(1 \le i \le k)$.

(2) For each t_j, the score of c having toxicity t_j was calculated by

$$P(c \rhd t_j) = \sum_{i=1}^{k} b_{ij}. \tag{5}$$

It is easy to observe that the score of c having toxicity t_j is the number of chemicals among $c_1, c_2, \ldots, c_k$ which have toxicity t_j. Since $c_1, c_2, \ldots, c_k$ are highly related to c, larger $P(c \rhd t_j)$ indicates that many closely related training chemicals of c have toxicity t_j, inducing that the probability of c having toxicity t_j is high. In particular, $P(c \rhd t_j) = 0$ suggests that the score of c having toxicity t_j is zero, inducing that the possibility of c having this toxicity is zero.

As mentioned in Section 2.1, the investigated problem is a multilabel classification problem. Only giving the most likely candidate toxicity is not enough. Fortunately, we can output a series of candidate toxicities according to the scores of the query chemical having 7 types of toxicity. The toxicity which receives the highest score is the most likely toxicity, while the toxicity receiving the second highest score is the second likely toxicity and so forth. For example, if the rank of seven scores for a certain query chemical c is

$$\begin{aligned} P(c \rhd t_1) \ge P(c \rhd t_4) \ge P(c \rhd t_2) > P(c \rhd t_3) = P(c \rhd t_5) \\ = P(c \rhd t_6) = P(c \rhd t_7) = 0, \end{aligned} \tag{6}$$

it suggests t_1 (i.e., acute toxicity) is the most likely toxicity for c, followed by t_4 (i.e., skin and eye irritation) and t_2 (i.e., mutagenicity), while the other types of toxicity are not predicted to be candidate toxicities for c. Furthermore, t_1 is called the first prediction, t_4 the second prediction, and so forth.

Table 2: Performance of the methods on $\mathbf{S}$ and $\mathbf{S}_c$.

Prediction order	Our method on $\mathbf{S}$[a]	Our method on $\mathbf{S}_c$[b]	Chen et al.'s method on $\mathbf{S}_c$[b]
1st	75.17%	75.40%	75.14%
2nd	43.52%	45.18%	49.87%
3rd	28.47%	29.76%	34.11%
4th	23.34%	24.15%	29.94%
5th	16.78%	17.98%	27.00%
6th	9.74%	10.24%	19.97%
7th	3.16%	3.16%	5.54%

[a] $\mathbf{S}$ is a chemical set consisting of 4,177 chemicals, which was used to examine our method.
[b] $\mathbf{S}_c$ is another chemical set consisting of 3,955 chemicals, which was used to compare our method with a previous method.

2.4. Accuracy Measurements. For a query chemical, the proposed method can provide a series of candidate toxicities. In view of this, we should calculate the accuracy for each order prediction. The kth prediction accuracy can be computed by [5, 15]

$$\mathrm{ACC}_k = \frac{CP_k}{N} \quad k = 1, 2, \ldots, 7, \tag{7}$$

where CP_k is the number of chemicals whose kth prediction is correct and N is the total number of chemicals that are predicted by the method. Since it is difficult to know the number of toxicities for a query chemical, the first prediction accuracy is the most important measure to evaluate the performance of the method. In addition, an effective prediction method for a multilabel classification problem should rank the candidate toxicities well; that is, prediction accuracies should follow a decreasing trend with the increasing of the prediction order.

Besides, to evaluate the performance of prediction method on the whole, another measurement was also adopted [5, 15]. It measures the proportion of the true toxicities covered by the first m predictions of chemicals, which can be calculated by

$$W_m = \frac{\sum_{i=1}^{N} \Psi_i^m}{N_i}, \tag{8}$$

where Ψ_i^m is the number of true toxicities of the ith chemical which are listed among its first m predictions and N_i is the total number of true toxicities of the ith chemical. Generally, m is always taken as the smallest integer bigger than or equal to the average number of toxicities of chemicals processed by the method; that is, $m = \lceil \sum_{i=1}^{N} N_i / N \rceil$. It is obvious that larger W_m indicates the true toxicities are arranged in the front of candidate toxicities.

3. Results and Discussion

3.1. Performance of the Method. For the 4,177 chemicals in $\mathbf{S}$, the prediction method was executed to identify their toxicities evaluated by jackknife test [15]. The seven prediction accuracies thus obtained by (7) are listed in Table 2,

Table 3: Chemicals with closest relationship of CID104975.

Compound ID	Tag of toxicity	Ontology information	Shortest path to CHEBI25957
CID995	t_1, t_2, t_3, and t_6	CHEBI:28851	CHEBI:25957, CHEBI:25959, CHEBI:25961, and CHEBI:28851
CID2236	t_1, t_2, t_3, and t_6	CHEBI:2825	CHEBI:25957, CHEBI:25959, CHEBI:25961, and CHEBI:2825
CID6763	t_1, t_2, and t_3	CHEBI:37454	CHEBI:25957, CHEBI:25959, CHEBI:25961, and CHEBI:37454
CID13257	t_2	CHEBI:35860	CHEBI:25957, CHEBI:25959, CHEBI:25961, and CHEBI:35860

column 2. It can be observed that the first prediction accuracy was 75.17%, the second one was 43.52%, and the third one was 28.47%. Furthermore, seven prediction accuracies always followed a decreasing trend with the increasing of the prediction order, indicating the proposed method arranged the candidate toxicities of all tested chemicals quite well. In addition, the average number of toxicities of chemicals in **S** was about 2.38. Thus, the first three predictions of all chemicals in **S** were collected, obtaining the accuracy of 61.87% by (8), which means the proportion of the true toxicities of chemicals in **S** covered by their first three predictions. All of these indicate that the proposed method is quite effective for identification of chemical toxicities.

3.2. Understanding the Method by Listing an Example. To better understand our method, this section listed an example. CID104975 is a chemical with toxicity t_2 (mutagenicity) and t_3 (tumorigenicity). Its ontology term is CHEBI:25957. According to the method, we computed the distance between CHEBI:25957 and ontology terms of other chemicals in **S**, thereby calculating the relationship between CID104975 and other chemicals by (2). Four chemicals, listed in Table 3, were found to be closely related to CID104975; they are CID995, CID2236, CID6763, and CID13257. Their toxicities and ontology terms are listed in Table 3, column 2 and column 3, respectively. By the method, the toxicity t_1 received 3 votes, t_2 4 votes, t_3 3 votes, t_6 2 votes, and other toxicities no votes. Accordingly, we obtained that the candidate toxicities for CID104975 were t_2, t_1, t_3, and t_6. It is obvious that the first and third predictions were correct, while the second prediction was incorrect.

3.3. Comparison of Other Methods. In this section, we employed another kind of chemical information, which has been applied for identification of chemical toxicities in Chen et al.'s study [5]. Their method used chemical-chemical interaction information, which has been deemed to be useful information for study of chemical-related problems [5, 15, 18, 19], to build the prediction method, and gave good performance.

To compare our method and Chen et al.'s method in a fair circumstance, a chemical set, consisting of 3,955 chemicals, was extracted from **S**, called $\mathbf{S_c}$, such that each chemical in $\mathbf{S_c}$ has both ontology information and interaction information; that is, each chemical can be predicted by these two methods. The number of chemicals in $\mathbf{S_c}$ on each type of toxicity is listed in Table 1, column 4, from which we can see that the distribution of 3,955 chemicals on seven types of toxicity is similar to chemicals in **S**. Also some chemicals have two or more toxicities. Our method and Chen et al.'s method were all executed on $\mathbf{S_c}$ with their performance being evaluated by jackknife test. Listed in Table 2, columns 3 and 4, are seven prediction accuracies. It can be seen that the first prediction accuracy of our method was 75.40%, which is little higher than 75.14% of Chen et al.'s method. However, with the increasing of prediction order, the prediction accuracies of Chen et al.'s method were higher than those obtained by our method. It is reasonable because the ontology information of chemicals is not very complete at present, which induces that many relations of ontology terms have not been detected. Furthermore, we also calculated the measurement defined in (8). Since the average number of toxicities of chemical in $\mathbf{S_c}$ was about 2.44, the first three predictions of chemicals in $\mathbf{S_c}$, which were obtained by two methods, were collected, thereby obtaining the accuracy of 61.70% for our method and 65.31% for Chen et al.'s method. It is also caused by the aforementioned reason. Although, if one considers more than one toxicity for a certain chemical, our method is not better than Chen et al.'s method, the first prediction accuracy of our method is higher than that of Chen et al.'s method, which is the most important one because one always pays more attention to the most likely toxicity for a chemical. In view of this, we believe that our method has superiority for identification of chemical toxicities.

4. Conclusions

This study gave a new prediction method to identify chemical toxicities. By utilizing the ontology information of chemicals reported in ChEBI, one can predict the toxicities of a certain chemical with quite high quality. It is hopeful that this method may promote the study of chemicals.

Conflict of Interests

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

References

[1] U. C. Dubach, B. Rosner, and T. Stürmer, "An epidemiologic study of abuse of analgesic drugs. Effects of phenacetin and salicylate on mortality and cardiovascular morbidity (1968 to

1987)," *The New England Journal of Medicine*, vol. 324, no. 3, pp. 155–160, 1991.

[2] "AstraZeneca Decides to Withdraw Exanta," 2006, http://www.astrazeneca.com/Media/Press-releases/Article/20060214–AstraZeneca-Decides-to-Withdraw-Exanta.

[3] M. Zheng, Z. Liu, C. Xue et al., "Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine," *Bioinformatics*, vol. 22, no. 17, pp. 2099–2106, 2006.

[4] Y. Wang, J. Lu, F. Wang et al., "Estimation of carcinogenicity using molecular fragments tree," *Journal of Chemical Information and Modeling*, vol. 52, no. 8, pp. 1994–2003, 2012.

[5] L. Chen, J. Lu, J. Zhang, K.-R. Feng, M.-Y. Zheng, and Y.-D. Cai, "Predicting chemical toxicity effects based on chemical-chemical interactions," *PLoS ONE*, vol. 8, no. 2, Article ID e56517, 2013.

[6] Accelrys Software Inc, *Accelrys Toxicity Database 2011.4*, Accelrys Software Inc., San Diego, Calif, USA, 2011.

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

[8] 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.

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

[10] C.-S. Yu, C.-W. Cheng, W.-C. Su et al., "CELLO2GO: a web server for protein subcellular localization prediction with functional gene ontology annotation," *PLoS ONE*, vol. 9, no. 6, Article ID e99368, 2014.

[11] C. Bettembourg, C. Diot, and O. Dameron, "Semantic particularity measure for functional characterization of gene sets using gene ontology," *PLoS ONE*, vol. 9, no. 1, Article ID e86525, 2014.

[12] K.-C. Chou and Y.-D. Cai, "A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology," *Biochemical and Biophysical Research Communications*, vol. 311, no. 3, pp. 743–747, 2003.

[13] D. S. Wishart, C. Knox, A. C. Guo et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, no. 1, pp. D901–D906, 2008.

[14] D. S. Wishart, D. Tzur, C. Knox et al., "HMDB: the human metabolome database," *Nucleic Acids Research*, vol. 35, no. 1, pp. D521–D526, 2007.

[15] L. Chen, W.-M. Zeng, Y.-D. Cai, K.-Y. Feng, and K.-C. Chou, "Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities," *PLoS ONE*, vol. 7, no. 4, Article ID e35254, 2012.

[16] P. Du, T. Li, and X. Wang, "Recent progress in predicting protein sub-subcellular locations," *Expert Review of Proteomics*, vol. 8, no. 3, pp. 391–404, 2011.

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

[18] L.-L. Hu, C. Chen, T. Huang, Y.-D. Cai, and K.-C. Chou, "Predicting biological functions of compounds based on chemical-chemical interactions," *PLoS ONE*, vol. 6, no. 12, Article ID e29491, 2011.

[19] L. Chen, J. Lu, T. Huang et al., "Finding candidate drugs for hepatitis C based on chemical-chemical and chemical-protein interactions," *PLoS ONE*, vol. 9, no. 9, Article ID e107767, 2014.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 281589, 8 pages
http://dx.doi.org/10.1155/2015/281589

Research Article

An Improved PID Algorithm Based on Insulin-on-Board Estimate for Blood Glucose Control with Type 1 Diabetes

Ruiqiang Hu and Chengwei Li

School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin 150001, China

Correspondence should be addressed to Ruiqiang Hu; ruiqianghu@hit.edu.cn

Received 16 April 2015; Revised 27 May 2015; Accepted 2 June 2015

Academic Editor: Tao Huang

Automated closed-loop insulin infusion therapy has been studied for many years. In closed-loop system, the control algorithm is the key technique of precise insulin infusion. The control algorithm needs to be designed and validated. In this paper, an improved PID algorithm based on insulin-on-board estimate is proposed and computer simulations are done using a combinational mathematical model of the dynamics of blood glucose-insulin regulation in the blood system. The simulation results demonstrate that the improved PID algorithm can perform well in different carbohydrate ingestion and different insulin sensitivity situations. Compared with the traditional PID algorithm, the control performance is improved obviously and hypoglycemia can be avoided. To verify the effectiveness of the proposed control algorithm, *in silico* testing is done using the UVa/Padova virtual patient software.

1. Introduction

Diabetes, a disorder of endocrine metabolism, is an incurable disease. Diabetes affects millions of people in the world, and it is a disease with considerable complications including retinopathy, nephropathy, peripheral neuropathy, and blindness [1]. According to a prediction produced by the International Diabetes Federation in 2014, approximately 387 million people suffered from diabetes worldwide by 2014 and about 592 million patients by 2035 [2]. Thus, the maintenance of blood glucose concentration in a normal range is of critical importance for diabetic.

Type 1 diabetes is mainly due to the reason that the β-cell of pancreas cannot secrete insulin. They must rely on exogenous insulin to regulate blood glucose concentration. Currently, patients with type 1 diabetes are treated with either multiple daily injections (MDI) or continuous subcutaneous insulin infusion (CSII) delivering via an insulin pump [3]. The CSII has shown more advantages than MDI method because of the increasing flexibility of diet, exercise, convenience, and precision [4]. Various open-loop insulin pumps that are available in the market are programmable to deliver the required amount of insulin. However, a fully automated closed-loop insulin infusion system that can deliver appropriate amounts of insulin to patients without any manual interference is developing [5]. The closed-loop system contains three main components, which are (1) continuous glucose monitoring (CGM), (2) intelligent controller, and (3) insulin pump.

For open-loop insulin pump, a bolus calculator is used to calculate bolus insulin doses that can help diabetic regulate the postprandial blood glucose concentration. The bolus calculator takes into account many factors, such as current blood glucose, target blood glucose, amount of carbohydrate ingested, insulin sensitivity, correction factor (CF), and insulin : carbohydrate ratio (I : C) as well as duration of insulin action ("insulin on board (IOB)") [4].

For closed-loop insulin pump, the real-time CGM system is already commercially available and the control algorithm is the key technique of precise insulin infusion. The control algorithm requires high robustness and reliability. There are various control algorithms including PID control [6], model predictive control (MPC) [7, 8], optimal control [9], adaptive control [10], and sliding mode control [11]. Among those control algorithms, the PID controller is widely used in industrial control systems. The PID controller is attractive

for blood glucose control based on the features of simple structure with few parameters, easy implementation, good adaptation, and robustness.

Many closed-loop control algorithms had not considered the IOB factor. A limitation of IOB can optimize the output of control algorithm and decrease the risk of hypoglycemia. As the open-loop insulin pump, the previous insulin administration may lead to hypoglycemia. So the IOB estimate is considered to limit the insulin infusion dose. In this paper, the improved PID control algorithm based on IOB estimate is introduced. Controller performance is evaluated in a simulation study under a physiological model and considered the carbohydrate ingestion and insulin sensitivity changed. Also, the UVa/Padova virtual patient software is used to verify the effectiveness of PID controller with IOB estimate.

The paper is organized as follows. In Section 2, a combinational complicated and detailed model of glucose-insulin kinetic is introduced, which is based on the Hovorka et al. and Dalla Man et al. model. The PID controller with IOB estimate is designed, and the performance is evaluated by simulation in Section 3. The *in silico* testing using ten virtual patients is discussed in Section 4. The final conclusion is located in Section 5.

2. Glucose-Insulin Mathematical Model

Mathematical models of glucose-insulin interactions have been studied for over the past 50 years. Simple linear models were proposed by Ackerman et al. [12]. More complicated nonlinear models were proposed in later studies. In many of these models, compartmental modeling approach has been used. In this approach, the body is divided into compartments representing different organs or parts of the body and mass balance equations are derived for each compartment. The compartmental minimal model of Bergman et al. [13] has been widely used in many studies. More complicated compartmental models proposed by Cobelli and Mari [14], Hovorka et al. [15], and Dalla Man et al. [16] have considered more compartments for better understanding the behavior of different parts of the body. These models for glucose-insulin interactions have been widely used in studying the physiological behavior of diabetic patients.

In this paper, the glucose and insulin metabolic model refers to the model developed by Hovorka et al. [15] and Dalla Man et al. [16, 17]. The Hovorka model is a nonlinear compartmental model with three subsystems for glucose, insulin, and insulin action. The carbohydrate digestion and absorption model refers to Dalla Man's model in this paper. The combinational model is close to a realistic patient model.

2.1. Glucose Subsystem. The glucose subsystem is divided into two compartments: masses of glucose in the accessible compartment and masses of glucose in the nonaccessible compartment. The core model is a two-compartment representation of glucose kinetics. Consider

$$\begin{aligned}
\dot{Q_1}(t) &= -\left[\frac{F_{01}^c}{V_G G(t)} + x_1(t)\right] Q_1(t) + k_{12} Q_2(t) - F_R \\
&\quad + U_G(t) + \mathrm{EGP}_0\left[1 - x_3(t)\right], \\
\dot{Q_2}(t) &= x_1(t) Q_1(t) - \left[k_{12} + x_2(t)\right] Q_2(t), \\
G(t) &= \frac{Q_1(t)}{V_G}, \\
F_{01}^c &= \begin{cases} F_{01}^c & \text{if } G \geq 81\ \text{mg/dL}, \\ \dfrac{F_{01} G}{4.5} & \text{otherwise}, \end{cases} \\
F_R &= \begin{cases} 0.003\,(G - 9)\,V_G & \text{if } G \geq 162\ \text{mg/dL}, \\ 0 & \text{otherwise}, \end{cases}
\end{aligned} \tag{1}$$

where Q_1 and Q_2 are the masses of glucose in the accessible and nonaccessible compartments, respectively. k_{12} is the transfer rate constant from the nonaccessible to the accessible compartment. V_G is the distribution volume of the accessible compartment. G is the glucose concentration. EGP_0 is the endogenous glucose production extrapolated to the zero insulin concentration. F_{01}^c is the total insulin-independent glucose flux, corrected for the ambient glucose concentration. F_R is the renal glucose clearance above the glucose threshold of 162 mg/dL. The gut absorption rate U_G is introduced in Section 2.4 of carbohydrate digestion and absorption model.

2.2. Insulin Subsystem. The insulin subsystem describes the insulin absorption and insulin action on glucose kinetics. The plasma insulin concentration $I(t)$ is described by

$$\dot{I}(t) = \frac{U(t)}{V_I} - k_e I(t), \tag{2}$$

where k_e is the fractional elimination rate and V_I is the distribution volume.

2.3. Insulin Action Subsystem. The three insulin actions on glucose kinetics are represented by

$$\begin{aligned}
\dot{x_1} &= -k_{a1} x_1(t) + k_{b1} I(t), \\
\dot{x_2} &= -k_{a2} x_2(t) + k_{b2} I(t), \\
\dot{x_3} &= -k_{a3} x_3(t) + k_{b3} I(t),
\end{aligned} \tag{3}$$

where x_1, x_2, and x_3 are the effects of insulin on glucose distribution/transport, glucose disposal, and endogenous glucose production, respectively; k_{a1}, k_{a2}, and k_{a3} are the deactivation rate constants; k_{b1}, k_{b2}, and k_{b3} are the activation rate constants.

The insulin sensitivities of glucose distribution/transport and glucose intracellular disposal are represented individually as follows:

$$\begin{aligned}
S_{\mathrm{IT}}^f &= \frac{k_{b1}}{k_{a1}}, \\
S_{\mathrm{ID}}^f &= \frac{k_{b2}}{k_{a2}}, \\
S_{\mathrm{IE}}^f &= \frac{k_{b3}}{k_{a3}}.
\end{aligned} \tag{4}$$

Table 1: IOB model parameter k_{DIA} for different durations of insulin action.

DIA (h)	2	3	4	5	6	7	8
$k_{DIA} \times 10^{-3}$	39	26	19.5	16.3	13	11.3	9.9

2.4. Carbohydrate Digestion and Absorption. The carbohydrate digestion and absorption model consists of three-compartment nonlinear model, two for the glucose in the stomach solid Q_{sto1} and liquid Q_{sto2} and one for the glucose in the intestinal tract Q_{gut}. Consider

$$\begin{aligned} \dot{Q_{sto1}}(t) &= -k_{gri} * Q_{sto1}(t) + D * \delta(t), \\ \dot{Q_{sto2}}(t) &= -k_{empt} * Q_{sto2}(t) + k_{gri} * Q_{sto1}(t), \\ \dot{Q_{gut}}(t) &= -k_{abs} * Q_{gut}(t) + k_{empt} * Q_{sto2}(t), \\ \dot{Q_{sto}}(t) &= Q_{sto1}(t) + Q_{sto2}(t), \end{aligned} \tag{5}$$

where D is the amount of carbohydrate to be ingested, $\delta(t)$ is the impulse function, k_{gri} is the rate of grinding coefficient in the stomach, k_{empt} is the rate of fractional coefficient with which the chyme enters the intestine, and k_{abs} is the rate constant of intestinal absorption.

3. PID Controller with IOB Estimate

3.1. Insulin-on-Board Estimate. The insulin on board is defined as the amount of administered insulin that is still active in the body. Some insulin pump estimates the IOB to correct the boluses in order to avoid hyper- or hypoglycemia [4]. The IOB estimate is based on the insulin action curves. Here the IOB estimation is represented by a two-compartment dynamical model:

$$\begin{aligned} \dot{C_1}(t) &= u(t) - k_{DIA} C_1(t), \\ \dot{C_2}(t) &= k_{DIA}\left(C_1(t) - C_2(t)\right), \\ \mathrm{IOB}(t) &= C_1(t) + C_2(t), \end{aligned} \tag{6}$$

where C_1 and C_2 are the two compartments and $u(t)$ is the insulin dose. The constant k_{DIA} is tuned for each patient so model replicates the corresponding DIA. Figure 1 shows the insulin activity curves obtained with model for typical DIA value, while Table 1 shows the corresponding values k_{DIA} for typical DIA values [18]. The insulin action is different among each individual; there are many factors, such as exercise, stress, illness, and heat. The different insulin action curves are provided by insulin pump to calculate insulin bolus.

The insulin duration ranges from 2 h to 8 h; diabetes patient should choose a reasonable duration time. If patient sets the duration of insulin action time less than the actual time, it will increase the risk of hypoglycemia. The insulin pump indicates that there has been no longer active IOB and will infuse more insulin dose to consume blood glucose. On the other hand, if patient sets the duration of insulin action time longer than the actual time, it will increase the risk of hyperglycemia. The patient will take a smaller insulin dose than is necessary to regulate the blood glucose back to the set value.

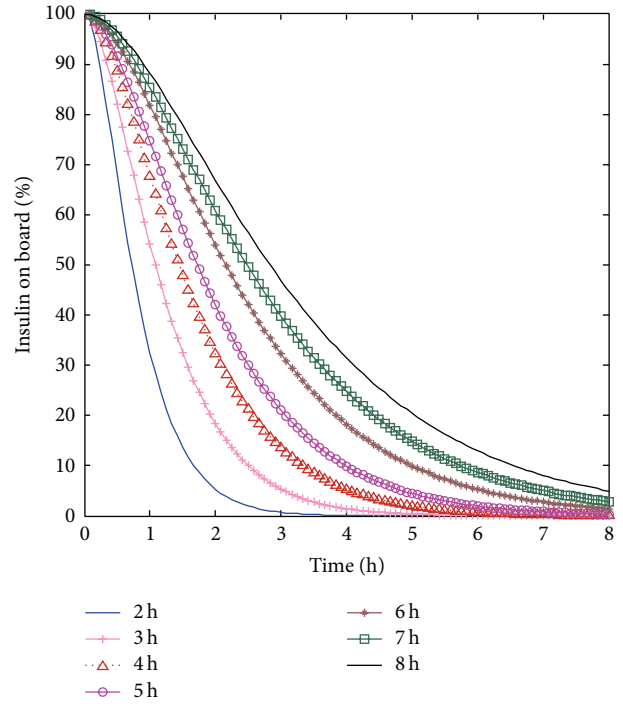


Figure 1: Estimated time profiles of insulin activity parameterized by DIA.

3.2. Design of PID Controller with IOB Estimate. The structure of the PID controller with IOB estimate is demonstrated in Figure 2, where G_o is the real blood glucose concentration of diabetes patient, G_m is the measured blood glucose by glucose sensor, G_t is the target blood glucose concentration, and U_I is the final insulin infusion rate.

The U_{PID} control law is as follows:

$$U_{PID} = U_0 + k_c\left[\left(G_m - G_t\right) + \frac{1}{\tau_I}\int_0^t \left(G_m - G_t\right)dt + \tau_D \frac{d\left(G_m - G_t\right)}{dt}\right], \tag{7}$$

where U_{PID} is the closed-loop control output. $G_m - G_t$ is the error of the target blood glucose and the measured blood glucose. U_0 is the basal insulin infusion rate. There are three adjustable parameters: proportional gain (k_c), integral time (τ_I), and derivative (τ_D).

The control output of insulin infusion rate U_I is based on the IOB estimate and the constraint output of insulin infusion. Figure 3 shows the inner structure of control algorithm. The output of controller is

$$U_I = kU_{PID}, \tag{8}$$

where the gain k is obtained as the average value of ω and the value is set as $0 \le k \le 1$.

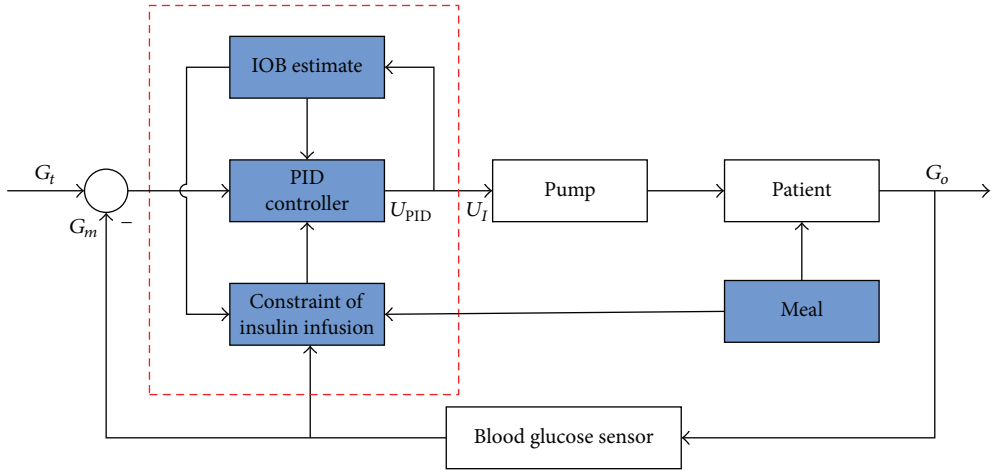


Figure 2: Structure of the improved PID controller with IOB estimate.

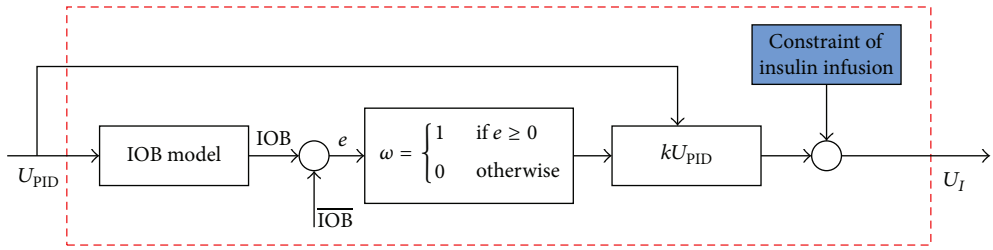


Figure 3: Structure of the IOB estimate.

The IOB estimate is based on the error of the IOB(t) (6) and $\overline{\text{IOB}}$ limit. In [19], the author proposed a method to calculate the $\overline{\text{IOB}}$ limit. The $\overline{\text{IOB}}$ value is obtained at time: (CHO+80 g)/(60 g/h), where CHO is the amount of carbohydrate intake. Although each meal is different for a patient, the corresponding limits are practically equal. For different duration of insulin action, the value of $\overline{\text{IOB}}$ is different. The error is

$$e = \overline{\text{IOB}} - \text{IOB}. \tag{9}$$

Based on (6) and (9), the time evolution of e is governed by

$$\frac{de}{dt} = k_{\text{DIA}} C_2 - \omega U_I, \tag{10}$$

that is,

$$\begin{aligned} \frac{de}{dt} &= k_{\text{DIA}} C_2 - U_I \quad \text{if } e \geq 0, \\ \frac{de}{dt} &= k_{\text{DIA}} C_2 \quad \text{if } e < 0, \end{aligned} \tag{11}$$

from (11), after infusing the insulin, the IOB increases quickly surpassing the $\overline{\text{IOB}}$; hence $e < 0$. The switching ω turns to 0. When IOB falls under the $\overline{\text{IOB}}$, e becomes a positive value and ω switches to 1. Under the control mode, ω switches between 0 and 1. We calculate the ω value during each 10 min period, and the gain k is the average value of ω. So the proposed control algorithm can decrease the insulin infusion rate and avoid the hypoglycemia event.

In this paper, the upper constraint output of PID controller is considered. The upper constraint is based on the IOB estimate, correction factor, and I : C ratio. The significance of upper constraint is to avoid overinfusion of insulin. It is calculated by the following condition:

$$\begin{aligned} &\text{If } I_{\text{CHO}} + I_G > \text{IOB}, \\ &U_{\max} = I_{\text{CHO}} + I_G - \text{IOB}, \\ &\text{Else} \quad U_{\max} = I_{\text{CHO}}, \end{aligned} \tag{12}$$

where $U_{\max}$ is the maximum constraint output of insulin infusion rate and I_{CHO} is the amount of insulin needed to compensate for a given meal and is calculated by

$$I_{\text{CHO}} = D \cdot (\text{I} : \text{C}), \tag{13}$$

where D is the mass of a given meal and I : C ratio is that 1 unit of insulin can consume the amount of CHO. I_G is the amount of insulin needed to correct for a positive deviation from the target glucose concentration and is calculated by the following condition:

$$\begin{aligned} &\text{If } G_m - G_t > 0, \\ &I_G = (G_m - G_t) \cdot \text{CF}, \\ &\text{Else} \quad I_G = 0, \end{aligned} \tag{14}$$

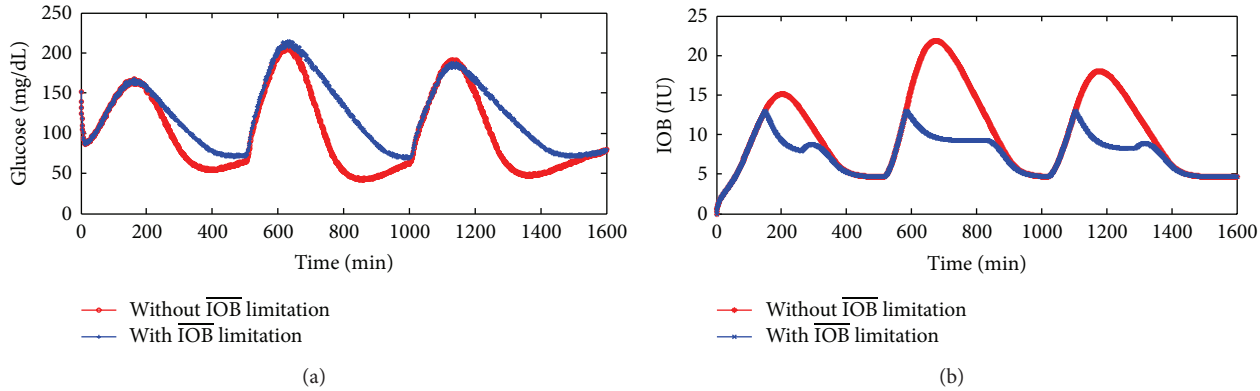


Figure 4: (a) Glucose responses profiles with IOB limitation {DIA(h) = 2 h, $\overline{IOB}$ = 13} and without IOB limitation; (b) IOB dose responses profiles with and without IOB limitation.

Table 2: Simulation conditions of glucose-insulin mathematical model.

Weight		75 kg	
DIA (h)		2 h~8 h	
$\overline{IOB}$		13~36	
Meal time	0 min	500 min	1000 min
CHO	40 g	60 g	50 g
I : C		1 U : 20 g	
CF		1 U : 80 mg/dL	

where G_m and G_t are the current measured and target blood glucose concentrations, respectively. CF is the correction factor.

3.3. Simulation Results. The proposed control algorithm is evaluated under the glucose-insulin mathematical model mentioned in Section 2. Table 2 shows the simulation conditions.

Figure 4(a) shows the glucose responses using the proposed PID controller with and without $\overline{IOB}$ limitation under the {DIA(h) = 2 h, $\overline{IOB}$ = 13} conditions. It can avoid the hypoglycemia event after three different meals ingested. When the estimated IOB reaches the limitation constraint, the switching law begins to take effect. The IOB dose falls below its limitation. Figure 4(b) shows the IOB dose responses with and without $\overline{IOB}$ limitation.

Figure 5 shows the glucose responses under the {DIA(h) = 2 h, $\overline{IOB}$ = 13} and {DIA(h) = 8 h, $\overline{IOB}$ = 36} conditions. Under different duration of insulin action, the value of $\overline{IOB}$ limitation needs to be calculated again. The simulation results indicate that the PID control with IOB estimate is effective and stabile for blood glucose control.

We all know that the insulin sensitivity (IS) is varied during a 24-hour period. The IS is employed for each of the three insulin sensitivity parameters in the Hovorka model. There are three IS values: IS_{nom}, IS_{min}, and IS_{max}. The maximum and minimum values of IS varied randomly on a daily basis following uniform distributions. IS_{max} is equal to $1.5IS_{nom}$. IS_{min} is equal to $0.5IS_{nom}$ [3]. In our simulations, IS increases $1.5IS_{nom}$ during 0–1000 min, and IS decreases $0.5IS_{nom}$ after 1000 min. To an extent, it can test the performance of controller. Figure 6 compares the glucose responses for insulin sensitivity changes under the proposed controller. It can avoid the hyperglycemia or hypoglycemia.

In order to evaluate the performance of control algorithm, the blood glucose index (BGI) and standard deviation (SD) are adopted. The BGI is a metric proposed by Kovatchev et al. [20], to evaluate the risk for hypoglycemia and hyperglycemia. BGI is equal to LBGI + HBGI, where LBGI and HBGI are low and high BG readings, respectively. SD is the standard deviation of glucose concentration. The statistical results are given in Table 3. Both PID controller and PID controller with IOB estimate are analyzed under the different insulin sensitivity. In all situations, the proposed controller has smaller BGI and SD values compared with PID controller. It demonstrates that the improved controller performs well.

4. *In Silico* Test on Virtual Patient

In order to evaluate the performance of the proposed PID controller, the test is performed on ten virtual subjects using the UVa/Padova virtual patient software. The patients are assumed to have three meals in a day. The multiple meals are 30 g CHO at 7 a.m., 50 g at 12 p.m., and 40 g at 6 p.m.

In the simulation, the control-variability grid analysis (CVGA) provides a summary of the quality of glucose regulation for a virtual subject [21]. CVGA plays an important role in the tuning of closed-loop glucose control algorithms and also in the comparison of their performance. Each subject presents by one data point for any given observation period. There are nine rectangular zones that are defined as follows: A-zone means accurate control, Lower B-zone means benign deviations into hypoglycemia, Upper B-zone means benign deviations into hyperglycemia, B-zone means benign control deviations, Lower C means overcorrection of hyperglycemia, Upper C means overcorrection of hypoglycemia, Lower D means failure to deal with hypoglycemia, Upper D means failure to deal with hyperglycemia, and E means erroneous control. Considering the sensor noise, the X-Y coordinates

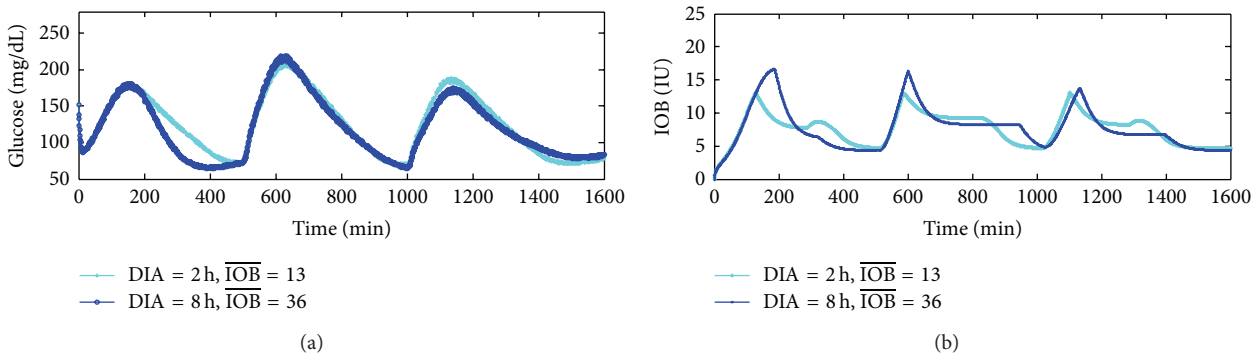


Figure 5: Glucose and IOB dose responses profiles under {DIA(h) = 2 h, $\overline{IOB}$ = 13} and {DIA(h) = 8 h, $\overline{IOB}$ = 36}, respectively.

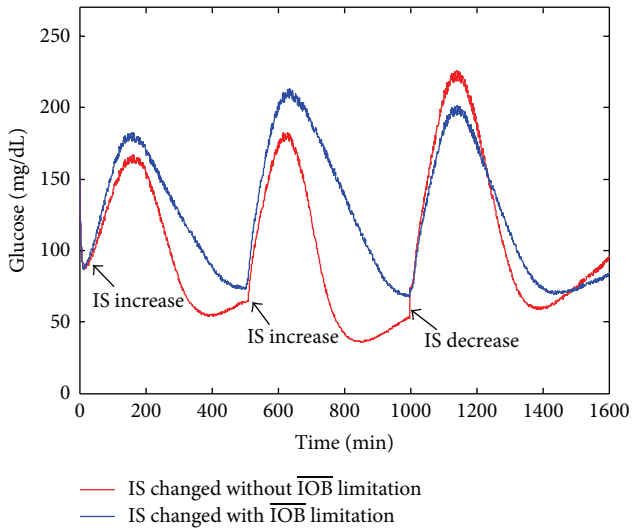


Figure 6: Glucose responses profiles under IS changed with and without IOB limitation.

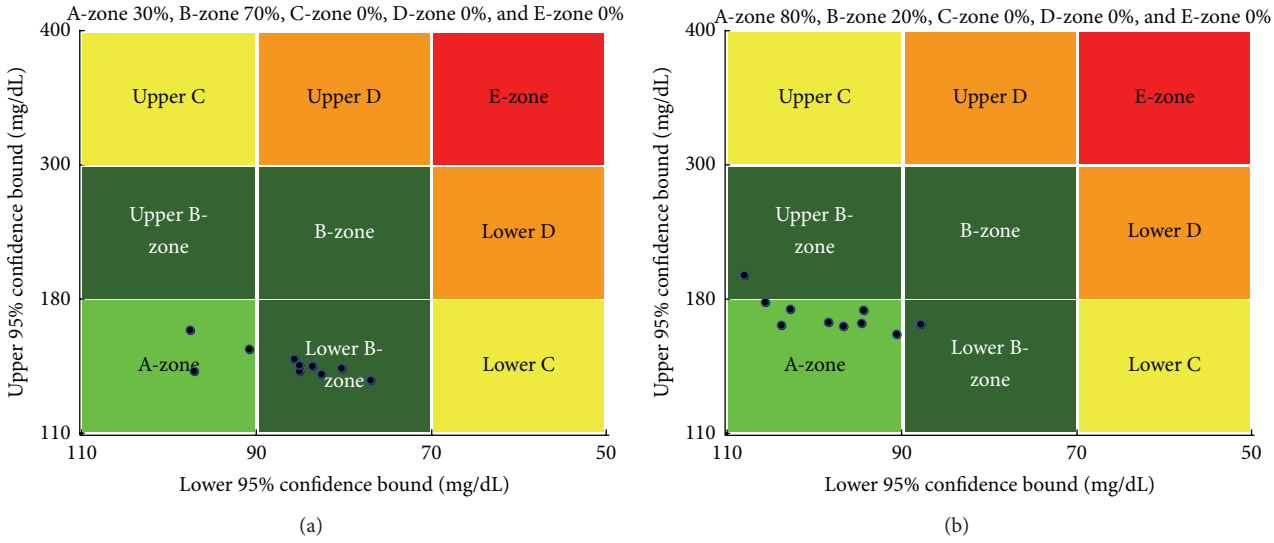


Figure 7: The control-variability grid analysis (CVGA) plot for the PID controller with (a) and without (b) IOB estimate.

Table 3: Results for blood glucose control under DIA (h) = 2 h, where insulin sensitivities are normal and changed.

Insulin sensitivity	Control algorithm	G_{max} (mg/dL)	G_{min} (mg/dL)	BGI	SD (mg/dL)
IS normal	With IOB limitation	214.8	72.6	4.2	42.2
	Without IOB limitation	210.7	41.8	9.8	48.7
IS changed	With IOB limitation	213.2	70.2	4.6	44.6
	Without IOB limitation	225.9	35.6	10.4	51.8

of CVGA would be the 95% confidence bound of a virtual patient's data.

The results indicate that 30% of virtual subjects are within A-zone and 70% of virtual subjects are within B-zone under the traditional PID controller as shown in Figure 7(a) and 80% of virtual subjects are within A-zone and 20% of virtual subjects are within B-zone under the PID controller with IOB estimate as shown in Figure 7(b). The results indicate that the PID controller with IOB estimate is effective and robust. The blood glucose can be regulated more accurately than the traditional PID controller. It is excellent performance in tight blood glucose control and avoiding the hypoglycemic.

5. Conclusions

An improved PID algorithm for blood glucose control is presented. The features of the proposed control algorithm are that the PID controller is based on the IOB estimate and the upper constraint. The control algorithm is evaluated using a combinational glucose-insulin mathematical model. The simulation results have demonstrated that the hypoglycemic events can be avoided and the glucose responses in a reasonable range under multimeal ingested and insulin sensitivity changed. The statistical results also indicate that the BGI and SD values are smaller compared with the traditional PID control. Based on the *in silico* test, the CVGA indicates that the proposed PID controller can regulate glucose in an accurate control range and reduce the risk of hypoglycemic. It is demonstrated to be very robust and effective. The simulations of this paper will provide useful theoretical basis for blood glucose control.

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 Fundamental Research Funds for the Central Universities, (Grant no. HIT. IBRSEM. 201307) and the program for Harbin City Science and Technology Innovation Talents of Special Fund Project (Grant no. 2014RFXXJ065).

References

[1] J. Li, Y. Kuang, and C. C. Mason, "Modeling the glucose-insulin regulatory system and ultradian insulin secretory oscillations with two explicit time delays," *Journal of Theoretical Biology*, vol. 242, no. 3, pp. 722–735, 2006.

[2] http://www.idf.org/.

[3] G. Marchetti, M. Barolo, L. Jovanovic, H. Zisser, and D. E. Seborg, "An improved PID switching control strategy for type 1 diabetes," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 3, pp. 857–865, 2008.

[4] H. Zisser, L. Robinson, W. Bevier et al., "Bolus calculator: a review of four 'smart' insulin pumps," *Diabetes Technology and Therapeutics*, vol. 10, no. 6, pp. 441–444, 2008.

[5] K. Mythreyi, S. C. Subramanian, and R. Krishna Kumar, "Nonlinear glucose-insulin control considering delays-Part II: control algorithm," *Control Engineering Practice*, vol. 28, no. 1, pp. 26–33, 2014.

[6] E. M. Watson, M. J. Chappell, F. Ducrozet, S. M. Poucher, and J. W. T. Yates, "A new general glucose homeostatic model using a proportional-integral-derivative controller," *Computer Methods and Programs in Biomedicine*, vol. 102, no. 2, pp. 119–129, 2011.

[7] H. Lee and B. W. Bequette, "A closed-loop artificial pancreas based on model predictive control: human-friendly identification and automatic meal disturbance rejection," *Biomedical Signal Processing and Control*, vol. 4, no. 4, pp. 347–354, 2009.

[8] L. Magni, D. M. Raimondo, C. Dalla Man, G. De Nicolao, B. Kovatchev, and C. Cobelli, "Model predictive control of glucose concentration in type I diabetic patients: an in silico trial," *Biomedical Signal Processing and Control*, vol. 4, no. 4, pp. 338–346, 2009.

[9] I. Y. S. Chávez, R. Morales-Menéndez, and S. O. M. Chapa, "Glucose optimal control system in diabetes treatment," *Applied Mathematics and Computation*, vol. 209, no. 1, pp. 19–30, 2009.

[10] M. Ottavian, M. Barolo, H. Zisser, E. Dassau, and D. E. Seborg, "Adaptive blood glucose control for intensive care applications," *Computer Methods and Programs in Biomedicine*, vol. 109, no. 2, pp. 144–156, 2013.

[11] A. G. Gallardo Hernández, L. Fridman, A. Levant et al., "High-order sliding-mode control for blood glucose: practical relative degree approach," *Control Engineering Practice*, vol. 21, no. 5, pp. 747–758, 2013.

[12] E. Ackerman, L. C. Gatewood, J. W. Rosevear, and G. D. Molnar, "Model studies of blood-glucose regulation," *The Bulletin of Mathematical Biophysics*, vol. 27, no. 1, pp. 21–37, 1965.

[13] R. N. Bergman, L. S. Phillips, and C. Cobelli, "Physiologic evaluation of factors controlling glucose tolerance in man. Measurement of insulin sensitivity and β-cell glucose sensitivity from the response to intravenous glucose," *Journal of Clinical Investigation*, vol. 68, no. 6, pp. 1456–1467, 1981.

[14] C. Cobelli and A. Mari, "Validation of mathematical models of complex endocrine-metabolic systems. A case study on a model of glucose regulation," *Medical and Biological Engineering and Computing*, vol. 21, no. 4, pp. 390–399, 1983.

[15] R. Hovorka, V. Canonico, L. J. Chassin et al., "Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes," *Physiological Measurement*, vol. 25, no. 4, pp. 905–920, 2004.

[16] C. Dalla Man, M. Camilleri, and C. Cobelli, "A system model of oral glucose absorption validation on gold standard data," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2472–2478, 2006.

[17] X. Gao, H. Ning, and Y. Wang, "Systematically *in silico* comparison of unihormonal and bihormonal artificial pancreas systems," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 712496, 10 pages, 2013.

[18] C. Ellingsen, E. Dassau, H. Zisser et al., "Safety constraints in an artificial pancreatic β cell: an implementation of model predictive control with insulin on board," *Journal of Diabetes Science and Technology*, vol. 3, no. 3, pp. 536–544, 2009.

[19] F. León-Vargas, F. Garelli, H. De Battista, and J. Vehí, "Postprandial blood glucose control using a hybrid adaptive PD controller with insulin-on-board limitation," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 724–732, 2013.

[20] B. P. Kovatchev, W. L. Clarke, M. Breton, K. Brayman, and A. McCall, "Quantifying temporal glucose variability in diabetes via continuous glucose monitoring: mathematical methods and clinical application," *Diabetes Technology & Therapeutics*, vol. 7, no. 6, pp. 849–862, 2005.

[21] L. Magni, D. M. Raimondo, C. Dalla Man et al., "Evaluating the efficacy of closed-loop glucose regulation via control-variability grid analysis," *Journal of Diabetes Science and Technology*, vol. 2, no. 4, pp. 630–635, 2008.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 549026, 8 pages
http://dx.doi.org/10.1155/2015/549026

Research Article

G2LC: Resources Autoscaling for Real Time Bioinformatics Applications in IaaS

Rongdong Hu,[1] Guangming Liu,[1,2] Jingfei Jiang,[1] and Lixin Wang[1]

[1] *School of Computer, National University of Defense Technology, Changsha 410073, China*
[2] *National Supercomputer Center, Tianjin 300457, China*

Correspondence should be addressed to Rongdong Hu; rongdonghu@nudt.edu.cn

Received 17 May 2015; Revised 22 June 2015; Accepted 23 June 2015

Academic Editor: Tao Huang

Cloud computing has started to change the way how bioinformatics research is being carried out. Researchers who have taken advantage of this technology can process larger amounts of data and speed up scientific discovery. The variability in data volume results in variable computing requirements. Therefore, bioinformatics researchers are pursuing more reliable and efficient methods for conducting sequencing analyses. This paper proposes an automated resource provisioning method, G2LC, for bioinformatics applications in IaaS. It enables application to output the results in a real time manner. Its main purpose is to guarantee applications performance, while improving resource utilization. Real sequence searching data of BLAST is used to evaluate the effectiveness of G2LC. Experimental results show that G2LC guarantees the application performance, while resource is saved up to 20.14%.

1. Introduction

With significant advances in high-throughput sequencing technologies and consequently the exponential expansion of biological data, bioinformatics encounters difficulties in analysis of vast amounts of data. The need for storing and processing large-scale genome data, easy access to analyses tools, and efficient data sharing and retrieval has presented significant challenges. At present, cloud computing is a promising solution to address these challenges. Cloud computing offers near-infinite amount of resources capacity at a competitive rate and allows users to obtain resources on demand with pay-as-you-go pricing model. The elasticity and enormous capacity of cloud make it possible for bioinformatics applications to return results in a real time way. It will help speed up the bioinformatics research process and relieve the storage pressure of massive data.

IaaS (Infrastructure as a Service), as one important form of cloud computing, mainly leverages the virtualization technology to create multiple VMs (Virtual Machines) on a physical host and can support rapid deployment of large-scale applications [1]. Cloud providers can reduce power consumption by consolidating various applications into a fewer number of physical hosts and switching idle hosts to low-power modes. However, virtualization also creates a new problem. The application performance relies on effective management of VM capacity. One essential requirement of cloud computing is providing reliable QoS defined in terms of SLA (Service Level Agreements). SLA violation will bring economic penalties to cloud providers. Therefore, they always strive to ensure the agreed performance of individual VM. It is nontrivial because of the complexity of applications, various resources usage patterns, shared underlying hardware infrastructure, and the performance correlation and interference among applications.

The focus of this work is on the performance of bioinformatics applications in IaaS. This work tries to take advantage of cloud elasticity to deal with changes in application loads. A resource autoscaling method, G2LC, is proposed to provide suitable processing power for bioinformatics application, to keep up with the changes of sequence length. The method is based on statistical learning load forecasting algorithm. Application performance is guaranteed by adjusting the forecasted results, while minimizing resource usage. Real traces data of BLAST, one of the most widely used bioinformatics programs for sequence searching, is used to evaluate

the effectiveness of G2LC. Experimental results show that G2LC can save more than 20% of the resources, while guaranteeing application performance.

The rest of this paper is organized as follows. Section 2 describes the background. Section 3 proposes the detailed design and implementation of G2LC and Section 4 presents the experimental evaluation. Section 5 examines the related work and Section 6 makes the conclusions.

2. Background

According to [2], Real Time Systems (RTS) are those whose correctness depends not only on the logical results but also on the time in which such results are produced. In this type of applications, completion or response is always constrained by time. Failing in accomplishing this requirement could result in serious implications. Depending on the flexibility of such constraints, real time applications are generally classified into hard, firm, and soft. Hard real time applications are those where the nonfulfillment of the time constraints leads to system failure. Firm real time applications have hard constraints, but they allow certain level of tolerance. In the case of soft real time applications, the nonfulfillment of deadlines degrades the performance of system but does not destroy it by failure or crash. In this study, bioinformatics applications will be considered as soft real time applications. Regardless of the sequence length change, the result is required to return as quickly as possible.

Cloud computing is inherently real time and more specifically soft real time. Most existing cloud applications have stringent timing and performance requirements, such as voice and object recognition, image and video retrieval, financial systems, log processing, advertisement placement, and personalized recommendations. These applications are becoming increasingly latency sensitive and operating under demanding workloads that require fast response, for which some violations of the timing constraints are acceptable. For example, if an email system responds slowly, users may switch to another service provider. While the failure of the system to respond quickly may lead to customer churn and reduction on service provider's profits, it does not cause any catastrophic consequences.

On the other hand, cloud computing is also very suitable for RTS. In order to accomplish the time constraints, RTS normally demands large amount of computing resources. Cloud computing can offer this scalability. The virtualization and the resulting decoupling of infrastructure and application offered by cloud make it possible to rapidly scale the infrastructure to meet the resource requirements of the real time applications. Popular social applications, such as Facebook and Twitter, make further enhancement to enable real time communication. Major search engines jump in real time war by providing real time search results. Force.com provides real time integration with external cloud services such as Amazon Web Services, Facebook, Google App Engine, and Twitter.

However, the advent of other critical cloud computing targets such as the improvement of cost efficiency is creating a challenging atmosphere to real time applications. Cloud providers require not only accomplishing performance and time constraints of the applications, but also improving the resources utilization of data centers. The objective is to increase their profits while QoS is guaranteed. This balance is fundamental for the real time cloud.

3. Method: G2LC

3.1. VM Vertical Scaling. In IaaS, applications share the underlying hardware by running in isolated VMs. Each VM, during its initialization, is configured with a certain amount of resources (such as CPU, memory, and disk). A key factor for improving utilization efficiency is resource provisioning. The objective of VM vertical scaling is to ensure that VM capacity is matched with the workload, while overprovisioning wastes costly resources and underprovisioning degrades application performance. Existing virtualization technologies can adjust the capacity of a live VM locally based on time division multiplexing to maximize the resource utilization, also referred to as VM resizing.

Implementation of any policy is accompanied by operating costs. Chen et al. [3] set up a simulated environment and perform a preliminary experiment to illustrate the VM vertical scaling effect on three types of applications (CPU-, memory-, and network I/O-intensive). Experimental results show that the application performance degradation during the VM vertical scaling is smaller than that during VM migration, and the time of performance degradation is also shorter. The VM vertical scaling avoids the unnecessary VM migration by reallocating the spare resources to the heavy-loaded VM in very short time. By comparison, although VM migration can solve the performance problem, it spends much more time to do VM transmission, which generates significant interferences to the other colocated applications. Particularly, the network-intensive application receives serious interference when doing the VM migration, which is because much of the network I/O is preempted by the migration.

This work will adopt VM vertical scaling to adjust the VM processing capacity according to load changes of application. In IaaS virtualization platform, there are many mature tools (such as Xen (http://xenproject.org/), KVM (http://www.linux-kvm.org/page/Main_Page), and VMware (http://www.vmware.com/)) available for system manager to perform the monitoring and vertical scaling operations. For instance, we can use the Xen *xm* command to collect the CPU utilization of VM and use the Xen credit scheduler to set the CPU capacity limit of VM for vertical scaling.

3.2. Load Forecasting. One essential requirement of a real time cloud is providing reliable QoS defined in terms of SLA. The sequences processed by bioinformatics applications often have very different lengths causing dynamic resources usage pattern. The consolidation of VMs can lead to performance degradation when an application encounters an increasing demand resulting in an unexpected rise of resources usage. This may lead to SLA violation—increasing response time. Overprovisioning may help to ensure SLA, but it leads to

inefficiency when the load decreases. The optimal strategy is to timely adjust resources provisioning according to the actual demands of the application. One precondition of this approach is to find out the future load.

In this work, we will adopt KSwSVR, proposed in our previous work, as our load prediction method [4]. It is based on statistical learning technology which is suitable for the complex and dynamic characteristics of the cloud computing environment. KSwSVR integrates an improved SVR (Support Vector Regression) algorithm and Kalman smoothing technology and does not require access to the internal details of application. The improved SVR gives more weight to more important data than standard SVR, using the historical information more reasonably. Kalman Smoother is employed to eliminate the noise of resources usage data coming from measurement error. In comparison with AR (Autoregression), BPNN (Back Propagation Neural Networks), and standard SVR, KSwSVR always has the minimum prediction error facing every type of resources and different predicted steps.

3.3. G2LC. We propose *G2LC* to improve the load forecasting-based resource autoscaling method with two adjustment mechanisms—*global gain* for predicted value and *local compensation* for error.

(i) SLA Definition. Resource utilization is commonly used by data center operators as a proxy for application performance because of the monotonic relationship between them and the fact that utilization is easily measurable at the OS level. In-depth researches on this relationship were also conducted, such as in [5, 6]. As it lies outside the sphere of our work, without loss of generality, we also adopt resource utilization to indicate QoS. In this study, SLA model of application is defined as follows:

$$\frac{1}{i-1}\sum_{j=1}^{i-1} F\left(x_j^{\text{alloc}} > x_j^{\text{use}}\right) = \text{cslaV}_i \le \text{slaV}$$
$$F(y) = \begin{cases} 0, & \text{if } y \text{ is true} \\ 1, & \text{if } y \text{ is false.} \end{cases} \quad (1)$$

x_j^{alloc} and x_j^{use} separately represent actual resources allocation value and real resources usage of application in time interval j. cslaV_i is the average SLA violation rate before interval i. It is the indication of the average QoS for VM and is restrained by slaV which is confirmed after the negotiation between cloud service providers and customers. slaV represents the user's tolerance for performance degradation. It is usually smaller than 5% for real time applications. As long as VM resource utilization is below 100%, that is, $x_j^{\text{alloc}} > x_j^{\text{use}}$, we judge that the resource is enough and there is no SLA violation.

Typically, users rent a VM with a certain capacity from IaaS providers. The resource configuration of the VM is unchanged at run time. This is currently a common practice, but it cannot effectively deal with the changing load. We take it as a comparison object in this work, and the resource saving rate is calculated as follows:

$$\text{resource saving rate} = \frac{\sum_{i=1}^{T}\left(x_{\text{Fixed},i} - x_{\text{G2LC},i}^{\text{alloc}}\right)}{\sum_{i=1}^{T} x_{\text{Fixed},i}}. \quad (2)$$

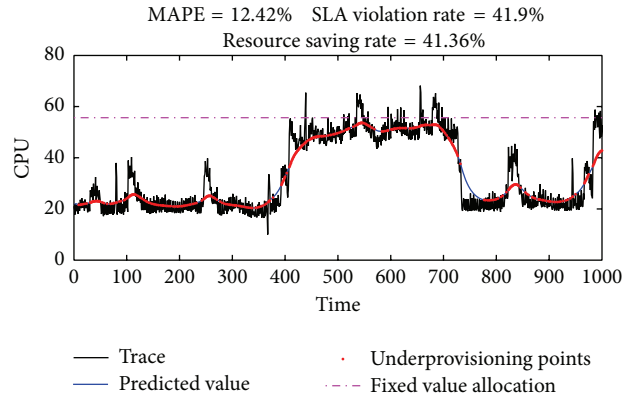


Figure 1: Resource autoscaling directly using predicted value.

(ii) Global Gain for Predicted Value. Even if the load prediction algorithm has high prediction accuracy, can we directly use the predicted value as the ultimate resource supply? We randomly select a piece of data from Google trace (https://code.google.com/p/googleclusterdata/) to analyze this issue. The data is linearly converted before test. Test results are shown in Figure 1.

If the resource allocated to VM is a fixed value, the CPU capacity must be more than 55.65 (horizontal dotted line) to keep the SLA violation rate below 5%. It can be seen from the figure that the loads continue to fluctuate, and the prediction accuracy of KSwSVR is acceptable. The MAPE (Mean Absolute Percentage Error) is only 12.42%. Compared with the fixed value allocation, the method based on the predicted value saves 41.36% of the resource usage in total. But the result of using the predicted value directly for the resource allocation is up to 41.9% SLA violation (red dots, underprovisioning points) rate, which is clearly unacceptable.

The root cause of this problem is the prediction error that any prediction algorithm cannot avoid. When the prediction target is variable, the error will be more notable. Therefore, we need to adjust the predicted value when making the VM scaling scheme.

Researchers from North Carolina State University, NetApp, and Google have studied the relationship between application performance and resource pressure (ratio of the total resource demand to the total resource allocation) [7]. They tested a web server and a database server. The result is shown in Figure 2. When the resource utilization of server exceeds 80%, application performance will seriously decline. If the target is to ensure the SLA violation rate less than 5%, the resource utilization of web servers and database servers must be kept, respectively, at 78% and 77% or less. The main reason is that current level of technology cannot effectively deal with load fluctuations. We need to provide a certain

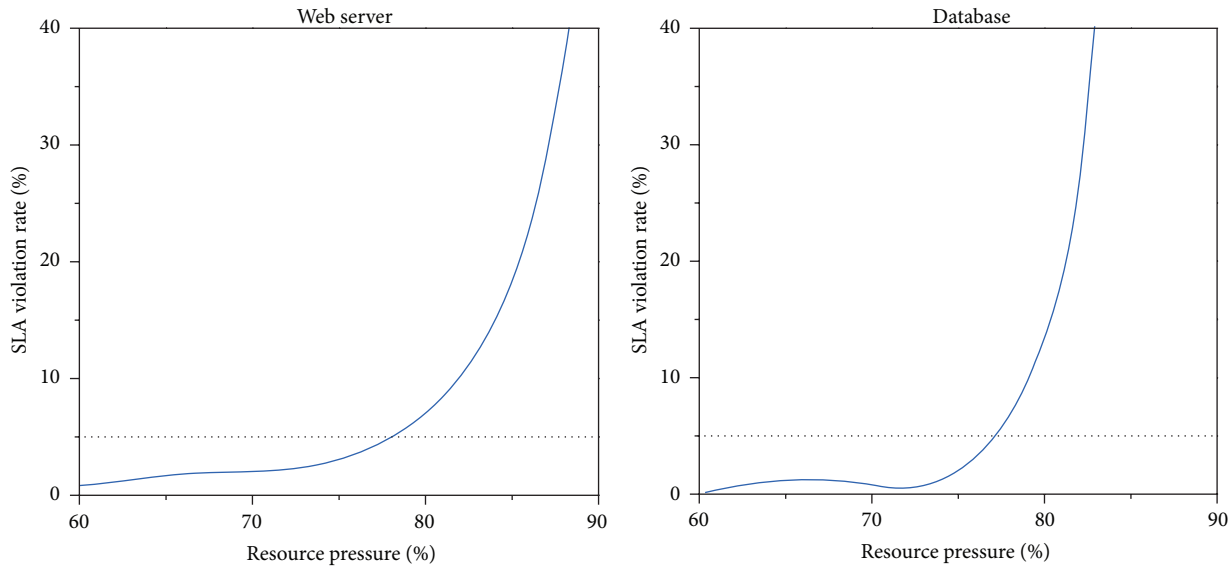


Figure 2: Relationship between application performance and resource pressure.

amount of redundant resources to maintain the application performance.

Inspired by this, we intend to adjust the forecast result $\widehat{x_t}$ by adding a gain coefficient $C_g > 1$, using more resources to meet the needs of real time applications:

$$x_t^{\text{alloc}} = C_g \widehat{x_t}. \tag{3}$$

We need to address the overprovisioning and underprovisioning problems in determining the amount of this part of redundant resources. For this purpose, we use an incremental traversal method to test the impact of the value of C_g on application performance and resource usage. The result is shown in Figure 3.

In the experiment, the value range of C_g is set to [1, 2]. In other words, the resource allocation amount increases from predicted value to its double. As C_g increases, more resources are added to application, and SLA violation rate decreases rapidly (*SLA violation rate*); that is, application performance is quickly enhanced. When $C_g = 1.3$, SLA violation rate drops below 5%. When $C_g > 1.6$, there is no SLA violation event. On the other hand, when C_g increases, the resource consumed by application also increases, and the cost advantage, relative to fixed value allocation, decreases linearly (*resource saving rate*). When C_g increases to 1.7, *resource saving rate* is reduced to zero; that is, the total resources usage of prediction-based dynamic allocation is quite equal to the one of fixed value allocation. If we continue to increase C_g, *resource saving rate* will become negative, and dynamic resource scaling will waste more resources.

If we only consider the resource utilization, the smaller the value of C_g, the better the management effect. However, to meet the application performance requirements, C_g should be set to 1.3 for the load in the experiment. That is, the resource utilization should be maintained at about $77\% \approx 1/1.3$. It is consistent with the conclusion of [7] cited before. We also randomly selected a number of other load data from

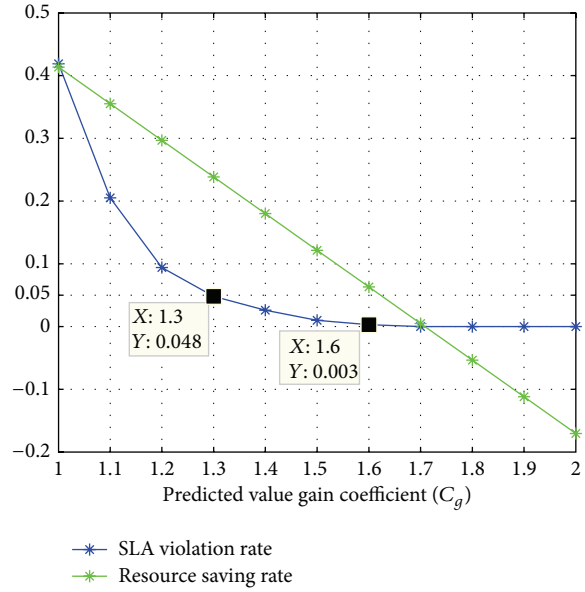


Figure 3: Impact of C_g on application performance and resource usage.

Google to test. The results showed that 1.3 is an ideal gain coefficient.

(iii) Local Compensation for Error. If the gain coefficient C_g is set to 1.3, the result of resource autoscaling based on load forecasting is shown in Figure 4.

As can be seen from the figure, after adding the gain coefficient C_g, the overall performance of the application is guaranteed well. Most of the time, because of the introduction of the gain coefficient, there is a significant gap between the resource provisioning curve and trace curve. This gap represents the wasted resource. SLA violation event mainly concentrated near the peak load, as it is difficult for prediction

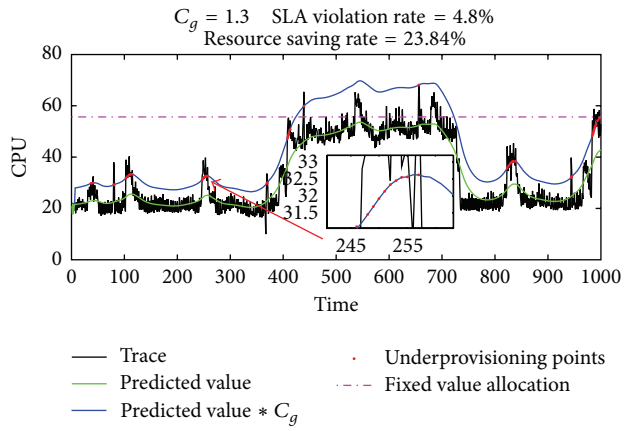


FIGURE 4: Resource autoscaling with gain coefficient C_g.

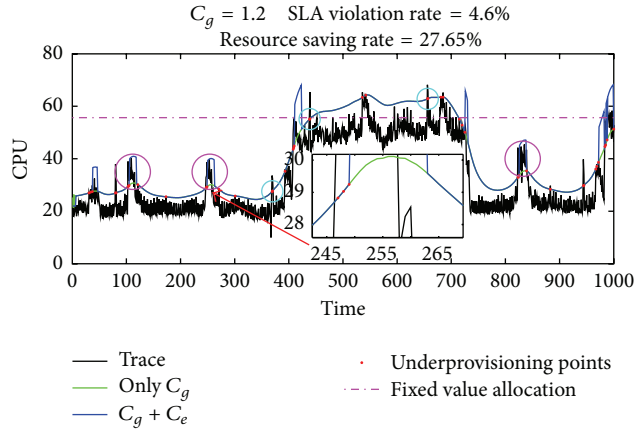


FIGURE 5: Resource autoscaling with C_g and C_e.

algorithm to deal with the temporary change of object. It is a problem that all current prediction algorithms cannot solve well. Therefore, predicted value needs postprocessing before being used. If we want to further improve the resource utilization, and also to ensure meeting application performance requirements, we need to amend the locality where the SLA violation events happen.

To this end, we further improve the resource scaling method and introduce a local error compensation mechanism to deal with the concentrated underprovisioning. The purpose is to reduce the SLA violation events as much as possible, providing space for further improving the resource utilization (by reducing gain coefficient C_g).

It should be noted that, in practice, IaaS service providers can perceive the VM underproviding based on VM resource utilization, but they cannot learn the specific deficiency. Therefore, we cannot directly use the difference between trace data and predicted value as the error compensation (which is actually the optimal solution).

We continue to introduce a local error compensation coefficient C_e based on (3):

$$x_t^{\text{alloc}} = C_e C_g \widehat{x_t}$$
$$C_e = \alpha^{\text{card}(v)-T_{\text{rd}}} \tag{4}$$
$$V = \left\{ i \mid x_{t-i}^{\text{use}} = x_{t-i}^{\text{alloc}},\ i \in \{1, 2, \ldots, w_e\} \right\},$$

where w_e is the windows width; that is, error compensation mechanism will take into account the resource usage of the past w_e periods to develop the resource scaling scheme of the next period. Because the VM resource usage cannot exceed the amount of its total resource, $x_{t-i}^{\text{use}} = x_{t-i}^{\text{alloc}}$ means lack of resource. V is a set of SLA violation events occurring in last w_e periods. card(V) denotes the elements number of set V. T_{rd} is a threshold value. Once the number of SLA violation events within the window exceeds the threshold, the error compensation mechanism will be triggered. $\alpha = 1.1$ is a constant; that is, error compensation amount increases at a rate of 10% each time. Figure 5 shows the dynamic resource autoscaling process under the combinational effect of global gain coefficient C_g and local error compensation coefficient C_e.

On the whole, compared with Figure 4, SLA violation rate reduces from 4.8% to 4.6%. Not only is the application performance improved slightly, but also more resource is saved. *Resource saving rate* (compared with fixed value allocation) increases from 23.84% (only C_g) to 27.65% (C_g + C_e). That is, the introduction of reasonable local error compensation mechanism creates space for reducing the gain coefficient C_g, while improving application performance and resource utilization.

Particularly, the most intuitive change generated by the reduction of C_g from 1.3 to 1.2 is the shrink of the gap between the resource provisioning curve and trace curve. In other words, less resource is wasted. Another significant change is that the resource provisioning curve is no longer as smooth as before. The introduction of local error compensation coefficient C_e and window w_e makes the resource management system respond rapidly to the load spikes (as shown in magenta circles). In addition, we introduce T_{rd} as a resource compensation mechanism trigger condition threshold, mainly to avoid the unnecessary compensation operation triggered by glitches (transient load peaks, as shown in cyan circles). It helps to improve resource utilization and enhances the system stability.

Another significant change is that although the number of SLA violation events (marked as red dots) does not reduce much, their distribution has changed a lot. In Figure 4, the SLA violation events concentrated near the peak load. In Figure 5, the red dots become decentralized. For end-user applications, they may encounter sporadic request response delay but will not suffer long time "fake system halt." This will help to improve the user experience.

So far, we have adjusted the predicted values at two levels: global gain and local compensation. The control process of resources autoscaling is shown in Figure 6. Monitor collects VM resources utilization data x_i^{use} and sends them to the predictor. Predictor predicts the resource consumption $\widehat{x_t}$ in the next control cycle. Finally, the resource scaling scheme x_t^{alloc} is figured out after the adjusting of global gain C_g and local compensation C_e.

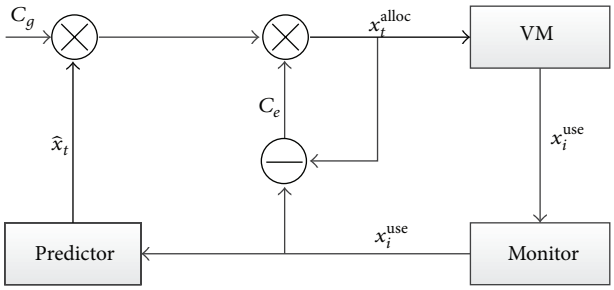


Figure 6: Control process of G2LC.

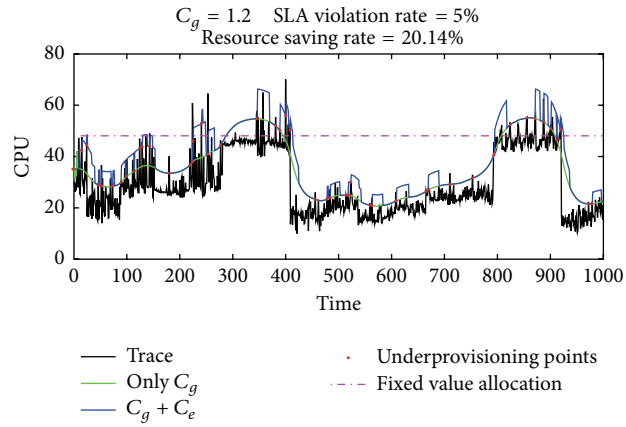


Figure 7: G2LC on BLAST trace.

4. Experiment with BLAST

Focusing on CPU utilization is a good way to understand the application performance, as it is typically proportional to the end-user productivity. Thus, CPU utilization can support greater transparency between cloud service providers and customers. Measuring and reporting CPU utilization is also a simple, affordable, and adequate way of gauging data center efficiency. Most importantly, many of the existing bioinformatics applications are compute-intensive applications. Hence, in this work, we focus on the CPU utilization of application.

It should be noted that the experiments in this paper mainly focus on CPU. So, the experimental conclusions surely apply to CPU-intensive applications. However, G2LC is also applicable to other types of applications (such as the memory-/disk-/network-intensive ones), because the existing virtualization technology can dynamically split these types of resources in a fine-grained way and the forecasting algorithm also applies to these resource objects.

4.1. Experiment Setup. BLAST (Basic Local Alignment Search Tool) [8] is one of the most widely used bioinformatics programs for sequence searching. It addresses a fundamental problem in bioinformatics research. BLAST is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences and identify library sequences that resemble the query sequence above a certain threshold. The heuristic algorithm it uses is much faster than other approaches, such as calculating an optimal alignment. This emphasis on speed is vital for making the algorithm practical on the huge genome databases currently available.

The effectiveness of G2LC is evaluated by using open real-world BLAST workload traces (http://ammatsun.acis.ufl.edu/amwiki/index.php/Prediction) rather than historical data generated by ourselves for the purpose of giving comparable and reproducible results.

The owners of the traces have comparatively assessed the suitability of several machine learning techniques for predicting spatiotemporal utilization of resources by BLAST [9]. They also extended Predicting Query Runtime (PQR) to the regression problem. BLAST was executed against the nonredundant (NR) protein sequence database from NCBI (National Center of Biotechnology Information). Given an input sequence, BLAST searches a database for similar sequences and calculates the best alignment of the matched sequences. Single nucleotide sequences of varying lengths served as input in the search process.

Different from their work focusing on run time prediction, this study is to guarantee the external performance of bioinformatics applications—returning the results in real time for different size of sequences (load). The traces provide the real search time of each sequence in nonvirtualization environments. If the search time of a sequence is longer, we believe it will need more computing resources in real time environment. Therefore, in the experiment, the search time attribute of sequence in traces is used as the load input. We expect to achieve a real time output by dynamic resource scaling.

4.2. Experimental Results. Experimental results are shown in Figure 7. With the change in the length of the sequence, the processing power of VM must be kept up with this change if we require BLAST to output the search result in a real time model. With the same parameters setting as before, G2LC not only guarantees the overall performance of BLAST, in this context, but also tries to minimize the gap between the resource provisioning curve and trace curve. The overall SLA violation rate is maintained at 5%. Compared with fixed value allocation with the same QoS, G2LC saved up to 20.14% of the resources.

To further analyze the effect of G2LC, we extract a small portion of the data to be described in detail, as shown in Figure 8. The global gain coefficient C_g makes the resource provisioning curve generally above the trace curve, guaranteeing the average performance of BLAST around the acceptable range. The introduction of local error compensation coefficient C_e makes the G2LC respond rapidly to the peak load growth (as shown in cyan circle). On the basis of C_g, C_e further reduces the probability of SLA violation event. In addition, the trigger threshold of resource compensation mechanism, T_{rd}, avoids the unnecessary compensation operation at spikes (as shown in magenta circles). It helps to save resource and enhances the system stability. But in some

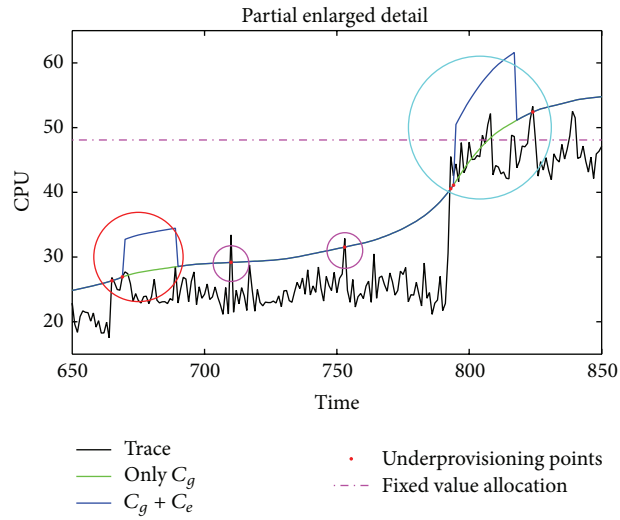


Figure 8: G2LC effect in detail.

cases, T_{rd} and window w_e will lead to a slight negative impact. As shown in red circle, once the number of SLA violation events in the window exceeds the threshold T_{rd}, the local compensation mechanism is triggered, regardless of whether the subsequent load increases. If subsequent load did not continue to increase, this will lead to a waste of resources. But its impact on the global effect is very slight, because the duration of resource waste cannot exceed the window width w_e.

5. Related Work

In this section, we briefly review some recent approaches on building and running bioinformatics applications on cloud platform.

As the field of bioinformatics expands, some researches have utilized cloud computing to deliver large computing capacity and on-demand scalability. Crossbow [10] is a cloud enabled tool that combines the aligner Bowtie and the SNP caller SOAPsnp and uses Hadoop for parallel computing. Rainbow [11] is a cloud-based software package that can assist in the automation of large-scale whole-genome sequencing (WGS) data analyses. It copies input datasets to Amazon S3 and utilizes Amazon's computing capabilities to run WGS data analyses pipelines. CloudMap [12] is a pipeline that greatly simplifies the analysis of mutant genome sequences from raw FASTQ reads to mapping plots and short lists of candidate mutations. CloudBurst [13] is a parallel read-mapping algorithm used for next-generation sequence data of the human genome and other reference genomes. It is implemented on a Hadoop-based cluster and aims to optimize the parallel execution. RSD-Cloud [14] runs a comparative genomics algorithm on Amazon EC2 for ortholog calculations across a wide selection of fully sequenced genomes. These projects focus on the solutions of specific problems by developing a tool or method.

Cloud BioLinux [15] is one of the early attempts to simplify the deployment and execution of bioinformatics applications on the cloud. It is a VM configured for high-performance bioinformatics using cloud platforms. At the beginning, over 135 bioinformatics tools have been deployed and configured on the VM. Li et al. [16] presented Hadoop-based applications employed in bioinformatics, covering next-generation sequencing and other biological domains. They described how to obtain an increase in performance by utilizing Hadoop on a cloud computing service and explored different alignment tools and applications that perform sequence alignment. Widera and Krasnogor [17] used Google App Engine computing platform as the computing resource. They introduced the method of building the computer generated protein models used in the protein structure prediction. The proposed Protein Models Comparator is their solution to the problem of large-scale model comparison and can be scaled for different data sizes. Hung and Hua [18] combined two different heterogeneous architectures, software architecture-Hadoop framework and hardware architecture-GPU, to develop a high performance cloud computing service, called Cloud-BLASTP, for protein sequence alignment. Cloud-BLASTP takes advantage of high performance, availability, reliability, and scalability. Cloud-BLASTP guarantees that all submitted jobs are properly completed, even when running job on an individual node or mapper experience failure.

Liu et al. [19] introduced a novel utility accrual scheduling algorithm for real time cloud computing services. The real time tasks are scheduled nonpreemptively with the objective to maximize the total utility. Two different time utility functions were proposed to model the real time applications for cloud computing that need not only to reward the early completions but also to penalize the abortions or deadline misses of real time tasks. Kim et al. [20] investigated power-aware provisioning of VMs for real time services. They modeled a real time service as a real time VM request and provisioned VMs using DVFS scheme. Several schemes were proposed to reduce power consumption by hard real time services and power-aware profitable provisioning of soft real time services.

In comparison to all these studies, G2LC is a general solution for bioinformatics applications to improve resource utilization in IaaS. It does not require access to the internal details of application and executes autoscaling scheme only based on the analysis of application resource utilization data. The purpose is to reduce service costs, while ensuring QoS at the same time.

6. Conclusions

With the rapid growth of next-generation sequencing technologies, more and more data have been discovered and published. To analyze such huge data, the computational performance becomes an important issue. The main focus of this work is on the performance of bioinformatics applications in IaaS. We try to take advantage of cloud elasticity to deal with changes in application loads, making it able to return a result in real time way. A resource autoscaling method, G2LC, is proposed to provide the right amount of resources,

to keep up with the changes of sequence length. A statistical learning-based algorithm is adopted for load forecasting. While minimizing resource usage, application performance is guaranteed by adjusting the forecasted results with global gain and local error compensation. Real BLAST trace data is used to evaluate the effectiveness of G2LC. Experimental results show that G2LC can save more than 20% of the resources, while guaranteeing application performance.

Conflict of Interests

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

Acknowledgment

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant no. 61303070.

References

[1] Z. Zhang, Z. Li, K. Wu et al., "VMThunder: fast provisioning of large-scale virtual machine clusters," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 12, pp. 3328–3338, 2014.

[2] J. A. Stankovic, "Misconceptions about real-time computing: a serious problem for next-generation systems," *Computer*, vol. 21, no. 10, pp. 10–19, 1988.

[3] W. Chen, X. Qiao, J. Wei, and T. Huang, "A two-level virtual machine self-reconfiguration mechanism for the cloud computing platforms," in *Proceedings of the 9th IEEE International Conference on Ubiquitous Intelligence & Computing (UIC '12) & 9th IEEE International Conference on Autonomic & Trusted Computing (ATC '12)*, pp. 563–570, IEEE, Fukuoka, Japan, September 2012.

[4] R. Hu, J. Jiang, G. Liu, and L. Wang, "Efficient resources provisioning based on load forecasting in cloud," *The Scientific World Journal*, vol. 2014, Article ID 321231, 12 pages, 2014.

[5] S. Kundu, R. Rangaswami, A. Gulati, M. Zhao, and K. Dutta, "Modeling virtualized applications using machine learning techniques," *ACM SIGPLAN Notices*, vol. 47, no. 7, pp. 3–14, 2012.

[6] H. Mi, H. Wang, Y. Zhou, M. R.-T. Lyu, and H. Cai, "Toward fine-grained, unsupervised, scalable performance diagnosis for production cloud computing systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 6, pp. 1245–1255, 2013.

[7] H. Nguyen, Z. Shen, and X. Gu, "Agile: elastic distributed resource scaling for infrastructure-as-a-service," in *Proceedings of the USENIX International Conference on Automated Computing (ICAC '13)*, San Jose, Calif, USA, June 2013.

[8] 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.

[9] A. Matsunaga and J. A. B. Fortes, "On the use of machine learning to predict the time and resources consumed by applications," in *Proceedings of the 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, pp. 495–504, IEEE Computer Society, Melbourne, Australia, May 2010.

[10] B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg, "Searching for SNPs with cloud computing," *Genome Biology*, vol. 10, no. 11, article R134, 2009.

[11] S. Zhao, K. Prenger, L. Smith et al., "Rainbow: a tool for large-scale whole-genome sequencing data analysis using cloud computing," *BMC Genomics*, vol. 14, no. 1, article 425, 2013.

[12] G. Minevich, D. S. Park, D. Blankenberg, R. J. Poole, and O. Hobert, "CloudMap: a cloud-based pipeline for analysis of mutant genome sequences," *Genetics*, vol. 192, no. 4, pp. 1249–1269, 2012.

[13] M. C. Schatz, "CloudBurst: highly sensitive read mapping with MapReduce," *Bioinformatics*, vol. 25, no. 11, pp. 1363–1369, 2009.

[14] D. P. Wall, P. Kudtarkar, V. A. Fusaro, R. Pivovarov, P. Patil, and P. J. Tonellato, "Cloud computing for comparative genomics," *BMC Bioinformatics*, vol. 11, no. 1, article 259, 2010.

[15] K. Krampis, T. Booth, B. Chapman et al., "Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community," *BMC Bioinformatics*, vol. 13, no. 1, article 42, 2012.

[16] X. Li, W. Jiang, Y. Jiang, and Q. Zou, "Hadoop applications in bioinformatics," in *Proceedings of the 7th IEEE Open Cirrus Summit (OCS '12)*, pp. 48–52, IEEE, Beijing, China, June 2012.

[17] P. Widera and N. Krasnogor, "Protein models comparator: scalable bioinformatics computing on the Google App Engine platform," http://arxiv.org/abs/1102.4293.

[18] C.-L. Hung and G.-J. Hua, "Local alignment tool based on Hadoop framework and GPU architecture," *BioMed Research International*, vol. 2014, Article ID 541490, 7 pages, 2014.

[19] S. Liu, G. Quan, and S. Ren, "On-line scheduling of real-time services for cloud computing," in *Proceedings of the 6th World Congress on Services (SERVICES '10)*, IEEE, Miami, Fla, USA, July 2010.

[20] K. H. Kim, A. Beloglazov, and R. Buyya, "Power-aware provisioning of virtual machines for real-time Cloud services," *Concurrency Computation Practice and Experience*, vol. 23, no. 13, pp. 1491–1505, 2011.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 462363, 12 pages
http://dx.doi.org/10.1155/2015/462363

Research Article

Identifying New Candidate Genes and Chemicals Related to Prostate Cancer Using a Hybrid Network and Shortest Path Approach

Fei Yuan,[1] You Zhou,[1] Meng Wang,[1] Jing Yang,[1] Kai Wu,[1] Changhong Lu,[2] Xiangyin Kong,[1] and Yu-Dong Cai[3]

[1]*Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences and Shanghai Jiao Tong University School of Medicine, Shanghai 200031, China*
[2]*Department of Mathematics, East China Normal University, Shanghai 200241, China*
[3]*College of Life Science, Shanghai University, Shanghai 200444, China*

Correspondence should be addressed to Xiangyin Kong; xykong@sibs.ac.cn and Yu-Dong Cai; cai_yud@126.com

Received 24 January 2015; Accepted 24 February 2015

Academic Editor: Lin Lu

Prostate cancer is a type of cancer that occurs in the male prostate, a gland in the male reproductive system. Because prostate cancer cells may spread to other parts of the body and can influence human reproduction, understanding the mechanisms underlying this disease is critical for designing effective treatments. The identification of as many genes and chemicals related to prostate cancer as possible will enhance our understanding of this disease. In this study, we proposed a computational method to identify new candidate genes and chemicals based on currently known genes and chemicals related to prostate cancer by applying a shortest path approach in a hybrid network. The hybrid network was constructed according to information concerning chemical-chemical interactions, chemical-protein interactions, and protein-protein interactions. Many of the obtained genes and chemicals are associated with prostate cancer.

1. Introduction

The prostate is a gland in the male reproductive system that surrounds the prostatic urethra and affects urinary function. Its secretion is a component of semen. Prostate cancer is a form of adenocarcinoma. Most prostate cancers grow slowly, while some grow relatively rapidly [1, 2]. In the early stage, some prostate cancer patients present no symptoms, while others display symptoms similar to benign prostatic hyperplasia. Advanced prostate cancer can spread to other parts of the body, including the bones and lymph nodes [3]. Prostate cancer can also affect sexual function, such as erection and ejaculation. It is the world's second most common cancer [1]. More than 80% of men will be diagnosed with prostate cancer by the age of 80 [4], but, due to its slow growth, most patients do not die from this disease.

Biopsy is necessary to confirm the diagnosis of prostate cancer. Ultrasound (US) and magnetic resonance imaging (MRI) can help determine whether the cancer has metastasized [2]. Prostate specific antigen (PSA) screening is widely used in the USA to diagnose prostate cancer at an earlier age and cancer stage [5]. Noninvasive detection methods are being developed, including detecting EN2 and PCA3 mRNA in the urine [6, 7]. BCL-2, Ki-67, and ERK5 may also be useful as markers [8–10]. Treatment options for prostate cancer include surgery, radiation therapy, hormone therapy, and chemotherapy [2].

Prostate cancer risk is associated with age, family disease history, and race. It is not monogenic; many genes are involved. For example, mutations in BRCA1 and BRCA2 have been implicated in prostate cancer, while they are also risk factors for ovarian cancer and breast cancer [11]. p53 mutations are more frequently observed after prostate cancer metastasis. Additionally, one copy of the tumor suppressor gene PTEN is lost in up to 70% of prostate cancer patients [12]. Genome-wide association studies have identified several

SNPs that affect prostate cancer risk [13–15]. The transcription factor RUNX2 can prevent prostate cancer cell apoptosis [16], and inhibition of X-linked inhibitor of apoptosis (XIAP) is being studied as a strategy to enhance apoptosis and prevent cancer cell proliferation [17]. Sexually transmissible infections (STI), such as HPV-16, HPV-18, and HSV-2, are significantly linked with prostate cancer [18–20].

Several chemicals have also been studied in prostate cancer. Zinc can change prostate cell metabolism to produce citrate, an important component of semen. This process requires a large amount of energy and prostate cancer cells that are devoid of zinc reserve energy for growth [21]. The prostate glands require androgens to work properly. Hormone therapies, including castration treatment (reduction of androgen/testosterone/DHT), are commonly used, but they are only effective in a subset of patients. Androgen receptor inhibition is effective in mouse studies [22]. More treatments are being tested to improve the survival of castration-resistant prostate cancer patients.

As discussed above, prostate cancer is a very complicated disease, and we have yet to identify all risk factors. Additional genes and chemicals remain to be discovered. While it is time consuming and expensive to identify genes or chemicals related to prostate cancer using traditional approaches, the development of computer science can overcome these obstacles by building effective computational methods. Here, we proposed an alternative computational method to identify new candidate genes and chemicals related to prostate cancer. To simultaneously investigate genes and chemicals, a hybrid network was constructed using chemical-chemical interactions and chemical-protein interactions from STITCH (search tool for interactions of chemicals) [23] and protein-protein interactions from STRING (search tool for the retrieval of interacting genes/proteins) [24]. By applying a shortest path approach in the hybrid network, we extracted genes and chemicals related to prostate cancer. To validate our model, several of the identified genes and chemicals were investigated in related prostate cancer literature.

2. Materials and Methods

2.1. Genes Related to Prostate Cancer. We collected genes related to prostate cancer using the following approaches: (I) 143 reviewed genes were chosen from UniProt (http://www.uniprot.org/, UniProt Release 2014_4) [25] using the search terms, "human," "prostatic cancer," and "reviewed"; (II) 86 genes were chosen from the TSGene Database (Tumor Suppressor Gene Database, http://bioinfo.mc.vanderbilt.edu/TSGene/cancer_type.cgi [26]) after the Entrez IDs were converted into their official symbols; and (III) 96 genes were retrieved from the NCI (National Cancer Institute, https://gforge.nci.nih.gov, released 2009.6) database [27]. After integrating the aforementioned 325 genes, we obtained 309 genes related to prostate cancer (Supplementary Material I; see Supplementary Material available online at http://dx.doi.org/10.1155/2015/462363).

2.2. Chemicals Related to Prostate Cancer. Chemicals related to prostate cancer were collected from the CTD (Comparative Toxicogenomics Database) (http://ctdbase.org/detail.go?type=disease&acc=MESH:D011471&view=chem, July 2014) [28]. These chemicals were manually assessed in the literature. Here, 177 chemicals with direct evidence of association with prostate cancer, such as "marker," "mechanism," or "therapeutic," were considered. Among these 177 chemicals, 106 were present in the hybrid network described below (see Section 2.3). Thus, we employed these 106 chemicals in this study (Supplementary Material I).

2.3. Hybrid Network. The hybrid network was constructed according to information based on chemical-chemical interactions, chemical-protein interactions, and protein-protein interactions. In brief, the chemical-chemical interactions and chemical-protein interactions were retrieved from STITCH (version 4.0, http://stitch.embl.de/) [23], and the protein-protein interactions were downloaded from STRING (version 9.1, http://www.string-db.org/) [24]. The obtained interactions include both known and predicted interactions. Thus, they can widely measure the associations between chemicals and proteins, and they have been widely used to investigate many chemical-related and protein-related problems [29–40]. In addition, to measure the strength of these interactions, each interaction was assigned a score in STITCH and STRING. The score of the chemical-chemical interaction between chemicals c_1 and c_2 was denoted by $S_{cc}(c_1, c_2)$, the score of the chemical-protein interaction between chemical c and protein p by $S_{cp}(c, p)$, and the score of the protein-protein interaction between proteins p_1 and p_2 by $S_{pp}(p_1, p_2)$. Due to the large number of chemicals, we only considered chemicals with KEGG (Kyoto Encyclopedia of Genes and Genomes) records [41] to reduce search space (i.e., chemicals occurring in the retrieved chemical-protein interactions and chemical-chemical interactions must be in KEGG).

The hybrid network used proteins and chemicals from the three types of interactions as nodes. Each edge represented one of the three types of interactions, and they were assigned a weight to indicate the strength of the interaction using the following equations:

$$w(e) = \begin{cases} 1000 - S_{pp}(p_1, p_2) & \text{If } n_1 \text{ and } n_2 \text{ represented proteins } p_1 \text{ and } p_2 \\ 1000 - S_{cp}(c, p) & \text{If } n_1 \text{ and } n_2 \text{ represented chemical } c \text{ and protein } p \\ 1000 - S_{cc}(c_1, c_2) & \text{If } n_1 \text{ and } n_2 \text{ represented chemicals } c_1 \text{ and } c_2. \end{cases} \quad (1)$$

Finally, we obtained a hybrid network consisting of 35,842 nodes, where 15,072 nodes represented chemicals and 20,770 nodes represented proteins. The size of the network, that is, the number of edges in the network, was 3,046,625, where 398,701 edges represented chemical-chemical interactions, 222,610 edges represented chemical-protein interactions, and 2,425,314 edges represented protein-protein interactions.

2.4. A Shortest Path Approach Used to Identify New Candidate Genes and Chemicals. Chemicals or proteins that comprise an interaction always have similar functions [31, 36, 42]. One chemical/protein and one chemical/protein that interact with a high score (low weight of the corresponding edge in the hybrid network) are more likely to share similar functions than those with a low score. Therefore, we can infer that chemicals/proteins occurring in a shortest path connecting the chemicals/proteins, n_1 and n_2, are likely to share functions with n_1 and n_2. Thus, we searched all the shortest paths connecting any pair of chemicals and proteins related to prostate cancer, and the corresponding chemicals and proteins occurring in these paths were considered candidate chemicals and genes. Simultaneously, the number of paths containing a certain candidate chemical or gene was termed "betweenness."

Some of the candidate chemicals and genes may be false positives, and some chemicals or proteins may have universal associations with other chemicals or proteins, so they are observed in the shortest paths connecting any pair of randomly selected chemicals or proteins. To control for these false positives, we randomly produced 1,000 chemical and protein sets, and each set had the same numbers of chemicals and proteins as the set consisting of chemicals and genes related to prostate cancer. For each set, we searched for the shortest paths connecting any pair of chemicals or proteins and counted the betweenness of the candidate chemicals and proteins based on these paths. Then, we counted the number of randomly produced sets in which the betweenness was larger than the set consisting of chemicals and genes related to prostate cancer for each candidate chemical or gene; the P value was defined as the aforementioned number divided by 1,000. Thus, a low P value for a certain candidate chemical or gene indicates strong linkage with prostate cancer.

3. Results and Discussion

3.1. Candidate Genes and Chemicals. As mentioned in Sections 2.1 and 2.2, we employed 309 genes and 106 chemicals related to prostate cancer. We searched all shortest paths connecting any of these genes. Based on the obtained paths, we extracted 595 candidate genes and 102 candidate chemicals and calculated their betweenness (Supplementary Material II). According to the method in Section 2.4, the P values of these candidate genes and chemicals were computed to control for false positives, which are also listed in Supplementary Material II. Then, we set the P value threshold as 0.05 to select for significant candidate genes and chemicals (i.e., candidate genes and chemicals with P values less than 0.05 were selected). Ultimately, 187 genes and 11 chemicals were selected (Supplementary Material III).

3.2. Analysis of Enriched KEGG Pathways of Significant Candidate Genes. As mentioned in Section 3.1, we obtained 187 significant candidate genes that were potentially related to prostate cancer pathogenesis. To analyze the relationship between these genes and prostate cancer, we employed a functional annotation tool, DAVID (Database for Annotation, Visualization and Integrated Discovery) [43], to understand their biological significance. The results of DAVID included the enrichment of the 187 significant candidate genes in KEGG pathways and GO terms (Supplementary Material IV and V, resp.).

In total, the 187 significant candidate genes shared 40 KEGG pathways. After sorting the 40 KEGG pathways according to their FDR (false discovery rate) adjusted P value (last column in Supplementary Material IV), we found that the top six pathways were highly associated with prostate cancer. Figure 1 shows these pathways, the number of genes among the 187 significant candidate genes that shared each pathway and the proportion of these genes among all genes sharing the pathway. Table 1 lists the FDR of these pathways.

The most enriched pathway was hsa05200: pathways in cancer, with 30 significant candidate genes sharing this pathway (see Figure 1) and an FDR of $2.08E-06$ (see Table 1, row 2). The fourth most enriched pathway was hsa05214: glioma, with 10 significant candidate genes sharing this pathway (see Figure 1) and an FDR of $3.03E-02$ (see Table 1, row 5). These results indicate that prostate cancer and other types of cancer share a common mechanism.

The second most enriched pathway was hsa04010: MAPK signaling pathway, with 27 significant candidate genes (see Figure 1) and an FDR of $2.15E-06$ (see Table 1, row 3). Mitogen-activated protein kinase (MAPK) pathways are evolutionarily conserved and link extracellular signals to fundamental cellular processes. Mutations in these pathways can affect Ras and B-Raf and play a critical role in cancer development [44].

The third most enriched pathway was hsa05215: prostate cancer, with 12 significant candidate genes (see Figure 1) and an FDR of $1.56E-02$ (see Table 1, row 4). This result shows that some of the candidate genes have already been grouped into the pathway which was drawn based on the previous knowledge of molecular interaction and reaction networks in prostate cancer.

The fifth most enriched pathway was hsa04722: neurotrophin signaling pathway, with 13 significant candidate genes (see Figure 1) and an FDR of $7.59E-02$ (see Table 1, row 6). Neurotrophins play a role in the survival of malignant prostate cells [45]. Neurotrophins include nerve growth factor (NGF), brain-derived neurotrophic factor (BDNF), neurotrophin 3 (NT-3), and neurotrophin 4/5 (NT4/5), and they bind with trk receptors. The survival of malignant prostate cells requires ectopic expression of trk B and trk C and continued expression of trk A. Trk inhibition has been suggested to be a drug therapeutic target [46].

The sixth most enriched pathway was hsa04310: Wnt signaling pathway, with 14 significant candidate genes (see Figure 1) and an FDR of $1.26E-01$ (see Table 1, row 7). The Wnt signaling pathway is involved in carcinogenesis and embryonic development. It acts as a common element in the regulation of stem cell renewal and the maintenance of many cellular systems. Disruption of this pathway is associated with cancer [47]. Mutations in components of this pathway, including APC, Axin, Axin2/conduction, and β-catenin, are found in a variety of cancers [48]. The Wnt signaling pathway plays a critical role in prostate cancer, as

Table 1: The top six KEGG pathways shared by 187 significant candidate genes.

Pathway ID	Pathway name	Genes sharing the pathway	FDR
hsa05200	Pathways in cancer	FGF6, FGFR2, TRAF2, FGFR1, PDGFB, WNT3A, MITF, NFKB1, TGFB1, CTNNB1, GLI1, MAX, WNT1, CASP3, RARA, HHIP, AXIN1, PIK3R2, CREBBP, CDK6, BIRC2, RALGDS, CCND1, PLCG1, NTRK1, MAPK3, PDGFRB, PTCH1, IKBKB, and GSTP1	$2.08E-06$
hsa04010	MAPK signaling pathway	FGFR2, FGF6, TRAF2, FGFR1, PDGFB, PPP3R1, NFKB1, TGFB1, ATF2, MAP3K7, MAX, TNFRSF1A, CASP3, MAP3K5, MAP3K3, MAP2K6, RASA1, FLNA, MAPK14, NTRK1, GADD45G, MAPK3, PDGFRB, HSPB1, IKBKB, CD14, and DUSP6	$2.15E-06$
hsa05215	Prostate cancer	FGFR2, FGFR1, CCND1, PDGFB, MAPK3, CREBBP, PDGFRB, NFKB1, IKBKB, GSTP1, CTNNB1, and PIK3R2	$1.56E-02$
hsa05214	Glioma	CCND1, PLCG1, PDGFB, CAMK2G, MAPK3, PDGFRB, CDK6, SHC1, CALM2, and PIK3R2	$3.03E-02$
hsa04722	Neurotrophin signaling pathway	CAMK2G, NFKB1, MAGED1, MAP3K5, MAP3K3, PLCG1, NTRK1, MAPK14, MAPK3, SHC1, IKBKB, CALM2, and PIK3R2	$7.59E-02$
hsa04310	Wnt signaling pathway	ROCK1, WNT3A, CAMK2G, CREBBP, CSNK2B, PPP3R1, CTNNB1, MAP3K7, WNT1, CCND1, CSNK1E, LRP6, LRP5, and AXIN1	$1.26E-01$

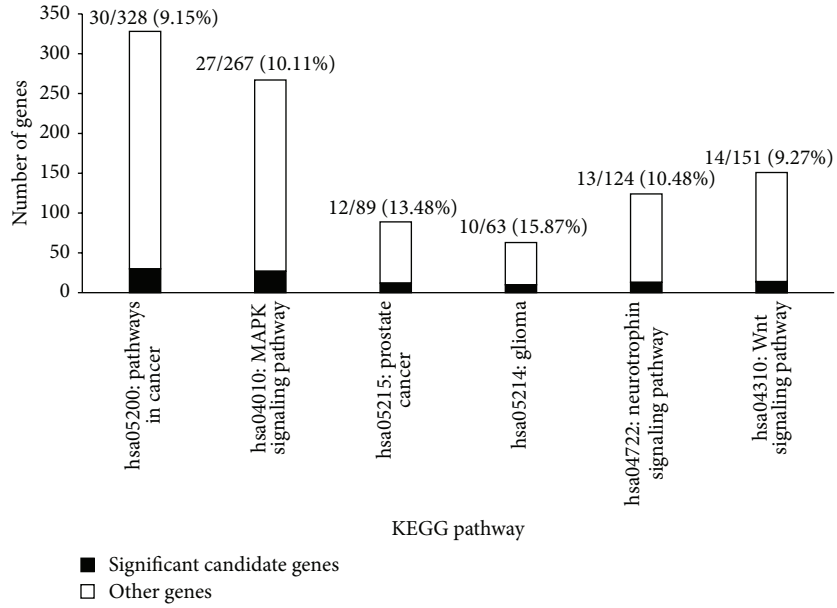


Figure 1: Top six pathways highly associated with prostate cancer analyzed by DAVID. The black part represents the number of significant candidate genes sharing the pathway; the white part represents the number of other genes sharing the pathway.

its key component, β-catenin, works as an androgen receptor (AR) cofactor. β-Catenin can significantly enhance androgen-stimulated transcriptional activation by the AR [49]. Abnormal expression of Wnt ligands and receptors may also contribute to the pathogenesis of prostate cancer [50].

3.3. Analysis of Enriched GO Terms of Significant Candidate Genes. In total, the 187 significant candidate genes enriched 576 GO terms (Supplementary Material V), and we investigated the top ten GO terms sorted by FDR. Figure 2 shows these GO terms, the number of genes among the 187 significant candidate genes that shared each GO term and the proportion of these genes among all genes sharing the GO term. Table 2 lists the FDR of these GO terms.

All of these ten GO terms were biological process (BP) GO terms, and four were associated with the regulation of cell proliferation and death: GO:0042127, regulation of cell proliferation (39 significant candidate genes sharing this GO term, refer to Figure 2) ("FDR" = $2.11E-09$, refer to Table 2); GO:0042981, regulation of apoptosis (35 significant candidate genes sharing this GO term, refer to Figure 2) ("FDR" = $1.38E-06$, refer to Table 2); GO:0043067, regulation of programmed cell death (35 significant candidate genes sharing this GO term, refer to Figure 2) ("FDR" = $1.79E-06$,

Table 2: The top ten GO terms shared by 187 significant candidate genes.

GO term ID	GO term	Genes sharing the GO term	FDR
GO:0042127	Regulation of cell proliferation	FGFR2, FGFR1, CCL2, PDGFB, NDN, MITF, STRN, GNRHR, VIPR1, FOXO4, GHRHR, TGFB1, CTNNB1, GLI1, MAGED1, CTTNBP2, VDR, CASP3, MYOCD, SFTPD, SHC1, MUC2, PTGER2, GNRH1, CDK6, LIG4, DBH, NTN1, CDKN1C, PRKCQ, CCND1, HNF4A, HGS, TGFBR3, PDGFRB, PTCH1, SST, ADRA1D, and LRP5	$2.11E-09$
GO:0010033	Response to organic substance	CGA, CCL2, PDGFB, LHCGR, NR3C1, FOXO4, TGFB1, GHRHR, CTNNB1, B2M, CTTNBP2, TNFRSF1A, CASP3, REN, RARA, SHC1, KCNMA1, GNRH1, CSNK2B, ESR1, DBH, BIRC2, PRKCQ, CCND1, HNF4A, MAPK14, ALDH2, HSD11B2, HSPB1, TGFBR3, PTCH1, IRF3, SST, and CD14	$3.38E-07$
GO:0042981	Regulation of apoptosis	TRAF2, C9, CCL2, MITF, PPP3R1, NFKB1, RRM2B, NR3C1, TGFB1, MAGED1, MAP3K7, VDR, BAK1, MAP3K5, CASP3, NQO1, TERT, MAP2K6, RASA1, TERF1, KCNMA1, MUC2, GNRH1, ROCK1, ESR1, LIG4, DBH, BIRC2, TNFRSF10B, NTRK1, UBC, HSPB1, IKBKB, SST, and GSTP1	$1.38E-06$
GO:0043067	Regulation of programmed cell death	TRAF2, C9, CCL2, MITF, PPP3R1, NFKB1, RRM2B, NR3C1, TGFB1, MAGED1, MAP3K7, VDR, BAK1, MAP3K5, CASP3, NQO1, TERT, MAP2K6, RASA1, TERF1, KCNMA1, MUC2, GNRH1, ROCK1, ESR1, LIG4, DBH, BIRC2, TNFRSF10B, NTRK1, UBC, HSPB1, IKBKB, SST, and GSTP1	$1.79E-06$
GO:0010941	Regulation of cell death	TRAF2, C9, CCL2, MITF, PPP3R1, NFKB1, RRM2B, NR3C1, TGFB1, MAGED1, MAP3K7, VDR, BAK1, MAP3K5, CASP3, NQO1, TERT, MAP2K6, RASA1, TERF1, KCNMA1, MUC2, GNRH1, ROCK1, ESR1, LIG4, DBH, BIRC2, TNFRSF10B, NTRK1, UBC, HSPB1, IKBKB, SST, and GSTP1	$1.97E-06$
GO:0009719	Response to endogenous stimulus	KCNMA1, CGA, CCL2, GNRH1, PDGFB, LHCGR, ESR1, FOXO4, DBH, BIRC2, TGFB1, GHRHR, CTNNB1, CTTNBP2, PRKCQ, CCND1, REN, ALDH2, TGFBR3, HSD11B2, RARA, SHC1, PTCH1, and SST	$4.75E-06$
GO:0016477	Cell migration	ICAM1, CCL2, ROCK1, PDGFB, NDN, NUP85, CDH2, CX3CL1, DBH, NTN1, TGFB1, CTTNBP2, WNT1, CKLF, LRP6, SFTPD, TGFBR3, PDGFRB, SCNN1B, and LRP5	$5.61E-06$
GO:0007242	Intracellular signaling cascade	TRAF2, FGFR1, CYP24A1, CCL2, LHCGR, NR3C1, VIPR1, FOXO4, GHRHR, CTNNB1, MAP3K7, VDR, MAP3K5, MAP3K3, REN, RARA, SHC1, RASA1, MAP2K6, CNKSR1, CCM2, ROCK1, ESR1, RALGDS, FLNA, PRKCQ, CCND1, NCOA1, TNFRSF10B, PLCG1, NEDD4, MAPK14, NTRK1, KRIT1, GADD45G, MAPK3, RAB5A, TGFBR3, IRF3, IKBKB, ADRA1D, GRB14, DUSP6	$1.26E-05$
GO:0009725	Response to hormone stimulus	KCNMA1, CGA, CCL2, GNRH1, PDGFB, LHCGR, ESR1, FOXO4, GHRHR, TGFB1, CTNNB1, CTTNBP2, PRKCQ, CCND1, REN, ALDH2, TGFBR3, HSD11B2, RARA, SHC1, PTCH1, SST	$2.16E-05$
GO:0048870	Cell motility	ICAM1, CCL2, ROCK1, PDGFB, NDN, NUP85, CDH2, CX3CL1, DBH, NTN1, TGFB1, CTTNBP2, WNT1, CKLF, LRP6, SFTPD, TGFBR3, PDGFRB, SCNN1B, LRP5	$3.20E-05$

refer to Table 2); and GO:0010941, regulation of cell death (35 significant candidate genes sharing this GO term, refer to Figure 2) ("FDR" = $1.97E-06$, refer to Table 2). Cell proliferation and apoptosis are both important biological processes that may lead to cancer if altered by gene mutation and other risk factors. An increasing number of studies have demonstrated that important genes and miRNAs that participate in these processes could be therapeutic targets. For instance, miR-145 functions as a tumor suppressor. By targeting FSCN1, miR-145 suppresses cell proliferation in prostate cancer, and it represents an important therapeutic target [51].

Three GO terms were associated with cell responses to stimulus: GO:0010033, response to organic substance (34 significant candidate genes sharing this GO term, refer to Figure 2) ("FDR" = $3.38E-07$, refer to Table 2); GO:0009719, response to endogenous stimulus 24 significant candidate genes sharing this GO term, refer to Figure 2) ("FDR" = $4.75E-06$, refer to Table 2); and GO:0009725, response to hormone stimulus (22 significant candidate genes sharing this GO term, refer to Figure 2) ("FDR" = $2.16E-05$, refer to Table 2). Sex hormones play an important role in the growth and development of the prostate [52]. Testosterone is implicated in the pathogenesis of prostate cancer [53]. Hormone therapy is currently used in the clinical treatment of prostate cancer, but it is only effective in a subset of patients. A recent study found no association between prediagnostic circulating sex hormones and lethal prostate cancer or total

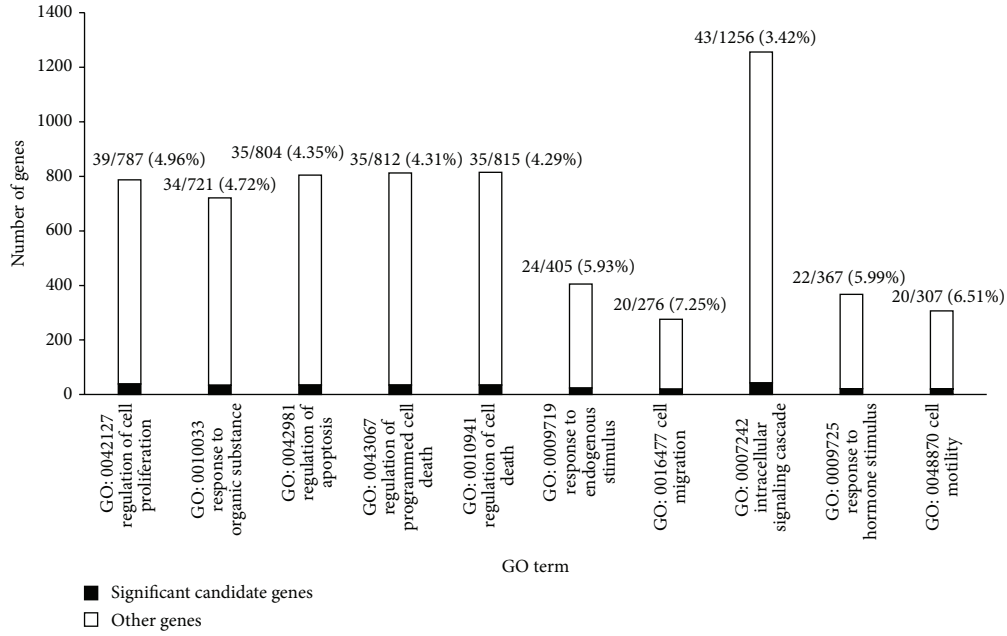


Figure 2: Top ten GO terms highly related to prostate cancer analyzed by DAVID. The black part represents the number of significant candidate genes sharing the GO term; the white part represents the number of other genes sharing the GO term.

mortality [54]. This topic remains debatable, and further prospective studies are needed. Small chemicals that can stimulate prostate cells also warrant further attention.

Two GO terms were associated with cell motility: GO:0016477, cell migration (20 significant candidate genes sharing this GO term, refer to Figure 2) ("FDR" = $5.61E-06$, refer to Table 2), and GO:0048870, cell motility (20 significant candidate genes sharing this GO term, refer to Figure 2) ("FDR" = $3.20E-05$, refer to Table 2). Metastatic prostate cancer often spreads to bone, but the lung and liver are also common sites. More symptoms may occur depending on the site of cancer spread.

The last term was GO:0007242: intracellular signaling cascade (43 significant candidate genes sharing this GO term, refer to Figure 2) ("FDR" = $1.26E-05$, refer to Table 2). A recent report demonstrated that activation of Stat3 signaling was essential for prostate cancer progression, and inhibition of this pathway may be a therapeutic strategy [55]. Downregulation of Notch-1 and Jagged-1 could inhibit prostate cancer cell growth, migration and invasion, and induce apoptosis via inactivation of the Akt, mTOR, and NF-κB signaling pathways [56].

3.4. Analysis of Significant Candidate Genes. In our study, 187 significant candidate genes were obtained (Supplementary Material III), where 42 genes were with P value 0. Among these 42 genes, 21 genes were found to be reported as prostate cancer related genes in some previous studies, which implies our method is quite effective. Please see Table 3 for the detailed information of these 21 genes. For the rest 21 significant candidate genes with P value 0, four of them (listed in rows 2–5 of Table 4) were deemed to be related to prostate cancer based on their current validated functions. They were discussed as below.

Table 3: 21 significant candidate genes with P value 0 which have been reported to be related to prostate cancer in previous studies.

Gene ID	Gene name	Betweenness	P value	Supporting references
ENSP00000320940	NCOA1	3886	0	[99]
ENSP00000262367	CREBBP	3569	0	[100]
ENSP00000340858	B2M	2564	0	[101]
ENSP00000287641	SST	1085	0	[102]
ENSP00000410294	FGFR2	1085	0	[103]
ENSP00000346294	S100A4	365	0	[104]
ENSP00000226413	GNRHR	363	0	[105]
ENSP00000263408	C9	363	0	[106]
ENSP00000264001	CKLF	363	0	[107]
ENSP00000293308	KRT8	363	0	[108]
ENSP00000294954	LHCGR	363	0	[109]
ENSP00000298772	TRIM13	363	0	[110]
ENSP00000330382	PDGFB	363	0	[111]
ENSP00000348775	ACOX3	363	0	[112]
ENSP00000361366	SFTPD	363	0	[113]
ENSP00000382166	CX3CR1	363	0	[114]
ENSP00000413720	CDKN1C	363	0	[115]
ENSP00000216862	CYP24A1	36	0	[116]
ENSP00000420168	GSTA2	20	0	[117]
ENSP00000276431	TNFRSF10B	11	0	[118]
ENSP00000263946	PKP1	1	0	[119]

PLCG1. PLCG1 (phospholipase C, gamma 1) encodes the enzyme required to catalyze the formation of inositol IP3

Table 4: Information regarding significant candidate genes and chemicals related to prostate cancer.

Gene or chemical ID	Gene or chemical name	Betweenness	*P* value
ENSP00000244007	PLCG1	2,110	0
ENSP00000227758	BIRC2	1,583	0
ENSP00000215479	AMELY	363	0
ENSP00000262809	ELL	363	0
CID000002519	Caffeine	371	0.028
CID000005566	Trifluoperazine	363	0.001
CID000060662	Mibefradil	363	0.013
CID000161930	Icilin	363	0
CID000065036	Allicin	2	0.024

(1,4,5-trisphosphate) and DAG (diacylglycerol) from phosphatidylinositol 4,5-bisphosphate. In this process, IP3 uses Ca^{2+} as a cofactor for nuclear translocation and the subsequent activation of downstream targets [57]. In our study, PLCG1 was highly related to prostate cancer, as demonstrated by its high betweenness (2,110; see row 2 of Table 4) and low *P* value (0; see row 2 of Table 4). Frequent mutations occur in the catalytic domain of PLCG1, which induce the activation of downstream signaling pathway and PLCG1 was sensitive to specific inhibition of CaN in CTCL (cutaneous T-cell lymphoma) [58]. Many receptors, such as EGF (epidermal growth factor) and PDGF (platelet-derived growth factor), are affected by PLCG1 [59, 60]. In addition, PLCG1 plays a key role in chemotaxis triggered by growth factor receptors, and it is involved in integrin-dependent cell motility in diverse types of cancer [61]. Research regarding the function of PLCG1 in prostate cancer is rare; we remind that PLCG1 is a diagnostic marker and a drug target in prostate cancer.

BIRC2. BIRC2 (baculoviral IAP repeat containing 2), also known as API1 or cIAP1 (cellular inhibitors of apoptosis), belongs to a protein family that binds TRAF1/2 (tumor necrosis factor receptor-associated factors) to inhibit apoptosis. In our study, BIRC2 was closely associated with human prostate cancer, and its betweenness and *P* value were 1,583 and 0, respectively (see row 3 of Table 4). ARC (caspase recruitment) regulates BIRC2, and BIRC2 expression is inverse to ARC in AML (acute myeloid leukemia) [62, 63]. In addition, in metastatic human colon and breast cancer cells, BIRC2 is the molecular target of ceramide, and the Smac mimetic, BV6, targets BIRC2 to induce apoptosis via the TNFα signaling pathway [64, 65]. However, the detailed mechanism of BIRC2 action remains unknown. We speculate that BIRC2 is a key apoptosis-associated factor in prostate cancer that warrants further experimentation.

AMELY. In prostate cancer, many driver genes are gender-related. In our study, a gender-related locus gene, AMELY (amelogenin Y-linked) (betweenness: 363, *P* value: 0; see row 4 of Table 4), was related to prostate cancer. AMELY, which belongs to the amelogenin family of extracellular matrix proteins, is a single copy gene locus on the Y chromosome (Yp11.2) [66, 67]. AMELY and its homolog, AMELX, are often used for gender identification [68]. Deletions of AMELY occur frequently in certain ethnic populations [69–71]. Research regarding AMELY function is rare, especially in human prostate cancer, but we believe that it may be a potential gender-related gene and a biomarker in human prostate cancer. In the future, more experiments and clinical samples are still needed to validate the importance of this gene in prostate cancer.

ELL. ELL, the eleven-nineteen lysine-rich leukemia gene, encodes an RNA polymerase II transcription elongation factor that suppresses transient pausing by RNA polymerase II and functions in the process of transcription [72–74]. ELL was significantly associated with prostate cancer, as demonstrated by its high betweenness (363, see row 5 of Table 4) and low *P* value (0, see row 5 of Table 4). ELL was initially identified as a partner gene fused to MLL in the t(11;19) (q23; p13.1) translocation in AML (acute myeloid leukemia) [75]. U19/Eaf2 is an androgen-response gene that forms nuclear speckles by binding to ELL *in vivo*. U19/Eaf2 is downregulated in human prostate cancer, and its overexpression induces prostate cancer cell apoptosis [76]. Direct evidence regarding the function of ELL in human prostate cancer is rare, but our data and previous studies suggest that ELL is an inducer of apoptosis and a putative target in human prostate cancer.

Besides, significant candidate genes that were not discussed here still may be related to prostate cancer. We listed them in Supplementary Material III and hope that they will be the useful information for further study on prostate cancer.

3.5. Analysis of Significant Candidate Chemicals. We also obtained 11 significant candidate chemicals involved in prostate cancer (Supplementary Material III). This section discusses the relationships between several candidate chemicals and prostate cancer. Information pertaining to the discussed chemicals is listed in rows 6–10 of Table 4.

Caffeine. The betweenness and *P* value of caffeine (PubChem ID: CID000002519) were 371 and 0.028, respectively (row 6 of Table 4). Caffeine is a bitter, white crystalline xanthine alkaloid that can be extracted from coffee, tea, and other sources. A complex relationship has been reported between caffeine and cancer. For example, Sarkaria et al. suggested that caffeine could cause checkpoint defects, and, as a result, it might be useful for cancer therapy [77]. This statement could be regarded as an evidence to support our result. However, Wilson et al. observed a strong inverse association between coffee consumption and the risk of lethal prostate cancer, but this association appeared to be related to noncaffeine components of coffee [78]. Michels et al. did not find a strong association between caffeine and colon or rectal cancer [79]. Thus, further studies are needed to determine whether caffeine is associated with prostate cancer.

Trifluoperazine. The betweenness and *P* value of trifluoperazine (PubChem ID: CID000005566) were 363 and 0.001, respectively (row 7 of Table 4). Trifluoperazine is a typical

antipsychotic medicine of the phenothiazine chemical class. Calmodulin (CaM) is critical for the proliferation and viability of cells, including cancer cells. Trifluoperazine inhibits CaM [80]. The antitumor properties of trifluoperazine have been reported in murine T-cell lymphomas, metastatic breast cancer, and prostatic cancer [81–84]. These reports support the robustness of our analysis.

Mibefradil. The betweenness and *P*-value of mibefradil (PubChem ID: CID000060662) were 363 and 0.013, respectively (row 8 of Table 4). Mibefradil is a blocker of the L/T-type calcium channel [85], which plays an essential role in regulating cell growth and proliferation [86]. Dysregulation of this channel may lead to tumor progression [87]. Blocking the T-type Ca^{2+} channel with mibefradil inhibits tumor cell proliferation and migration in multiple types of tumors, including human astrocytoma, neuroblastoma, glioblastoma, and breast cancer cells [85, 87–89]. Our results suggest that mibefradil represents a new candidate chemical for prostate cancer.

Icilin. The betweenness and *P* value of icilin (PubChem ID: CID000161930) were 363 and 0, respectively (row 9 of Table 4). Icilin is an artificial superagonist of the transient receptor potential M8 (TRPM8) ion channel. Cold and cooling agents activate TRPM8, inducing a cooling sensation. TRPM8 is a tumor marker for diagnosis and a target for cancer therapy. TRPM8 expression increases in the early stages of prostate cancer, and it is involved in prostate cell apoptosis [90]. Direct activation of TRPM8 by icilin inhibits prostate cancer by reducing cancer cell motility [91]. Taken together with previous studies, our results suggest that icilin is closely related to prostate cancer, and it may be a promising drug.

Allicin. The betweenness and *P* value of allicin (PubChem ID: CID000065036) were 2 and 0.024, respectively (row 10 of Table 4). Allicin is a garlic extract with antibacterial properties. The antitumor ability of allicin can be traced back to the early 1960s [92]. Currently, many studies have reported that garlic and its extracts can prevent cancer, such as skin cancer [93], hepatocarcinoma [94], and so forth [95, 96]. Garlic may work by enhancing repair DNA synthesis (RDS), depressing nitrosamine formation and reducing carcinogen bioactivation [97, 98]. The correlation between allicin and prostate cancer may provide novel insight for future research.

4. Conclusions

This work provided an alternative computational method to investigate prostate cancer. Several candidate genes and chemicals were extracted using this method, and analysis of the literature confirmed that they are related to prostate cancer. We hope that the results of this study will lead to the validation of these genes and chemicals.

Conflict of Interests

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

Authors' Contribution

Fei Yuan and You Zhou contributed equally to this work.

Acknowledgments

This work was supported by the National Basic Research Program of China (2011CB510101, 2011CB510102), the National Natural Science Foundation of China (31371335, 91230201), and the Innovation Program of Shanghai Municipal Education Commission (12ZZ087).

References

[1] B. W. Stewart and C. P. Wild, *World Cancer Report 2014*, International Agency for Research on Cancer, Lyon, France, WHO Press, Geneva, Switzerland, 2014.

[2] National Cancer Institute, *Prostate Cancer Treatment (PDQ)*, National Cancer Institute, 2014.

[3] R. W. Ruddon, *Cancer Biology*, Oxford University Press, Oxford, UK, 4th edition, 2007.

[4] D. G. Bostwick and L. Cheng, *Urologic Surgical Pathology*, 2nd edition, 2007.

[5] J. M. Fitzpatrick, "Management of localized prostate cancer in senior adults: the crucial role of comorbidity," *BJU International, Supplement*, vol. 101, supplement 2, pp. 16–22, 2008.

[6] R. Morgan, A. Boxall, A. Bhatt et al., "Engrailed-2 (EN2): a tumor specific urinary biomarker for the early diagnosis of prostate cancer," *Clinical Cancer Research*, vol. 17, no. 5, pp. 1090–1098, 2011.

[7] A. Bourdoumis, A. G. Papatsoris, M. Chrisofos, E. Efstathiou, A. Skolarikos, and C. Deliveliotis, "The novel prostate cancer antigen 3 (PCA3) biomarker," *International Brazilian Journal of Urology*, vol. 36, no. 6, pp. 665–669, 2010.

[8] S. D. Catz and J. L. Johnson, "BCL-2 in prostate cancer: a minireview," *Apoptosis*, vol. 8, no. 1, pp. 29–37, 2003.

[9] S. Chakravarthi, P. Thanikachalam, H. S. Nagaraja, D. L. Wee Yang, and N. I. Bukhari, "Assessment of proliferative index and its association with Ki-67 antigen molecule expression in nodular hyperplasia of prostate," *Indian Journal of Science & Technology*, vol. 2, no. 8, pp. 1–4, 2009.

[10] A. K. Ramsay, S. R. C. McCracken, M. Soofi et al., "ERK5 signalling in prostate cancer promotes an invasive phenotype," *British Journal of Cancer*, vol. 104, no. 4, pp. 664–672, 2011.

[11] J. P. Struewing, P. Hartge, S. Wacholder et al., "The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews," *The New England Journal of Medicine*, vol. 336, no. 20, pp. 1401–1408, 1997.

[12] Scientists Discover Anti-Cancer Mechanism That Arrests Early Prostate Cancer, 2005, http://www.sciencedaily.com/releases/2005/08/050804074959.htm.

[13] R. A. Eeles, Z. Kote-Jarai, G. G. Giles et al., "Multiple newly identified loci associated with prostate cancer susceptibility," *Nature Genetics*, vol. 40, no. 3, pp. 316–321, 2008.

[14] G. Thomas, K. B. Jacobs, M. Yeager et al., "Multiple loci identified in a genome-wide association study of prostate cancer," *Nature Genetics*, vol. 40, no. 3, pp. 310–315, 2008.

[15] H. C. Whitaker, Z. Kote-Jarai, H. Ross-Adams et al., "The rs10993994 risk allele for prostate cancer results in clinically relevant changes in microseminoprotein-beta expression in

tissue and urine," *PLoS ONE*, vol. 5, no. 10, Article ID e13363, 2010.

[16] I. Leav, J. Plescia, H. L. Goel et al., "Cytoprotective mitochondrial chaperone TRAP-1 as a novel molecular target in localized and metastatic prostate cancer," *The American Journal of Pathology*, vol. 176, no. 1, pp. 393–401, 2010.

[17] S.-I. Watanabe, Y. Miyata, S. Kanda et al., "Expression of X-linked inhibitor of apoptosis protein in human prostate cancer specimens with and without neo-adjuvant hormonal therapy," *Journal of Cancer Research and Clinical Oncology*, vol. 136, no. 5, pp. 787–793, 2010.

[18] A. V. Sarma, J. C. McLaughlin, L. P. Wallner et al., "Sexual behavior, sexually transmitted diseases and prostatitis: the risk of prostate cancer in black men," *The Journal of Urology*, vol. 176, no. 3, pp. 1108–1113, 2006.

[19] M. Hisada, C. S. Rabkin, H. D. Strickler, W. E. Wright, R. E. Christianson, and B. J. van den Berg, "Human papillomavirus antibody and risk of prostate cancer," *Journal of the American Medical Association*, vol. 283, no. 3, pp. 340–341, 2000.

[20] L. K. Dennis, J. A. Coughlin, B. C. McKinnon et al., "Sexually transmitted infections and prostate cancer among men in the U.S. military," *Cancer Epidemiology Biomarkers and Prevention*, vol. 18, no. 10, pp. 2665–2671, 2009.

[21] L. C. Costello and R. B. Franklin, "The clinical relevance of the metabolism of prostate cancer; zinc and tumor suppression: connecting the dots," *Molecular Cancer*, vol. 5, article 17, 2006.

[22] N. V. Narizhneva, N. D. Tararova, P. Ryabokon et al., "Small molecule screening reveals a transcription-independent pro-survival function of androgen receptor in castration-resistant prostate cancer," *Cell Cycle*, vol. 8, no. 24, pp. 4155–4167, 2009.

[23] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: interaction networks of chemicals and proteins," *Nucleic Acids Research*, vol. 36, no. 1, pp. D684–D688, 2008.

[24] L. J. Jensen, M. Kuhn, M. Stark et al., "STRING 8—a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D412–D416, 2009.

[25] U. Consortium, "Update on activities at the Universal Protein Resource (UniProt) in 2013," *Nucleic Acids Research*, vol. 41, no. D1, pp. D43–D47, 2013.

[26] M. Zhao, J. Sun, and Z. Zhao, "TSGene: a web resource for tumor suppressor genes," *Nucleic Acids Research*, vol. 41, no. 1, pp. D970–D976, 2013.

[27] S. McNeil, A. Budhu, N. Grantees et al., *Imaging*, National Cancer Institute, 2013.

[28] A. P. Davis, C. G. Murphy, R. Johnson et al., "The comparative toxicogenomics database: update 2013," *Nucleic Acids Research*, vol. 41, no. 1, pp. D1104–D1114, 2013.

[29] L. Chen, J. Lu, T. Huang et al., "Finding candidate drugs for hepatitis C based on chemical-chemical and chemical-protein interactions," *PLoS ONE*, vol. 9, no. 9, Article ID e107767, 2014.

[30] L.-L. Hu, C. Chen, T. Huang, Y.-D. Cai, and K.-C. Chou, "Predicting biological functions of compounds based on chemical-chemical interactions," *PLoS ONE*, vol. 6, no. 12, Article ID e29491, 2011.

[31] L. L. Hu, T. Huang, X. Shi, W.-C. Lu, Y.-D. Cai, and K.-C. Chou, "Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties," *PLoS ONE*, vol. 6, no. 1, Article ID e14556, 2011.

[32] L. Zhang, "Sequence-based prediction of protein-protein interactions using random tree and genetic algorithm," in *Intelligent Computing Technology*, vol. 7389 of *Lecture Notes in Computer Science*, pp. 334–341, 2012.

[33] L. Chen, J. Lu, N. Zhang, T. Huang, and Y.-D. Cai, "A hybrid method for prediction and repositioning of drug Anatomical Therapeutic Chemical classes," *Molecular BioSystems*, vol. 10, no. 4, pp. 868–877, 2014.

[34] E.-H. Yap, T. Rosche, S. Almo, and A. Fiser, "Functional clustering of immunoglobulin superfamily proteins with protein-protein interaction information calibrated hidden markov model sequence profiles," *Journal of Molecular Biology*, vol. 426, no. 4, pp. 945–961, 2014.

[35] B.-Q. Li, T. Huang, L. Liu, Y.-D. Cai, and K.-C. Chou, "Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network," *PLoS ONE*, vol. 7, no. 4, Article ID e33393, 2012.

[36] L. Chen, W.-M. Zeng, Y.-D. Cai, K.-Y. Feng, and K.-C. Chou, "Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities," *PLoS ONE*, vol. 7, no. 4, Article ID e35254, 2012.

[37] C.-W. Tung, "Acquiring decision rules for predicting ames-negative hepatocarcinogens using chemical-chemical interactions," in *Pattern Recognition in Bioinformatics*, vol. 8626 of *Lecture Notes in Computer Science*, pp. 1–9, Springer International Publishing, 2014.

[38] E. Klipp, R. C. Wade, and U. Kummer, "Biochemical network-based drug-target prediction," *Current Opinion in Biotechnology*, vol. 21, no. 4, pp. 511–516, 2010.

[39] M. Re and G. Valentini, "Network-based drug ranking and repositioning with respect to DrugBank therapeutic categories," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 6, pp. 1359–1371, 2013.

[40] C.-W. Tung and J.-L. Jheng, "Interpretable prediction of non-genotoxic hepatocarcinogenic chemicals," *Neurocomputing*, vol. 145, pp. 68–74, 2014.

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

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

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

[44] A. S. Dhillon, S. Hagan, O. Rath, and W. Kolch, "MAP kinase signalling pathways in cancer," *Oncogene*, vol. 26, no. 22, pp. 3279–3290, 2007.

[45] C. A. Dionne, A. M. Camoratto, J. P. Jani et al., "Cell cycle-independent death of prostate adenocarcinoma is induced by the trk tyrosine kinase inhibitor CEP-751 (KT6587)," *Clinical Cancer Research*, vol. 4, no. 8, pp. 1887–1898, 1998.

[46] A. T. Weeraratna, J. T. Arnold, D. J. George, A. DeMarzo, and J. T. Isaacs, "Rational basis for Trk inhibition therapy for prostate cancer," *Prostate*, vol. 45, no. 2, pp. 140–148, 2000.

[47] T. Reya and H. Clevers, "Wnt signalling in stem cells and cancer," *Nature*, vol. 434, no. 7035, pp. 843–850, 2005.

[48] B. Lustig and J. Behrens, "The Wnt signaling pathway and its role in tumor development," *Journal of Cancer Research and Clinical Oncology*, vol. 129, no. 4, pp. 199–221, 2003.

[49] C. I. Truica, S. Byers, and E. P. Gelmann, "Beta-catenin affects androgen receptor transcriptional activity and ligand specificity," *Cancer Research*, vol. 60, no. 17, pp. 4709–4713, 2000.

[50] M. Verras and Z. Sun, "Roles and regulation of Wnt signaling and β-catenin in prostate cancer," *Cancer Letters*, vol. 237, no. 1, pp. 22–32, 2006.

[51] M. Fuse, N. Nohata, S. Kojima et al., "Restoration of miR-145 expression suppresses cell proliferation, migration and invasion in prostate cancer by targeting FSCN1," *International Journal of Oncology*, vol. 38, no. 4, pp. 1093–1101, 2011.

[52] T. Imamoto, H. Suzuki, M. Yano et al., "The role of testosterone in the pathogenesis of prostate cancer," *International Journal of Urology*, vol. 15, no. 6, pp. 472–480, 2008.

[53] A. W. Roddam, N. E. Allen, P. Appleby, and T. J. Key, "Endogenous sex hormones and prostate cancer: a collaborative analysis of 18 prospective studies," *Journal of the National Cancer Institute*, vol. 100, no. 3, pp. 170–183, 2008.

[54] B. Gershman, I. M. Shui, M. Stampfer et al., "Prediagnostic circulating sex hormones are not associated with mortality for men with prostate cancer," *European Urology*, vol. 65, no. 4, pp. 683–689, 2014.

[55] Z. Ni, W. Lou, E. S. Leman, and A. C. Gao, "Inhibition of constitutively activated Stat3 signaling pathway suppresses growth of prostate cancer cells," *Cancer Research*, vol. 60, no. 5, pp. 1225–1228, 2000.

[56] Z. Wang, Y. Li, S. Banerjee et al., "Down-regulation of Notch-1 and Jagged-1 inhibits prostate cancer cell growth, migration and invasion, and induces apoptosis via inactivation of Akt, mTOR, and NF-κB signaling pathways," *Journal of Cellular Biochemistry*, vol. 109, no. 4, pp. 726–736, 2010.

[57] F. Macian, "NFAT proteins: key regulators of T-cell development and function," *Nature Reviews Immunology*, vol. 5, no. 6, pp. 472–484, 2005.

[58] J. P. Vaque, G. Gómez-López, V. Monsálvez et al., "PLCG1 mutations in cutaneous T-cell lymphomas," *Blood*, vol. 123, no. 13, pp. 2034–2043, 2034.

[59] J. Kassis, D. A. Lauffenburger, T. Turner, and A. Wells, "Tumor invasion as dysregulated cell motility," *Seminars in Cancer Biology*, vol. 11, no. 2, pp. 105–117, 2001.

[60] A. Wells, "Tumor invasion: role of growth factor-induced cell motility," *Advances in Cancer Research*, vol. 78, pp. 31–101, 1999.

[61] N. P. Jones, J. Peak, S. Brader, S. A. Eccles, and M. Katan, "PLCγ1 is essential for early events in integrin signalling required for cell motility," *Journal of Cell Science*, vol. 118, no. 12, pp. 2695–2706, 2005.

[62] P. Y. Mak, D. H. Mak, V. Ruvolo et al., "Apoptosis repressor with caspase recruitment domain modulates second mitochondrial-derived activator of caspases mimetic-induced cell death through BIRC2/MAP3K14 signalling in acute myeloid leukaemia," *British Journal of Haematology*, vol. 167, no. 3, pp. 376–384, 2014.

[63] B. Z. Carter, P. Y. Mak, D. H. Mak et al., "Synergistic targeting of AML stem/progenitor cells with IAP antagonist birinapant and demethylating agents," *Journal of the National Cancer Institute*, vol. 106, no. 2, Article ID djt440, 2014.

[64] E. Varfolomeev, J. W. Blankenship, S. M. Wayson et al., "IAP antagonists induce autoubiquitination of c-IAPs, NF-kappaB activation, and TNFalpha-dependent apoptosis," *Cell*, vol. 131, no. 4, pp. 669–681, 2007.

[65] A. V. Paschall, M. A. Zimmerman, C. M. Torres et al., "Ceramide targets xIAP and cIAP1 to sensitize metastatic colon and breast cancer cells to apoptosis induction to suppress tumor progression," *BMC Cancer*, vol. 14, no. 1, article 24, 2014.

[66] E. C. Lau, T. K. Mohandas, L. J. Shapiro, H. C. Slavkin, and M. L. Snead, "Human and mouse amelogenin gene loci are on the sex chromosomes," *Genomics*, vol. 4, no. 2, pp. 162–168, 1989.

[67] E. C. Salido, P. H. Yen, K. Koprivnikar, L.-C. Yu, and L. J. Shapiro, "The human enamel protein gene amelogenin is expressed from both the X and the Y chromosomes," *The American Journal of Human Genetics*, vol. 50, no. 2, pp. 303–316, 1992.

[68] K. M. Sullivan, A. Mannucci, C. P. Kimpton, and P. Gill, "A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin," *BioTechniques*, vol. 15, no. 4, pp. 636–641, 1993.

[69] Y. M. Chang, R. Perumal, P. Y. Keat, R. Y. Y. Yong, D. L. C. Kuehn, and L. Burgoyne, "A distinct Y-STR haplotype for Amelogenin negative males characterized by a large Y_p11.2 (DYS458-MSY1-*AMEL-Y*) deletion," *Forensic Science International*, vol. 166, no. 2-3, pp. 115–120, 2007.

[70] W. Lattanzi, M. C. Di Giacomo, G. M. Lenato et al., "A large interstitial deletion encompassing the amelogenin gene on the short arm of the Y chromosome," *Human Genetics*, vol. 116, no. 5, pp. 395–401, 2005.

[71] Y. M. Chang, L. A. Burgoyne, and K. Both, "Higher failures of amelogenin sex test in an Indian population group," *Journal of Forensic Sciences*, vol. 48, no. 6, pp. 1309–1313, 2003.

[72] B. J. Elmendorf, A. Shilatifard, Q. Yan, J. W. Conaway, and R. C. Conaway, "Transcription factors TFIIF, ELL, and Elongin negatively regulate SII-induced nascent transcript cleavage by non-arrested RNA polymerase II elongation intermediates," *The Journal of Biological Chemistry*, vol. 276, no. 25, pp. 23109–23114, 2001.

[73] J. C. Eissenberg, J. Ma, M. A. Gerber, A. Christensen, J. A. Kennison, and A. Shilatifard, "dELL is an essential RNA polymerase II elongation factor with a general role in development," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 15, pp. 9894–9899, 2002.

[74] M. Gerber, J. Ma, K. Dean, J. C. Eissenberg, and A. Shilatifard, "*Drosophila* ELL is associated with actively elongating RNA polymerase II on transcriptionally active sites in vivo," *The EMBO Journal*, vol. 20, no. 21, pp. 6104–6114, 2001.

[75] M. J. Thirman, D. A. Levitan, H. Kobayashi, M. C. Simon, and J. D. Rowley, "Cloning of ELL, a gene that fuses to MLL in a t(11;19)(q23;p13.1) in acute myeloid leukemia," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 25, pp. 12110–12114, 1994.

[76] W. Xiao, F. Jiang, and Z. Wang, "ELL binding regulates U19/Eaf2 intracellular localization, stability, and transactivation," *Prostate*, vol. 66, no. 1, pp. 1–12, 2006.

[77] J. N. Sarkaria, E. C. Busby, R. S. Tibbetts et al., "Inhibition of ATM and ATR kinase activities by the radiosensitizing agent, caffeine," *Cancer Research*, vol. 59, no. 17, pp. 4375–4382, 1999.

[78] K. M. Wilson, J. L. Kasperzyk, J. R. Rider et al., "Coffee consumption and prostate cancer risk and progression in the health professionals follow-up study," *Journal of the National Cancer Institute*, vol. 103, no. 11, pp. 876–884, 2011.

[79] K. B. Michels, W. C. Willett, C. S. Fuchs, and E. Giovannucci, "Coffee, tea, and caffeine consumption and incidence of colon and rectal cancer," *Journal of the National Cancer Institute*, vol. 97, no. 4, pp. 282–292, 2005.

[80] D. Bar-Sagi and J. Prives, "Trifluoperazine, a calmodulin antagonist, inhibits muscle cell fusion," *The Journal of Cell Biology*, vol. 97, no. 5, part 1, pp. 1375–1380, 1983.

[81] S. Naftalovich, E. Yefenof, and Y. Eilam, "Antitumor effects of ketoconazole and trifluoperazine in murine T-cell lymphomas," *Cancer Chemotherapy and Pharmacology*, vol. 28, no. 5, pp. 384–390, 1991.

[82] E. Cifuentes, J. M. Mataraza, B. A. Yoshida et al., "Physical and functional interaction of androgen receptor with calmodulin in prostate cancer cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 2, pp. 464–469, 2004.

[83] G. T. Budd, R. M. Bukowski, A. Lichtin, L. Bauer, P. Van Kirk, and R. Ganapathi, "Phase II trial of doxorubicin and trifluoperazine in metastatic breast cancer," *Investigational New Drugs*, vol. 11, no. 1, pp. 75–79, 1993.

[84] A. Sivanandam, S. Murthy, K. Chinnakannu et al., "Calmodulin protects androgen receptor from calpain-mediated breakdown in prostate cancer cells," *Journal of Cellular Physiology*, vol. 226, no. 7, pp. 1889–1896, 2011.

[85] K. Hayashi, S. Wakino, Y. Ozawa et al., "Role of protein kinase C in Ca channel blocker-induced renal arteriolar dilation in spontaneously hypertensive rats—studies in the isolated perfused hydronephrotic kidney," *Keio Journal of Medicine*, vol. 54, no. 2, pp. 102–108, 2005.

[86] B. Ciapa, D. Pesando, M. Wilding, and M. Whitaker, "Cell-cycle calcium transients driven by cyclic changes in inositol trisphosphate levels," *Nature*, vol. 368, no. 6474, pp. 875–878, 1994.

[87] J. Pottle, C. Sun, L. Gray, and M. Li, "Exploiting MCF-7 cells' calcium dependence with interlaced therapy," *Journal of Cancer Therapy*, vol. 4, no. 7, pp. 32–40, 2013.

[88] A. Panner, L. L. Cribbs, G. M. Zainelli, T. C. Origitano, S. Singh, and R. D. Wurster, "Variation of T-type calcium channel protein expression affects cell division of cultured tumor cells," *Cell Calcium*, vol. 37, no. 2, pp. 105–119, 2005.

[89] Y. Zhang, J. Zhang, D. Jiang et al., "Inhibition of T-type Ca^{2+} channels by endostatin attenuates human glioblastoma cell proliferation and migration," *British Journal of Pharmacology*, vol. 166, no. 4, pp. 1247–1260, 2012.

[90] B. Beck, G. Bidaux, A. Bavencoffe et al., "Prospects for prostate cancer imaging and therapy using high-affinity TRPM8 activators," *Cell Calcium*, vol. 41, no. 3, pp. 285–294, 2007.

[91] D. Gkika and N. Prevarskaya, "TRP channels in prostate cancer: the good, the bad and the ugly?" *Asian Journal of Andrology*, vol. 13, no. 5, pp. 673–676, 2011.

[92] J. A. Dipaolo and C. Carruthers, "The effect of allicin from garlic on tumor growth," *Cancer Research*, vol. 20, pp. 431–434, 1960.

[93] H.-C. Wang, J. Pao, S.-Y. Lin, and L.-Y. Sheen, "Molecular mechanisms of garlic-derived allyl sulfides in the inhibition of skin cancer progression," *Annals of the New York Academy of Sciences*, vol. 1271, no. 1, pp. 44–52, 2012.

[94] C.-L. Zhang, T. Zeng, X.-L. Zhao, L.-H. Yu, Z.-P. Zhu, and K.-Q. Xie, "Protective effects of garlic oil on hepatocarcinoma induced by N-nitrosodiethylamine in rats," *International Journal of Biological Sciences*, vol. 8, no. 3, pp. 363–374, 2012.

[95] J.-Y. Hu, Y.-W. Hu, J. J. Zhou, M.-W. Zhang, D. Li, and S. Zheng, "Consumption of garlic and risk of colorectal cancer: an updated meta-analysis of prospective studies," *World Journal of Gastroenterology*, vol. 20, no. 41, pp. 15413–15422, 2014.

[96] A. Y. Nasr and H. A. Saleh, "Aged garlic extract protects against oxidative stress and renal changes in cisplatin-treated adult male rats," *Cancer Cell International*, vol. 14, article 92, 2014.

[97] G. N. Lvova and G. D. Zasukhina, "Modification of repair DNA synthesis in mutagen-treated human fibroblasts during adaptive response and the antimutagenic effect of garlic extract," *Genetika*, vol. 38, no. 3, pp. 306–309, 2002.

[98] J. A. Milner, "Mechanisms by which garlic and allyl sulfur compounds suppress carcinogen bioactivation. Garlic and carcinogenesis," *Advances in Experimental Medicine and Biology*, vol. 492, pp. 69–81, 2001.

[99] H. E. Mäki, K. K. Waltering, M. J. Wallén et al., "Screening of genetic and expression alterations of SRC1 gene in prostate cancer," *Prostate*, vol. 66, no. 13, pp. 1391–1398, 2006.

[100] J. Bouchal, F. R. Santer, P. P. S. Höschele, E. Tomastikova, H. Neuwirt, and Z. Culig, "Transcriptional coactivators p300 and CBP stimulate estrogen receptor-beta signaling and regulate cellular events in prostate cancer," *Prostate*, vol. 71, no. 4, pp. 431–437, 2011.

[101] M. Abdul and N. Hoosein, "Changes in beta-2 microglobulin expression in prostate cancer," in *Urologic Oncology: Seminars and Original Investigations*, Elsevier, New York, NY, USA, 2000.

[102] M. Ruscica, P. Magni, L. Steffani et al., "Characterization and sub-cellular localization of SS1R, SS2R, and SS5R in human late-stage prostate cancer cells: effect of mono- and bi-specific somatostatin analogs on cell growth," *Molecular and Cellular Endocrinology*, vol. 382, no. 2, pp. 860–870, 2014.

[103] W. Chen, G.-M. Wang, J.-M. Guo, L.-A. Sun, and H. Wang, "NGF/γ-IFN inhibits androgen-independent prostate cancer and reverses androgen receptor function through downregulation of FGFR2 and decrease in cancer stem cells," *Stem Cells and Development*, vol. 21, no. 18, pp. 3372–3380, 2012.

[104] Y.-W. Kwon, I. H. Chang, K. D. Kim et al., "Significance of S100A2 and S100A4 expression in the progression of prostate adenocarcinoma," *The Korean Journal of Urology*, vol. 51, no. 7, pp. 456–462, 2010.

[105] L. Sviridonov, M. Dobkin-Bekman, B. Shterntal et al., "Differential signaling of the GnRH receptor in pituitary gonadotrope cell lines and prostate cancer cell lines," *Molecular and Cellular Endocrinology*, vol. 369, no. 1-2, pp. 107–118, 2013.

[106] Q. Hong, E. Kuo, L. Schultz, R. J. Boackle, and N.-S. Chang, "Conformationally altered hyaluronan restricts complement classical pathway activation by binding to C1q, C1r, C1s, C2, C5 and C9, and suppresses WOX1 expression in prostate DU145 cells," *International Journal of Molecular Medicine*, vol. 19, no. 1, pp. 173–179, 2007.

[107] S. di Meo, I. Airoldi, C. Sorrentino, A. Zorzoli, S. Esposito, and E. di Carlo, "Interleukin-30 expression in prostate cancer and its draining lymph nodes correlates with advanced grade and stage," *Clinical Cancer Research*, vol. 20, no. 3, pp. 585–594, 2014.

[108] J. Feng, J. Sun, S.-T. Kim et al., "A genome-wide survey over the ChIP-on-chip identified androgen receptor-binding genomic regions identifies a novel prostate cancer susceptibility locus at 12q13.13," *Cancer Epidemiology Biomarkers & Prevention*, vol. 20, no. 11, pp. 2396–2403, 2011.

[109] S. Xiong, Q. Wang, S. V. Liu et al., "Effects of luteinizing hormone receptor signaling in prostate cancer cells," *The Prostate*, vol. 75, no. 2, pp. 141–150, 2015.

[110] G. Botchkina and I. Ojima, *Prostate and Colon Cancer Stem Cells as a Target for Anti-Cancer Drug Development*, INTECH, 2011.

[111] A. J. Najy, J. J. Won, L. S. Movilla, and H.-R. C. Kim, "Differential tumorigenic potential and matriptase activation between PDGF B versus PDGF D in prostate cancer," *Molecular Cancer Research*, vol. 10, no. 8, pp. 1087–1097, 2012.

[112] S. Zha, S. Ferdinandusse, J. L. Hicks et al., "Peroxisomal branched chain fatty acid beta-oxidation pathway is upregulated in prostate cancer," *Prostate*, vol. 63, no. 4, pp. 316–323, 2005.

[113] O. Kankavi, M. Baykara, M. I. Eren Karanis, C. I. Bassorgun, H. Ergin, and M. A. Ciftcioglu, "Evidence of surfactant protein A and D expression decrement and their localizations in human prostate adenocarcinomas," *Renal Failure*, vol. 36, no. 2, pp. 258–265, 2014.

[114] W. L. Jamieson, S. Shimizu, J. A. D'Ambrosio, O. Meucci, and A. Fatatis, "CX3CR1 is expressed by prostate epithelial cells and androgens regulate the levels of CX3CL1/fractalkine in the bone marrow: potential role in prostate cancer bone tropism," *Cancer Research*, vol. 68, no. 6, pp. 1715–1722, 2008.

[115] A. P. Singh, S. Bafna, K. Chaudhary et al., "Genome-wide expression profiling reveals transcriptomic variation and perturbed gene networks in androgen-dependent and androgen-independent prostate cancer cells," *Cancer Letters*, vol. 259, no. 1, pp. 28–38, 2008.

[116] W. Luo, A. R. Karpf, K. K. Deeb et al., "Epigenetic regulation of vitamin D 24-hydroxylase/CYP24A1 in human prostate cancer," *Cancer Research*, vol. 70, no. 14, pp. 5953–5962, 2010.

[117] B. Ning, C. Wang, F. Morel et al., "Human glutathione S-transferase A2 polymorphisms: variant expression, distribution in prostate cancer cases/controls and a novel form," *Pharmacogenetics*, vol. 14, no. 1, pp. 35–44, 2004.

[118] O. R. Saramäki, K. P. Porkka, R. L. Vessella, and T. Visakorpi, "Genetic aberrations in prostate cancer by microarray analysis," *International Journal of Cancer*, vol. 119, no. 6, pp. 1322–1329, 2006.

[119] K. Knerr, K. Ackermann, T. Neidhart, and W. Pyerin, "Bone metastasis: osteoblasts affect growth and adhesion regulons in prostate tumor cells and provoke osteomimicry," *International Journal of Cancer*, vol. 111, no. 1, pp. 152–159, 2004.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 715639, 11 pages
http://dx.doi.org/10.1155/2015/715639

Research Article

Identifying Novel Candidate Genes Related to Apoptosis from a Protein-Protein Interaction Network

Baoman Wang,[1] Fei Yuan,[1] Xiangyin Kong,[1] Lan-Dian Hu,[1] and Yu-Dong Cai[2]

[1]*Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[2]*College of Life Science, Shanghai University, Shanghai 200444, China*

Correspondence should be addressed to Xiangyin Kong; xykong@sibs.ac.cn, Lan-Dian Hu; ldhu2013@163.com, and Yu-Dong Cai; cai_yud@126.com

Received 1 June 2015; Accepted 29 June 2015

Academic Editor: Lin Lu

Apoptosis is the process of programmed cell death (PCD) that occurs in multicellular organisms. This process of normal cell death is required to maintain the balance of homeostasis. In addition, some diseases, such as obesity, cancer, and neurodegenerative diseases, can be cured through apoptosis, which produces few side effects. An effective comprehension of the mechanisms underlying apoptosis will be helpful to prevent and treat some diseases. The identification of genes related to apoptosis is essential to uncover its underlying mechanisms. In this study, a computational method was proposed to identify novel candidate genes related to apoptosis. First, protein-protein interaction information was used to construct a weighted graph. Second, a shortest path algorithm was applied to the graph to search for new candidate genes. Finally, the obtained genes were filtered by a permutation test. As a result, 26 genes were obtained, and we discuss their likelihood of being novel apoptosis-related genes by collecting evidence from published literature.

1. Introduction

Apoptosis, an efficient cell death program, plays an important role in maintaining strictly regulated organismal homeostasis and involves the interaction of multiple factors. Since the mid-nineteenth century, cell death has been widely studied, and researchers have learned that all physiological processes of multicellular organisms involve cell death, particularly during embryogenesis and metamorphosis [1, 2]. The first, the second, and the third PCD are the primary forms of apoptosis. The well-known caspase-dependent apoptosis is the first PCD. During the process of the second PCD, some vacuoles appear that have two membranes and autophagy functions; however, we know little regarding the third PCD. The second and the third PCD belong to caspase-independent apoptosis [3]. In the first, second, or third PCD, apoptosis maintains organism homeostasis and helps organism survival by defending against exogenous or endogenous toxic compounds. The intrinsic and extrinsic pathways have been well studied as the typical apoptotic processes [4–6]. Activated cell surface receptors mediate extrinsic apoptosis and transmit apoptotic signals through the combination of receptors and ligands. Death receptors consist of the tumor necrosis factor receptor gene superfamily, such as *TNFR-1*, *Fas/CD95*, and the *TRAIL* receptors *DR-4* and *DR-5* [7]. The first type PCD cells can bring about caspase-dependent apoptosis pathways [8]. A caspase cascade that is extreme enough to execute cell death cannot be generated from activated receptors in the second type PCD cells, and the signal amplification depends on mitochondria-dependent apoptotic pathways. Mitochondria, which are the central regulator of intrinsic apoptosis pathways and communicate with organelles, can connect the different apoptosis pathways [4]. The apoptosis pathway also involves some ion channels. The calcium channel represents the typical ion channel, and calcium ion concentration in the cytosol participates in signal transduction, cell death, and proliferation. Moreover, calcium channel opening or closing controls cell fate.

Organisms regulate their development and maintain through sophisticated interplay between cells. During development, organisms produce excess cells that finally go through PCD and contribute to the formation of organic

structures [9]. In interdigital mesenchymal tissue, the formation of independent digits through massive cell death is a typical example of PCD in development [10]. Apoptosis processes possess great biological significance, being involved in differentiation, development, proliferation, regulation, and so forth. Therefore, a variety of pathological conditions present dysregulation or dysfunction of the apoptotic program. Disorders in apoptosis can induce cancer, viral infection, and autoimmune disease; however, abnormal apoptosis will induce AIDS and neurodegenerative disease [11]. Multiple internal and external stimuli, such as ligands binding cell surface receptors, treatment with cytotoxic drugs or irradiation, DNA damage, contradictory cell cycle signaling, death signals, or a lack of survival signals can trigger apoptosis. The initiation, mediation, or execution of apoptosis involves many factors and once the genes encoding these factors mutate, the death machinery can be dysfunctional. Moreover, researchers have found that several mutations in apoptosis genes induce human diseases as initiating or contributing factors [12]. The excessive proliferation induced by the activation of oncogenes and disorders in apoptosis checkpoints have become primary factors in tumorigenesis over the last years [13].

Apoptosis contributes to maintaining the balance of homeostasis by normal cell death [3]. Necrosis can result in inflammation, but apoptosis yields few side effects. As such, apoptosis can be therapeutic targets to treat some diseases, for example, obesity [14], cancer, and neurodegenerative diseases. Therefore, the identification of key apoptosis-related genes before disease occurrence will greatly help in the prevention and treatment of disease. However, it is inefficient to discover novel apoptosis-related genes using traditional experiments. Building effective computational methods can highly increase this efficiency. We therefore proposed a computational method to identify apoptosis-related genes in this study. Twenty-six new genes were identified, which were related to the biological processes of apoptosis by analyzing previously published literature.

2. Materials and Methods

2.1. Genes Related to Apoptosis. Previously known apoptosis-related genes were obtained from KEGG [15], a database resource for understanding high-level functions and utilities of biological systems from molecular-level information. In detail, 86 human genes were extracted from the information in the pathway hsa04210: Apoptosis-Homo sapiens (human) from the website: http://www.genome.jp/dbget-bin/www_bget?hsa04210. The names of these genes are available in Supplementary Material I available online at http://dx.doi.org/10.1155/2015/715639.

2.2. Method to Identify Novel Candidate Genes. To identify the novel candidate genes related to apoptosis, we used protein-protein interaction information that was retrieved from the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins, Version 9.0, http://www.string-db.org/) database [16] to construct a weighted graph. The weighted graph's construction procedures were the same as those in [17–19]. Here, we gave the brief description of these procedures; readers can refer to these studies for additional details. From the obtained file (protein.links.v9.0.txt.gz) retrieved from STRING, we extracted all protein-protein interactions of human. Each extracted protein-protein interaction of human consists of two proteins, represented by Ensembl IDs, and one score that evaluates the strength of the interaction with range between 150 and 999. The constructed graph took proteins, collected from all obtained protein-protein interactions of humans, as nodes and two nodes were adjacent if and only if the corresponding proteins can comprise an interaction. Obviously, each edge represented a protein-protein interaction of human. In addition, each edge was assigned a weight, which was defined as 1,000 minus the interaction score of the corresponding interaction.

The shortest path algorithm, Dijkstra's algorithm [20], was executed on the constructed graph to search for the shortest paths connecting any two known apoptosis-related genes. According to the definition of edge weight, consecutive genes in a shortest path were in an interaction with high interaction score, meaning they are more likely to share similar functions. The obtained shortest paths were used to calculate the betweenness of each node/gene in the constructed graph, which is defined as the number of shortest paths containing a certain node/gene as an inner node. Then we excluded genes with betweenness equal to zero and apoptosis-related genes. The remaining genes were further filtered with a permutation test. 500 gene sets were produced by randomly selecting genes in the constructed graph and these gene sets had the same sizes of the apoptosis-related gene set. For each gene set, all shortest paths connecting any two genes in the set were searched in the graph. The betweenness for each remaining genes was calculated based on these paths. Accordingly, for each remaining gene, there were 500 betweenness on 500 randomly produced sets and one betweenness on the apoptosis-related gene set. Another measurement, permutation FDR, was calculated for each remaining gene, which was defined as the ratio of the number of randomly produced gene sets in which the betweenness was larger than that of the known apoptosis-related gene set and the total number of randomly produced gene sets (500). Genes with permutation FDRs less than 0.05 were finally selected as significantly associated with apoptosis.

3. Results and Discussions

Based on 86 known apoptosis-related genes, some candidate genes can be obtained according to the method described in Section 2.2. The detailed procedure and result of each step is illustrated in Figure 1.

3.1. Results of the Method. The shortest paths connecting any pair of the 86 human genes related to apoptosis were searched in a constructed weighted graph. We discovered 114 genes with betweenness greater than zero (Supplementary Material II). Additionally, a permutation test excluded false discoveries that had high betweenness and little relationship with apoptosis by calculating permutation FDR for each of 114 candidate genes and setting 0.05 as the threshold. We finally

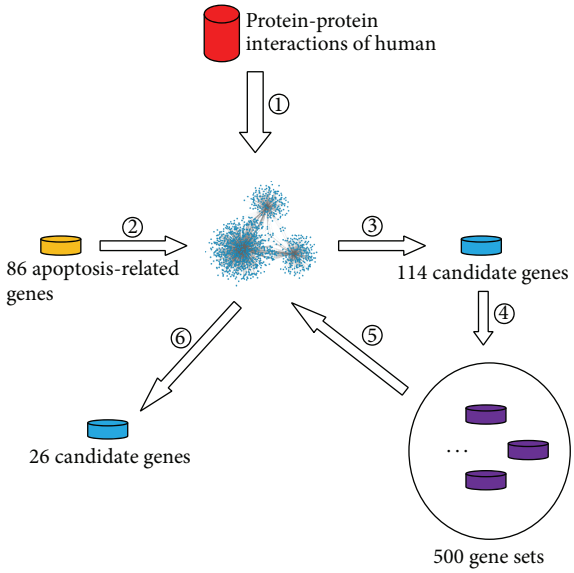


Figure 1: The procedure and result of each step in our method. ① Protein-protein interactions of human were used to construct a graph; ② all shortest paths connecting 86 known apoptosis-related genes were searched in the constructed graph; ③ 114 candidate genes were extracted from the obtained shortest paths; ④ 500 gene sets were randomly produced to execute permutation test; ⑤ for each gene set, all shortest paths connecting genes in the set were searched in the constructed graph; ⑥ 26 candidate genes were finally obtained by calculating permutation FDR.

obtained 26 genes, which are listed in Table 1. These genes are termed as significant candidate genes in the remaining parts of the paper.

3.2. Analysis of Significant Candidate Genes. We finally obtained 26 significant candidate genes for apoptosis. The following paragraphs gave detailed discussion on the relationships between these genes and apoptosis.

3.2.1. TRAF6 and TNFRSF1B. *TRAF6* (betweenness: 509, permutation FDR: <0.002; refer to Table 1, row 1) possesses unique receptor-binding specificity, which is vital as the signaling mediator in TNF receptor superfamily and the IL-1R/Toll-like receptor superfamily signaling pathway [21]. Because of its central convergence in different signal pathways, *TRAF6* is involved in regulating cell death, survival, and cellular responses to stress. Ample studies have shown that TRAF6 is involved in various cell apoptosis conditions. Most of these studies have suggested that TRAF6 regulates cell apoptosis by mediating the caspase-associated signaling pathway. In summary, *TRAF6* acts as a bifurcation point of the survival and death pathways. The subtle regulation by *TRAF6* in the imbalance of survival and death will finally determine cell fate and be a therapeutic target. In our study, one of these genes, *TNFRSF1B* (*TNFR2*) (betweenness: 86, permutation FDR: <0.002; refer to Table 1, row 2), belongs to the TNF receptor superfamily. For a long time, we have known little regarding TNF-induced signaling through TNFRSF1B and the mechanism of TNFR2-mediated cell death. A previous study demonstrated that TNFR2 triggers cell death in the presence of RIP, whereas without RIP [22] NF-κB is activated by TNFR2. Recently, researchers have identified that TNFRSF1B communicates with the JNK and TNF-induced NF-κB signaling pathways in vascular endothelial cells (EC). Understanding the TNFR2-mediated apoptotic and JNK signaling pathways may offer novel therapeutic targets to treat vascular diseases in EC [23]. This observation gives us more confidence in the accuracy of our calculation method. From the above description, we know that *TRAF6* plays an important role in the signaling mediated by the TNF receptor superfamily. Therefore, we hypothesize that these two genes may have synergistic effects on apoptosis regulation, which requires further validation.

3.2.2. IQGAP1. *IQGAP1* (betweenness: 289, permutation FDR: <0.002; refer to Table 1, row 3) belongs to a member of the *IQGAP* family and induces multiple cellular functions by interacting with its target proteins. Previous studies have found that IQGAP1 has interactions not only with cell adhesion molecules but also with cytoskeletal components and integrates multiple signaling pathways to regulate cell morphology and motility. Compared with normal tissue, IQGAP1 is overexpressed in colorectal carcinoma [24–26], breast cancer [26], astrocytoma, and head and neck squamous cell carcinoma [27], enhancing cell proliferation, migration, and invasion. Recently, research has indicated that it is also closely related to cell survival and apoptosis. ERK plays a vital role in several biological processes, particularly those involving cellular proliferation, differentiation, survival, and apoptosis [28]. In a mouse model of cardiac hypertrophy, IQGAP1 regulates Melusin-dependent cardiomyocyte hypertrophy and apoptosis via activation of MEK/ERK [29]. In addition, the interaction between RNase L and IQGAP1 can promote ECyd-induced apoptosis [30]. Taken together, we can speculate that *IQGAP1* plays a vital role in apoptosis and survival through signaling pathways, such as the MEK/ERK pathways and their partner proteins.

3.2.3. FURIN. *FURIN* (betweenness: 252, permutation FDR: <0.002; refer to Table 1, row 4) is a cellular endoprotease and participates in embryo formation and the maturation of proprotein substrates, which includes extra-cellular-matrix proteins, receptors, and other protease systems. Few studies have reported a direct relationship between *FURIN* and apoptosis, but Yang et al. have recently suggested that *FURIN* may be involved in regulating the proliferation and apoptosis of granulosa cells, because after *FURIN* was knocked down, the apoptosis of the granulosa cells was significantly increased from large antral/preovulatory follicles through downregulation of the antiapoptotic proteins XIAP and p-AKT [31]. Moreover, FURIN can activate massive proprotein substrates and is ubiquitously expressed and participates in many physiological and pathological processes. In our study, because *FURIN* displayed a higher betweenness value, we speculate that FURIN may indirectly be involved in the regulation of apoptosis through activating proprotein substrates in some signaling pathways.

Table 1: 26 significant candidate genes, their betweenness, and permutation FDRs.

Row number	Ensembl ID	Gene name	Betweenness	Permutation FDR
1	ENSP00000316840	TRAF6	509	<0.002
2	ENSP00000365435	TNFRSF1B	86	<0.002
3	ENSP00000268182	IQGAP1	289	<0.002
4	ENSP00000268171	FURIN	252	<0.002
5	ENSP00000327850	NFATC1	95	<0.002
6	ENSP00000379330	NFATC2	238	<0.002
7	ENSP00000315615	AKAP5	195	<0.002
8	ENSP00000267169	DIABLO	138	<0.002
9	ENSP00000216160	TAB1	120	<0.002
10	ENSP00000246533	CAPNS1	86	<0.002
11	ENSP00000380349	CAPN3	86	<0.002
12	ENSP00000286355	ADCY8	167	0.002
13	ENSP00000382834	NPRL3	83	0.008
14	ENSP00000288840	SMAD6	25	0.014
15	ENSP00000304895	IRS1	238	0.016
16	ENSP00000237596	PKD2	3	0.016
17	ENSP00000296871	CSF2	65	0.02
18	ENSP00000404503	BBC3	8	0.02
19	ENSP00000277541	NOTCH1	249	0.022
20	ENSP00000360683	PTPN1	93	0.024
21	ENSP00000327048	MAF	54	0.026
22	ENSP00000189444	NFKB2	8	0.034
23	ENSP00000349467	CALM1	244	0.038
24	ENSP00000258682	CAMK2B	179	0.04
25	ENSP00000264122	CBLB	5	0.042
26	ENSP00000360266	JUN	211	0.044

3.2.4. NFATC1 and NFATC2. NFATC1 (betweenness: 95, permutation FDR: <0.002; refer to Table 1, row 5) and NFATC2 (betweenness: 238, permutation FDR: <0.002; refer to Table 1, row 6) are the most famous NFAT factors in peripheral T cells and have similar function but different modes of expression. *NFATC2* belongs to the nuclear factor of activated T cells family and is a transcription factor involved in differentiation in lymphocytes. Many studies have demonstrated that *NFATC2* is involved in the regulation of apoptosis. In a *NFATC2*−/− mice model, *NFATC2*−/− cells not only presented an increase in apoptosis but also presented hyperproliferation [32]. Researchers have demonstrated that overexpression or activation of *NFAT1* can induce cell death in different cell types, for example, T lymphocytes, Burkitt's lymphoma, megakaryocytes, and fibroblasts [33–35]. Moreover, the calcineurin/NFATC2 pathway has an antiapoptotic role in melanoma cells. Apoptosis is induced by NFAT1 through cooperation with the Ras/Raf/MEK/ERK pathway and upregulates TNF-α expression in NIH3T3 fibroblasts [36]. Overexpression of *NFATC1* increases *TRAIL* expression in HT29 and Caco-2 cells and also induces FasL [37] and TNF-α expression upregulation in several cell types. For some time, the members of *NFAT* family have been considered to be redundant proteins. Nevertheless, in the regulation of cell proliferation and apoptosis, different roles for the *NFAT* family were identified by analyzing mice deficient for NFAT proteins. As transcription factors, the promoter regions of diverse activation-inducible genes all contain binding sites for NFAT proteins [32]. These activation-inducible genes include cytokines IL-2, IL-4, IL-5, and IFN-γ and cell surface proteins [33, 38, 39], suggesting that these transcription factors may participate in controlling the cell cycle and apoptosis [40, 41].

Constitutively active NFAT1 (CA-NFAT1) and NFAT2 short isoform (CA-NFAT2/A) mutants localize in the nucleus, bind DNA with high affinity, and activate endogenous *NFAT* target genes [34, 42]. Remarkably, in cell apoptosis, cycle, and transformation regulation, the abnormal expression of the CA-NFAT1 and CA-NFAT2 short isoform in NIH 3T3 fibroblasts presented opposite phenotypes. The NFAT2 short isoform acted as a repressor of cell death and a positive regulator of cell proliferation. Conversely, *NFAT1* increased cell death and repressed the cell cycle. In summary, the *NFAT1* and *NFAT2* genes present opposing roles in regulation of the cell cycle and apoptosis. Moreover, the *NFAT1* and *NFAT2* short isoform genes play dual roles as tumor suppressor or oncogene. The cell phenotype was transformed by CA-NFAT2; however, CA-NFAT1 could suppress the transformation, suggesting that different family members might have complementary functions, and the complementary functions might determine whether the cell

lives or dies. This observation also suggests that the cellular threshold levels of each NFAT protein and protein isoform determine the expression of a particular set of target genes, and ultimately, this process determines the fate of the cell. However, more work is necessary to help us better understand the physiological role of the balance between the NFAT1 and NFAT2 short isoforms.

3.2.5. AKAP5. Betweenness of this gene was 195 and its permutation FDR was <0.002 (refer to Table 1, row 7). A-kinase anchoring proteins (AKAPs) mediate the localization of the c-AMP-dependent protein kinase (PKA) and other signaling enzymes. No studies have yet indicated that *AKAP5* is directly related to apoptosis. We hypothesize that *AKAP5* may be involved in cell apoptosis by forming complexes with protein kinases, phosphatases, or scaffold proteins. The assembly and localization of signaling complexes are coordinated by scaffold, anchoring, and adaptor proteins to provide efficiency and specificity in signal transduction [43]. It therefore seems reasonable that defects in anchoring protein genes or pathophysiological changes in AKAP signaling complexes may underlie certain damages in cells or tissues. Studies have shown that AKAP5 can form a complex with IQGAP1, and the complex also contributes to the c-AMP/PKA signaling pathway. Because IQGAP1 is a scaffold protein and is involved in apoptosis regulation, it is not surprising that *AKAP5* is also linked to apoptosis. In addition, AKAP5 can interact with ADCY8 to regulate Ca^{2+}-dependent c-AMP synthesis in pancreatic and neuronal systems [44]. In the calcium signaling pathway, ADCY8 catalyzes the formation of c-AMP, which phosphorylates PKA to induce the endoplasmic reticulum to release Ca^{2+}, resulting in the expression of genes such as *CALM* and *CAMK* that cause cell proliferation and apoptosis. PKA also inhibits phosphorylation of BAD and suppresses apoptosis. Further experimental verification is necessary to test these hypotheses.

3.2.6. DIABLO. *DIABLO* (betweenness: 138, permutation FDR: <0.002; refer to Table 1, row 8), also called *Smac*, is a factor that has been shown to exit mitochondria in response to apoptotic stimuli and potentiate caspase activity. The function of *DIABLO* has been elaborated in detail. The inhibitory effect on both initiator and effector caspases is relieved by Smac through interacting with multiple IAPs [45–48], finally promoting apoptosis. Therefore, *Smac/DIABLO* may play a significant role in diagnostic and therapeutic features in cancer. Increasing data suggests that chemoradiation-resistance to apoptosis may result from decreased levels of Smac/DIABLO in advanced colon cancer [49]. In addition, numerous studies have observed that Smac mRNA expression is significantly lower in melanoma, prostate cancer, lung cancer, gastric cancer, colon cancer, and so forth [50–52]. Therefore, the design and development of small-molecule Smac mimetics as novel therapy targets is promising.

3.2.7. TAB1. The betweenness of this gene was 120 and its permutation FDR was <0.002 (refer to Table 1, row 9). The TAB1 protein is a regulator of the MAP kinase kinase kinase MAP3K7/TAK1 and can mediate various intracellular signaling pathways. This protein interacts and activates TAK1 kinase. TAK1 mediates multiple inducible transcription factors, such as NF-κB and JNKs [53], which contribute to the development of the embryo, cell survival, and innate immunity. The inhibition of TAK1 activity will suppress cancer cell death, and TAB1 interacts with TAK1 and promotes its autophosphorylation. The interaction between TAB1 and TAK1 therefore controls biological processes, particularly apoptosis. Research has also demonstrated that greater amounts of Xenopus *TAB1* (*xTAB1*) and *xTAK1* mRNAs injected into early embryos can result in cell death [54]. Many investigators have reported that XIAP not only functionally interacts with the BMP receptor but also with the adapter molecule TAB1, and in the presence of the transforming growth factor β1 (TGF-β1), TAK1 activates JNK1 and p38 as an upstream MAP3 kinase. The XIAP/TAK1-mediated activation of JNK1 depends on TAB1, and the XIAP/TAK1-mediated activation of JNK1 is involved in protection against apoptosis [55]. The proapoptotic pathway TAB1/p38 also mediates apoptosis [56]. In TRAIL-induced apoptotic pathways, the blockade of TAB1 activity enhances apoptosis through the activation of a caspase cascade. In addition, the BIR1 (a domain of XIAP)/TAB1 interaction is crucial for XIAP-induced NF-κB activation [57]. Taken together, we find that *TAB1* plays a vital role in regulating apoptosis and survival, consistent with our expectation.

3.2.8. CAPNS1 and CAPN3. CAPNS1 (betweenness: 86, permutation FDR: <0.002; refer to Table 1, row 10), as a common small regulatory subunit of calpains, is required to maintain the stability and activity of calpains. Some studies have reported that the BCL-2, procaspase 3, and Bax families are all calpain substrates and have confirmed a role for calpain during B and T cell development and apoptosis [58–60]. Because CAPNS1 is a common small regulatory subunit of calpains and contributes to maintaining the stability and activity of calpains, we presume that *CAPNS1* may indirectly be involved in the regulation of apoptosis. In addition, some recent studies have demonstrated that CAPNS1 participates in signaling pathways as a partner. In the Ras signaling pathway, CAPNS1 binds the RasGAP-SH3 domain in *K-Ras* (V12) oncogenic cells, modulating migration and cell survival, and the interaction between CAPNS1 and PP2A-Akt affects FoxO3A-dependent cell death [61]. In addition, calpain 3 belongs to the calpain family of calcium-dependent intracellular proteases, which also plays important roles in regulating apoptosis. The generation of the limb-girdle muscular dystrophy type 2A (LGMD2A) involves *CAPNS1* mutation. The muscular biopsy specimens of LGMD2A patients show that lack of calpain 3 causes IκBα accumulation and prevents NF-κB nuclear translocation, ultimately resulting in apoptosis. Moreover, deficiency in CAPN3 (betweenness: 86, permutation FDR: <0.002; refer to Table 1, row 11) is also associated with downregulation of the antiapoptotic factor c-FLIP and myonuclear apoptosis in LGMD2A muscles [62]. Whether *CAPNS1* and *CAPN3* interact or coordinately regulate apoptosis still must be studied.

3.2.9. ADCY8. The betweenness of this gene was 167 and its permutation FDR was 0.002 (refer to Table 1, row 12). Adenylate cyclase is a membrane-bound enzyme that contributes to the formation of cyclic AMP from ATP. Although no research has identified its direct relationship with apoptosis, it is an important member of the calcium signaling pathway. In the c-AMP signaling pathway, ADCY8 catalyzes the formation of c-AMP, which in turn induces the activation of PKA. PKA can promote the expression of many genes and influence the endoplasmic reticulum in regulating calcium concentration. In the regulation of signal transduction, the concentration of calcium in the cytosol plays a vital role and is involved in cell death and proliferation. In addition, calcium also can trigger cytochrome c release that does not depend on Bcl-2. In addition to its involvement in apoptosis, the calcium ion is also involved in many other signal pathways by controlling the ion channel's opening and closing [3]. Therefore, we presume that *ADCY8* may serve as a bridge between extracellular stimuli and apoptosis.

3.2.10. NPRL3. The betweenness of this gene was 83 and its permutation FDR was 0.008 (refer to Table 1, row 13). So far, we know little about the function of the encoded protein of NPRL3. However, its homolog NPR3 has been recently investigated in yeast and Drosophila [63]. Studies in yeast have demonstrated that an amino acid starvation signal to the target of rapamycin complex 1 (TORC1) can be mediated by the Npr2/3 complex, and artificially inhibiting TORC1 by rapamycin can rescue proliferation defects observed in npr2Δ and npr3Δ cells [63]. In addition, a study has demonstrated that in the female germ line in Drosophila TORC1 signaling can be inhibited by NPRL2 and NPRL3 in the absence of amino acids. In young egg chambers, apoptosis is inhibited by NPRL2 and NPRL3 by downregulating TORC1 activity in the condition of lack of nutrients. In addition, TORC1 is a key regulator of cell growth in response to amino acid availability [64]. Thus, these data suggest that TORC1 activity remains particularly high during periods of amino acid scarcity or other stress circumstances, and subsequently a cell death program will be initiated by a metabolic checkpoint. Nevertheless, the role of *NPRL3* in apoptosis in humans requires further research.

3.2.11. SMAD. The betweenness of this gene was 25 and its permutation FDR was 0.014 (refer to Table 1, row 14). The SMAD6 protein is a member of the *SMAD* family. SMAD proteins can mediate multiple signaling pathways through their roles as signal transducers and transcriptional modulators and negatively regulate BMP and TGF-beta/activin-signaling. Members of the TGF-beta family regulate multiple cellular processes, including cell proliferation, differentiation, organization, migration, and death. In addition, SMAD6 and SMAD2 predict overall survival in oral squamous cell carcinoma patients. However, the role of aberrant TGF-beta signaling is not clear [65]. In addition, in the lung adenocarcinoma cell line H1299, knockdown of *SMAD6* upregulates the plasminogen activator inhibitor-1 and phosphorylates SMAD2/3, finally activating TGF-beta signaling. Furthermore, because of the *SMAD6* knockdown, the JNK pathway is also activated and the phosphorylation of Rb-1 is reduced, causing G0-G1 cell apoptosis and arrest [66]. According to the above description, SMAD6 is a key factor in lung cancer cell growth and survival. Therefore, targeted inactivation of SMAD6 may open a new road for treating lung cancer. Moreover, when some lymphoma cell lines were exposed to TGF-β, Bcl-xl and Bcl-2 were downregulated, whereas Bax was upregulated. Furthermore, the mRNAs of *SMAD6* and *SMAD7* displayed significant upregulation [67]. These results indicated that the induction of apoptotic pathways may depend on alteration of the gene expression and protein levels. Another study has demonstrated that the TRAF6-TAK1-p38 MAPK/JNK pathway, a noncanonical TGF-β pathway, can be induced by TGF-β1; however, this process can be negatively regulated through the SMAD6 but not SMAD7. K63-linked poly-ubiquitination of TRAF6 can be abolished through the TGF-β1-induced SMAD6 in primary hepatocytes and AML-12 mouse liver cells. In addition, in cell culture or animal models, phosphorylation of TAK1 and p38 MAPK/JNK is maintained and apoptosis increased after knockdown of SMAD6 or A29, suggesting an important role of the SMAD6-A20 axis in negative regulation of the TGF-β1-TRAF6-TAK1-p38 MAPK/JNK pathway [68]. Recent research has shown that galangin can induce autophagy by activating the TGF-β receptor/SMAD pathway in HepG2 cells. In this process, SMAD6 and SMAD7 expression levels both decreased [69]. Taken together, SMAD6, as a negative regulator, participates in TGF-beta mediated apoptosis.

3.2.12. IRS1. The betweenness of this gene was 238 and its permutation FDR was 0.016 (refer to Table 1, row 15). The insulin receptor tyrosine kinase can phosphorylate the IRS1 protein. This gene mutates in type II diabetes with susceptibleness to insulin resistance. IRS1, as a member of the PI3K/AKT signaling pathway, regulates cell survival and apoptosis. A common Arg972 polymorphism in *IRS-1* affects the PI3-kinase/Akt survival pathway, which in turn results in resistance to the antiapoptotic effects of insulin. In addition, the Arg972 polymorphism also impairs human β-cell survival [70]. A report has observed that PTPL1 dephosphorylates IRS1, and PTPL1 expression can block the IRS-1/PI3K/Akt signaling pathway [71], finally inhibiting the insulin-like growth factor-I effect on cell survival and apoptosis.

3.2.13. PKD2. The protein encoded by *PKD2* (betweenness: 3, permutation FDR: 0.016; refer to Table 1, row 16) is a transmembrane protein and a calcium-permeable cation channel. In addition, *PKD2* is also responsible for transporting calcium signaling in renal epithelial cells. Calcium concentration changes can induce a series of cell biology processes to occur, such as activation of the MAPK signaling pathway [72], eventually controlling cell survival and apoptosis. In addition, calcium also activates the JNK pathway, which can subsequently stimulate Bax activation [3]. Taken together, we speculate that *PKD2* may regulate calcium cations through opening and closing calcium channels, thereby triggering physiological and pathological change and finally deciding cell fate.

3.2.14. CSF2. The protein encoded by *CSF2* (betweenness: 65, permutation FDR: 0.02; refer to Table 1, row 17) is a cytokine that regulates the differentiation and function of granulocytes and macrophages. Recently, some studies have reported that *CSF2* is associated with apoptosis. CSF2 can block apoptosis in bovine embryos through interaction with genes controlling apoptosis [73]. In addition, in advanced atherosclerosis, *GM-CSF* promotes macrophage apoptosis and plaque necrosis through IL-23 signaling [74].

3.2.15. BBC3 (PUMA). *BBC3* (betweenness: 8, permutation FDR: 0.02; refer to Table 1, row 18) belongs to a member of the *BCL-2* family. These family members are also in the BH3-only proapoptotic subclass. This protein induces mitochondrial outer membrane permeabilization and apoptosis through cooperating with direct activator proteins. As mentioned above, DIABLO is released from mitochondria, which can be increased through mitochondrial outer membrane permeabilization enhancement. In addition, *BBC3* was identified 12 years ago, mediates p53-dependent and p53-independent apoptosis, and is also involved in the intrinsic apoptosis pathway [75]. In the induction of apoptosis, a key regulatory step is PUMA binding to the inhibitory members of the Bcl-2 family (Bcl-2-like proteins), such as Bax/Bak, via its BH3 domain, which induces Smac/DIABLO release from mitochondria, finally resulting in intrinsic apoptosis [76].

3.2.16. NOTCH1. The betweenness of this gene was 249 and its permutation FDR was 0.022 (refer to Table 1, row 19). *NOTCH1* encodes a member of the NOTCH family. NOTCH signaling participates in maintaining the balance of cell proliferation, differentiation, and apoptosis; therefore, disorders in NOTCH signaling may induce tumorigenesis. The active form of NOTCH1, NOTCH1-ICN, is involved in many cell processes, such as T/B cell development and apoptosis, progression, and deterioration of various cancers. NOTCH1 induces resistance to glucocorticoid-induced apoptosis in developing thymocytes through downregulation of SRG3 expression [77]. In addition, downregulated expression of NOTCH1 promotes apoptosis and cell growth inhibition in pancreatic cancer cells [78]. Apoptosis is induced and cell proliferation is inhibited after the NOTCH1 signaling pathway is activated in the human esophageal squamous cell carcinoma cell line EC9706 [79]. Moreover, NOTCH1 also regulates apoptosis through participating in survival and apoptosis pathways. For example, NOTCH1 signaling can inhibit Akt/Hdm2-mediated p53 degradation and sensitizes human hepatocellular carcinoma (HCC) cells to TRAIL-induced apoptosis. NOTCH1 also inhibits apoptosis through activation of the PI3K-PKB/Akt pathway [80, 81].

3.2.17. PTPN1 (PTP1B). The protein encoded by *PTPN1* (betweenness: 93, permutation FDR: 0.024; refer to Table 1, row 20) belongs to the protein tyrosine phosphatase (PTP) family. PTPs have been well-known to regulate many cellular events, such as cell growth, differentiation, motility, and proliferation [82]. PTP-1B can also regulate the phosphorylation status of apoptotic proteins [83]. Therefore, research suggests that the apoptosis of hepatocytes caused by serum withdrawal can be protected through PTP1B deficiency [84], whereas its overexpression increases cellular events and results in apoptotic cell death. In cardiomyocytes, hypoxia/reoxygenation-induced apoptosis is also reduced by siRNA targeted to *PTP1B*. PTP1B deficiency is also involved in protecting against Fas-induced hepatic failure [85].

3.2.18. MAF (c-MAF). The betweenness of this gene was 54 and its permutation FDR was 0.026 (refer to Table 1, row 21). The MAF protein is a transcription factor containing a leucine zipper that can bind DNA. Because its folding type includes homodimer and heterodimer, it can transactivate target genes to participate in cellular processes. Recently, some studies have demonstrated that c-Maf can interact with c-Myb, downregulate Bcl-2 expression, increase cell death in peripheral CD4 cells [86], and transactivate the tumor suppressor gene *p53* in vitro. The apoptosis of primary cell lines is induced by overexpression of c-Maf via a p53-dependent mechanism [87]. In addition, c-Maf enhances apoptosis through transactivating caspase 6 in peripheral CD8 cells [88]. Taken together, we observed that MAF is closely related to apoptosis.

3.2.19. NFKB2. The betweenness of this gene was 8 and its permutation FDR was 0.034 (refer to Table 1, row 22). *NFKB2* encodes a subunit of the transcription factor complex NF-κB. NF-κB significantly functions in regulating the immune response; however, in almost the same manner, it also induces proliferation, inflammation, and regulation of apoptosis [89]. For example, variation in NF-κB activity results in mitochondrial apoptosis after infecting cells with pathological prion proteins. In addition, NFKB2, as a subunit of NF-κB, is involved in the MAPK signaling pathway, which finally regulates proliferation, inflammation, and antiapoptosis. Nevertheless, these hypotheses require further validation.

3.2.20. CALM1. CALM1 (betweenness: 244, permutation FDR: 0.038; refer to Table 1, row 23) is one member of the EF-hand calcium-binding protein family, and its function is regulated by calcium. No report has so far observed a relationship between *CALM1* and apoptosis; however, calcium concentration changes can induce a series of cell biological processes to happen. Because calcium activates the MAPK and JNK signaling pathways, we presume that calcium activates CALM1, which in turn activates downstream gene expression, potentially including apoptosis-related genes, such as *BCL-2*, finally regulating cell survival and apoptosis. In addition, in the STRING analysis, CALM1 has interactions with IQGAP1 and ADCY8. As mentioned above, *IQGAP1* and *ADCY8* have been well characterized to participate in apoptosis. In summary, these arguments all support our results.

3.2.21. CAMK2B. The betweenness of this gene was 179 and its permutation FDR was 0.04 (refer to Table 1, row 24). The product of *CAMK2B* is a member of the Ca^{2+}/calmodulin-dependent protein kinase subfamily. *CAMK2B* is involved in several pathways, such as the melanogenesis pathway and the neurotrophin signaling pathway. Because *CAMK2B* is

downstream of the gene *CALM1*, CALM1 phosphorylates and thereby activates CAMK2B. In addition, CAMK2B is also involved in the Wnt/Ca^{2+} pathway, which dephosphorylates the NFAT transcription factor family, which in turn induces the expression of genes such as CD40L, CTLA-4, and FasL, finally participating in the process of cell fate decision. Recently, research has demonstrated that CAMK2B protects neurons from homocysteine-induced apoptosis with the involvement of the HIF-1α signal pathway [90].

3.2.22. CBLB. CBLB (betweenness: 5, permutation FDR: 0.042; refer to Table 1, row 25) is an ubiquitin ligase. CBLB can ubiquitinate other proteins to influence biological processes. Cbl-b contributes to the apoptosis induced by the chemotherapy in rat basophilic leukemia cells through inhibiting PI3K/Akt activation and increasing MEK/ERK activation [91]. Downregulating *Cbl-b* by shRNA resulted in strongly activating ERK, JNK, and p38 MAPK [92] and upregulating DR4 and DR5 in the presence of bufalin in MDA-MB-231 and MCF-7 cells. Moreover, the ubiquitin ligase Cbl-b negatively regulates the PI3K/Akt pathway. Therefore, we presume that Cbl-b indirectly mediates cell survival and apoptosis, which requires further experimental exploration.

3.2.23. JUN. The betweenness of this gene was 211 and its permutation FDR was 0.044 (refer to Table 1, row 26). JUN regulates gene expression by interacting directly with target DNA sequences. JNK/P38 MAP kinase pathway is crucial in regulating apoptosis, proliferation, differentiation, and inflammation. JUN can be activated by JNK, and JNK in turn transactivates downstream gene expression to perform these functions. However, the details of this regulatory mechanism require further research.

4. Conclusion

This contribution attempted to provide a better comprehension of apoptosis by identifying novel apoptosis-related genes. An existing computational method was applied with a weighted graph, constructed by protein-protein interaction information, to search for possible genes related to apoptosis. The analyses of the obtained genes further suggest that they are related to apoptosis.

Conflict of Interests

The authors declare that there is no conflict of interests.

Authors' Contribution

Baoman Wang and Fei Yuan contributed equally to this work.

Acknowledgments

This contribution was supported by the National Basic Research Program of China (2011CB510101 and 2011CB510102), the National Natural Science Foundation of China (31371335), and the Innovation Program of Shanghai Municipal Education Commission (12ZZ087).

References

[1] A. Glücksmann, "Cell deaths in normal vertebrate ontogeny," *Biological Reviews*, vol. 26, no. 1, pp. 59–86, 1951.

[2] R. A. Lockshin and Z. Zakeri, "Programmed cell death and apoptosis: origins of the theory," *Nature Reviews Molecular Cell Biology*, vol. 2, no. 7, pp. 545–550, 2001.

[3] Z. Hongmei, "Extrinsic and intrinsic apoptosis signal pathway review," in *Apoptosis and Medicine*, T. M. Ntuli, Ed., chapter 1, InTech, 2012.

[4] L. Portt, G. Norman, C. Clapp, M. Greenwood, and M. T. Greenwood, "Anti-apoptosis and cell survival: a review," *Biochimica et Biophysica Acta—Molecular Cell Research*, vol. 1813, no. 1, pp. 238–259, 2011.

[5] M. R. Sprick and H. Walczak, "The interplay between the Bcl-2 family and death receptor-mediated apoptosis," *Biochimica et Biophysica Acta*, vol. 1644, no. 2-3, pp. 125–132, 2004.

[6] L. Duprez, E. Wirawan, T. V. Berghe, and P. Vandenabeele, "Major cell death pathways at a glance," *Microbes and Infection*, vol. 11, no. 13, pp. 1050–1062, 2009.

[7] E. E. Varfolomeev and A. Ashkenazi, "Tumor necrosis factor: an apoptosis JuNKie?" *Cell*, vol. 116, no. 4, pp. 491–497, 2004.

[8] G. Scaffidi, I. Schmitz, J. Zha, S. J. Korsmeyer, P. H. Krammer, and M. E. Peter, "Differential modulation of apoptosis sensitivity in CD95 type I and type II cells," *The Journal of Biological Chemistry*, vol. 274, no. 32, pp. 22532–22538, 1999.

[9] P. Meier, A. Finch, and G. Evan, "Apoptosis in development," *Nature*, vol. 407, no. 6805, pp. 796–801, 2000.

[10] V. Zuzarte-Luís and J. M. Hurlé, "Programmed cell death in the developing limb," *International Journal of Developmental Biology*, vol. 46, no. 7, pp. 871–876, 2002.

[11] E. H. Lindberg, J. Schmidt-Mende, A. M. Forsblom, B. Christensson, B. Fadeel, and B. Zhivotovsky, "Apoptosis in refractory anaemia with ringed sideroblasts is initiated at the stem cell level and associated with increased activation of caspases," *British Journal of Haematology*, vol. 112, no. 3, pp. 714–726, 2001.

[12] L. Müllauer, P. Gruber, D. Sebinger, J. Buch, S. Wohlfart, and A. Chott, "Mutations in apoptosis genes: a pathogenetic factor for human disease," *Mutation Research—Reviews in Mutation Research*, vol. 488, no. 3, pp. 211–231, 2001.

[13] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, 2000.

[14] Y. Zhang and C. Huang, "Targeting adipocyte apoptosis: a novel strategy for obesity therapy," *Biochemical and Biophysical Research Communications*, vol. 417, no. 1, pp. 1–4, 2012.

[15] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Research*, vol. 40, no. 1, pp. D109–D114, 2012.

[16] L. J. Jensen, M. Kuhn, M. Stark et al., "STRING 8—a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Research*, vol. 37, no. 1, pp. D412–D416, 2009.

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

[18] M. Jiang, Y. Chen, Y. Zhang et al., "Identification of hepatocellular carcinoma related genes with k-th shortest paths in a protein-protein interaction network," *Molecular BioSystems*, vol. 9, no. 11, pp. 2720–2728, 2013.

[19] B.-Q. Li, J. You, L. Chen et al., "Identification of lung-cancer-related genes with the shortest path approach in a protein-protein interaction network," *BioMed Research International*, vol. 2013, Article ID 267375, 8 pages, 2013.

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

[21] J. Y. Chung, Y. C. Park, H. Ye, and H. Wu, "All TRAFs are not created equal: common and distinct molecular mechanisms of TRAF-mediated signal transduction," *Journal of Cell Science*, vol. 115, no. 4, pp. 679–688, 2002.

[22] F. X. Pimentel-Muiños and B. Seed, "Regulated commitment of TNF receptor signaling: a molecular switch for death or activation," *Immunity*, vol. 11, no. 6, pp. 783–793, 1999.

[23] W. Ji, Y. Li, T. Wan et al., "Both internalization and AIP1 association are required for tumor necrosis factor receptor 2-mediated JNK signaling," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 32, no. 9, pp. 2271–2279, 2012.

[24] C. D. White, M. D. Brown, and D. B. Sacks, "IQGAPs in cancer: a family of scaffold proteins underlying tumorigenesis," *FEBS Letters*, vol. 583, no. 12, pp. 1817–1824, 2009.

[25] K. Nabeshima, Y. Shimao, T. Inoue, and M. Koono, "Immunohistochemical analysis of IQGAP1 expression in human colorectal carcinomas: its overexpression in carcinomas and association with invasion fronts," *Cancer Letters*, vol. 176, no. 1, pp. 101–109, 2002.

[26] L. Jadeski, J. M. Mataraza, H.-W. Jeong, Z. Li, and D. B. Sacks, "IQGAP1 stimulates proliferation and enhances tumorigenesis of human breast epithelial cells," *The Journal of Biological Chemistry*, vol. 283, no. 2, pp. 1008–1017, 2008.

[27] V. Patel, B. L. Hood, A. A. Molinolo et al., "Proteomic analysis of laser-captured paraffin-embedded tissues: a molecular portrait of head and neck cancer progression," *Clinical Cancer Research*, vol. 14, no. 4, pp. 1002–1014, 2008.

[28] M. Roy, Z. Li, and D. B. Sacks, "IQGAP1 binds ERK2 and modulates its activity," *The Journal of Biological Chemistry*, vol. 279, no. 17, pp. 17329–17337, 2004.

[29] C. D. White, H. H. Erdemir, and D. B. Sacks, "IQGAP1 and its binding proteins control diverse biological functions," *Cellular Signalling*, vol. 24, no. 4, pp. 826–834, 2012.

[30] A. Sato, T. Naito, A. Hiramoto et al., "Association of RNase L with a Ras GTPase-activating-like protein IQGAP1 in mediating the apoptosis of a human cancer cell-line," *FEBS Journal*, vol. 277, no. 21, pp. 4464–4473, 2010.

[31] X. Yang, Q. Wang, Z. Gao et al., "Proprotein convertase furin regulates apoptosis and proliferation of granulosa cells in the rat ovary," *PLoS ONE*, vol. 8, no. 2, Article ID e50479, 2013.

[32] M. S. Caetano, A. Vieira-de-Abreu, L. K. Teixeira, M. B. F. Werneck, M. A. Barcinski, and J. P. B. Viola, "NFATC2 transcription factor regulates cell cycle progression during lymphocyte activation: evidence of its involvement in the control of cyclin gene expression," *The FASEB Journal*, vol. 16, no. 14, pp. 1940–1942, 2002.

[33] A. Rao, C. Luo, and P. G. Hogan, "Transcription factors of the NFAT family: regulation and function," *Annual Review of Immunology*, vol. 15, pp. 707–747, 1997.

[34] B. K. Robbs, A. L. S. Cruz, M. B. F. Werneck, G. P. Mognol, and J. P. B. Viola, "Dual roles for NFAT transcription factor genes as oncogenes and tumor suppressors," *Molecular and Cellular Biology*, vol. 28, no. 23, pp. 7168–7181, 2008.

[35] L. S. Arabanian, S. Kujawski, I. Habermann, G. Ehninger, and A. Kiani, "Regulation of fas/fas ligand-mediated apoptosis by nuclear factor of activated T cells in megakaryocytes," *British Journal of Haematology*, vol. 156, no. 4, pp. 523–534, 2012.

[36] V. Perotti, P. Baldassari, I. Bersani et al., "NFATc2 is a potential therapeutic target in human melanoma," *Journal of Investigative Dermatology*, vol. 132, no. 11, pp. 2652–2660, 2012.

[37] S. Jayanthi, X. Deng, B. Ladenheim et al., "Calcineurin/NFAT-induced up-regulation of the Fas ligand/Fas death pathway is involved in methamphetamine-induced neuronal apoptosis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 3, pp. 868–873, 2005.

[38] C. J. Holtz-Heppelmann, A. Algeciras, A. D. Badley, and C. V. Paya, "Transcriptional regulation of the human FasL promoter-enhancer region," *The Journal of Biological Chemistry*, vol. 273, no. 8, pp. 4416–4423, 1998.

[39] K. M. Latinis, L. A. Norian, S. L. Eliason, and G. A. Koretzky, "Two NFAT transcription factor binding sites participate in the regulation of CD95 (Fas) ligand expression in activated human T cells," *Journal of Biological Chemistry*, vol. 272, no. 50, pp. 31427–31434, 1997.

[40] L. D. S. Carvalho, L. K. Teixeira, N. Carrossini et al., "The NFAT1 transcription factor is a repressor of cyclin A2 gene expression," *Cell Cycle*, vol. 6, no. 14, pp. 1789–1795, 2007.

[41] J. P. B. Viola, L. D. S. Carvalho, B. P. F. Fonseca, and L. K. Teixeira, "NFAT transcription factors: from cell cycle to tumor development," *Brazilian Journal of Medical and Biological Research*, vol. 38, no. 3, pp. 335–344, 2005.

[42] H. Okamura, J. Aramburu, C. García-Rodríguez et al., "Concerted dephosphorylation of the transcription factor NFAT1 induces a conformational switch that regulates transcriptional activity," *Molecular Cell*, vol. 6, no. 3, pp. 539–550, 2000.

[43] S. F. Oliveria, L. L. Gomez, and M. L. Dell'Acqua, "Imaging kinase–AKAP79–phosphatase scaffold complexes at the plasma membrane in living cells using FRET microscopy," *The Journal of Cell Biology*, vol. 160, no. 1, pp. 101–112, 2003.

[44] D. Willoughby, N. Masada, S. Wachten et al., "AKAP79/150 interacts with AC8 and regulates Ca^{2+}-dependent cAMP synthesis in pancreatic and neuronal systems," *The Journal of Biological Chemistry*, vol. 285, no. 26, pp. 20328–20342, 2010.

[45] Y. Shi, "A structural view of mitochondria-mediated apoptosis," *Nature Structural & Molecular Biology*, vol. 8, no. 5, pp. 394–401, 2001.

[46] C. Du, M. Fang, Y. Li, L. Li, and X. Wang, "Smac, a mitochondrial protein that promotes cytochrome c-dependent caspase activation by eliminating IAP inhibition," *Cell*, vol. 102, no. 1, pp. 33–42, 2000.

[47] J. Chai, C. Du, J.-W. Wu, S. Kyin, X. Wang, and Y. Shi, "Structural and biochemical basis of apoptotic activation by Smac/DIABLO," *Nature*, vol. 406, no. 6798, pp. 855–862, 2000.

[48] S. M. Srinivasula, P. Datta, X.-J. Fan, T. Fernandes-Alnemri, Z. Huang, and E. S. Alnemri, "Molecular determinants of the caspase-promoting activity of Smac/DIABLO and its role in the death receptor pathway," *The Journal of Biological Chemistry*, vol. 275, no. 46, pp. 36152–36157, 2000.

[49] Y. M. Anguiano-Hernandez, A. Chartier, and S. Huerta, "Smac/DIABLO and colon cancer," *Anti-Cancer Agents in Medicinal Chemistry*, vol. 7, no. 4, pp. 467–473, 2007.

[50] A. M. Verhagen, P. G. Ekert, M. Pakusch et al., "Identification of DIABLO, a mammalian protein that promotes apoptosis by binding to and antagonizing IAP proteins," *Cell*, vol. 102, no. 1, pp. 43–53, 2000.

[51] X. D. Zhang, X. Y. Zhang, C. P. Gray, T. Nguyen, and P. Hersey, "Tumor necrosis factor-related apoptosis-inducing ligand-induced apoptosis of human melanoma is regulated by Smac/DIABLO release from mitochondria," *Cancer Research*, vol. 61, no. 19, pp. 7339–7348, 2001.

[52] J. P. Carson, M. Behnam, J. N. Sutton et al., "Smac is required for cytochrome c-induced apoptosis in prostate cancer LNCaP cells," *Cancer Research*, vol. 62, no. 1, pp. 18–23, 2002.

[53] Y. S. Roh, J. Song, and E. Seki, "TAK1 regulates hepatic cell survival and carcinogenesis," *Journal of Gastroenterology*, vol. 49, no. 2, pp. 185–194, 2014.

[54] K. Yamaguchi, S.-I. Nagai, J. Ninomiya-Tsuji et al., "XIAP, a cellular member of the inhibitor of apoptosis protein family, links the receptors to TAB1–TAK1 in the BMP signaling pathway," *The EMBO Journal*, vol. 18, no. 1, pp. 179–187, 1999.

[55] M. Germana Sanna, J. da Silva Correia, O. Ducrey et al., "IAP suppression of apoptosis involves distinct mechanisms: the TAK1/JNK1 signaling cascade and caspase inhibition," *Molecular and Cellular Biology*, vol. 22, no. 6, pp. 1754–1766, 2002.

[56] B. Fiedler, R. Feil, F. Hofmann et al., "cGMP-dependent protein kinase type I inhibits TAB1-p38 mitogen-activated protein kinase apoptosis signaling in cardiac myocytes," *Journal of Biological Chemistry*, vol. 281, no. 43, pp. 32831–32840, 2006.

[57] S. E. Martin, Z.-H. Wu, K. Gehlhaus et al., "Rnai screening identifies tak1 as a potential target for the enhanced efficacy of topoisomerase inhibitors," *Current Cancer Drug Targets*, vol. 11, no. 8, pp. 976–986, 2011.

[58] B. Fadeel, B. Zhivotovsky, and S. Orrenius, "All along the watchtower: on the regulation of apoptosis regulators," *The FASEB Journal*, vol. 13, no. 13, pp. 1647–1657, 1999.

[59] K. M. McGinnis, M. E. Gnegy, Y. H. Park, N. Mukerjee, and K. K. W. Wang, "Procaspase-3 and poly (ADP) ribose polymerase (PARP) are calpain substrates," *Biochemical and Biophysical Research Communications*, vol. 263, no. 1, pp. 94–99, 1999.

[60] M. K. T. Squier and J. J. Cohen, "Calpain, an upstream regulator of thymocyte apoptosis," *The Journal of Immunology*, vol. 158, no. 8, pp. 3690–3697, 1997.

[61] C. Bertoli, T. Copetti, E. W.-F. Lam, F. Demarchi, and C. Schneider, "Calpain small-1 modulates Akt/FoxO3A signaling and apoptosis through PP2A," *Oncogene*, vol. 28, no. 5, pp. 721–733, 2009.

[62] B. Benayoun, S. Baghdiguian, A. Lajmanovich et al., "NF-κB-dependent expression of the antiapoptotic factor c-FLIP is regulated by calpain 3, the protein involved in limb-girdle muscular dystrophy type 2A," *The FASEB Journal*, vol. 22, no. 5, pp. 1521–1529, 2008.

[63] T. K. Neklesa and R. W. Davis, "A genome-wide screen for regulators of TORC1 in response to amino acid starvation reveals a conserved Npr2/3 complex," *PLoS Genetics*, vol. 5, no. 6, Article ID e1000515, 2009.

[64] Y. Wei and M. A. Lilly, "The TORC1 inhibitors Nprl2 and Nprl3 mediate an adaptive response to amino-acid starvation in Drosophila," *Cell Death & Differentiation*, vol. 21, no. 9, pp. 1460–1468, 2014.

[65] F. R. R. Mangone, F. Walder, S. Maistro et al., "Smad2 and Smad6 as predictors of overall survival in oral squamous cell carcinoma patients," *Molecular Cancer*, vol. 9, article 106, 2010.

[66] H.-S. Jeon, T. Dracheva, S.-H. Yang et al., "SMAD6 contributes to patient survival in non-small cell lung cancer and its knockdown reestablishes TGF-β homeostasis in lung cancer cells," *Cancer Research*, vol. 68, no. 23, pp. 9686–9692, 2008.

[67] M. Bakhshayesh, F. Zaker, M. Hashemi, M. Katebi, and M. Solaimani, "TGF- β1-mediated apoptosis associated with SMAD-dependent mitochondrial Bcl-2 expression," *Clinical Lymphoma, Myeloma and Leukemia*, vol. 12, no. 2, pp. 138–143, 2012.

[68] S. M. Jung, J.-H. Lee, J. Park et al., "Smad6 inhibits non-canonical TGF-β1 signalling by recruiting the deubiquitinase A20 to TRAF6," *Nature Communications*, vol. 4, article 2562, 2013.

[69] Y. Wang, J. Wu, B. Lin et al., "Galangin suppresses HepG2 cell proliferation by activating the TGF-β receptor/Smad pathway," *Toxicology*, vol. 326, pp. 9–17, 2014.

[70] M. Federici, M. L. Hribal, M. Ranalli et al., "The common Arg972 polymorphism in insulin receptor substrate-1 causes apoptosis of human pancreatic islets," *The FASEB Journal*, vol. 15, no. 1, pp. 22–24, 2001.

[71] M. Dromard, G. Bompard, M. Glondu-Lassis, C. Puech, D. Chalbos, and G. Freiss, "The putative tumor suppressor gene PTPN13/PTPL1 induces apoptosis through insulin receptor substrate-1 dephosphorylation," *Cancer Research*, vol. 67, no. 14, pp. 6806–6813, 2007.

[72] P. Igarashi and S. Somlo, "Genetics and pathogenesis of polycystic kidney disease," *Journal of the American Society of Nephrology*, vol. 13, no. 9, pp. 2384–2398, 2002.

[73] B. Loureiro, L. J. Oliveira, M. G. Favoreto, and P. J. Hansen, "Colony-stimulating factor 2 inhibits induction of apoptosis in the bovine preimplantation embryo," *American Journal of Reproductive Immunology*, vol. 65, no. 6, pp. 578–588, 2011.

[74] M. Subramanian, E. B. Thorp, and I. Tabas, "Identification of a non-growth factor role for GM-CSF in advanced atherosclerosis: promotion of macrophage apoptosis and plaque necrosis through IL-23 signaling," *Circulation Research*, vol. 116, no. 2, pp. e13–e24, 2015.

[75] P. Hikisz and Z. M. Kiliańska, "Puma, a critical mediator of cell death—one decade on from its discovery," *Cellular & Molecular Biology Letters*, vol. 17, no. 4, pp. 646–669, 2012.

[76] J. E. Chipuk and D. R. Green, "PUMA cooperates with direct activator proteins to promote mitochondrial outer membrane permeabilization and apoptosis," *Cell Cycle*, vol. 8, no. 17, pp. 2692–2696, 2009.

[77] Y. I. Choi, S. H. Jeon, J. Jang et al., "Notch1 confers a resistance to glucocorticoid-induced apoptosis on developing thymocytes by down-regulating SRG3 expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 18, pp. 10267–10272, 2001.

[78] Z. Wang, Y. Zhang, Y. Li, S. Banerjee, J. Liao, and F. H. Sarkar, "Down-regulation of Notch-1 contributes to cell growth inhibition and apoptosis in pancreatic cancer cells," *Molecular Cancer Therapeutics*, vol. 5, no. 3, pp. 483–493, 2006.

[79] Z. Lu, H. Liu, L. Xue, P. Xu, T. Gong, and G. Hou, "An activated Notch1 signaling pathway inhibits cell proliferation and induces apoptosis in human esophageal squamous cell carcinoma cell line EC9706," *International Journal of Oncology*, vol. 32, no. 3, pp. 643–651, 2008.

[80] A. Rangarajan, R. Syal, S. Selvarajah, O. Chakrabarti, A. Sarin, and S. Krishna, "Activated Notch1 signaling cooperates with papillomavirus oncogenes in transformation and generates resistance to apoptosis on matrix withdrawal through PKB/Akt," *Virology*, vol. 286, no. 1, pp. 23–30, 2001.

[81] P. Nair, K. Somasundaram, and S. Krishna, "Activated Notch1 inhibits p53-induced apoptosis and sustains transformation by

human papillomavirus type 16 E6 and E7 oncogenes through a PI3K-PKB/Akt-dependent pathway," *Journal of Virology*, vol. 77, no. 12, pp. 7106–7112, 2003.

[82] N. K. Tonks and B. G. Neel, "Combinatorial control of the specificity of protein tyrosine phosphatases," *Current Opinion in Cell Biology*, vol. 13, no. 2, pp. 182–195, 2001.

[83] H. Song, Z. Zhang, and L. Wang, "Small interference RNA against PTP-1B reduces hypoxia/reoxygenation induced apoptosis of rat cardiomyocytes," *Apoptosis*, vol. 13, no. 3, pp. 383–393, 2008.

[84] G. Taheripak, S. Bakhtiyari, M. Rajabibazl, P. Pasalar, and R. Meshkani, "Protein tyrosine phosphatase 1B inhibition ameliorates palmitate-induced mitochondrial dysfunction and apoptosis in skeletal muscle cells," *Free Radical Biology and Medicine*, vol. 65, pp. 1435–1446, 2013.

[85] V. Sangwan, G. N. Paliouras, A. Cheng, N. Dubé, M. L. Tremblay, and M. Park, "Protein-tyrosine phosphatase 1B deficiency protects against Fas-induced hepatic failure," *The Journal of Biological Chemistry*, vol. 281, no. 1, pp. 221–228, 2006.

[86] S. Peng, S. Lalani, J. W. Leavenworth, I.-C. Ho, and M. E. Pauza, "c-Maf interacts with c-Myb to down-regulate Bcl-2 expression and increase apoptosis in peripheral CD4 cells," *European Journal of Immunology*, vol. 37, no. 10, pp. 2868–2880, 2007.

[87] T. K. Hale, C. Myers, R. Maitra, T. Kolzau, M. Nishizawa, and A. W. Braithwaite, "Maf transcriptionally activates the mouse p53 promoter and causes a p53-dependent cell death," *The Journal of Biological Chemistry*, vol. 275, no. 24, pp. 17991–17999, 2000.

[88] S. Peng, H. Wu, Y.-Y. Mo, K. Watabe, and M. E. Pauza, "C-Maf increases apoptosis in peripheral CD8 cells by transactivating *Caspase 6*," *Immunology*, vol. 127, no. 2, pp. 267–278, 2009.

[89] S. Bourteele, K. Oesterle, A. O. Weinzierl et al., "Alteration of NF-kappaB activity leads to mitochondrial apoptosis after infection with pathological prion protein," *Cellular Microbiology*, vol. 9, no. 9, pp. 2202–2217, 2007.

[90] M. Fang, C. Feng, Y.-X. Zhao, and X.-Y. Liu, "Camk2b protects neurons from homocysteine-induced apoptosis with the involvement of HIF-1α signal pathway," *International Journal of Clinical and Experimental Medicine*, vol. 7, no. 7, pp. 1659–1668, 2014.

[91] X. Qu, Y. Li, J. Liu et al., "Cbl-b promotes chemotherapy-induced apoptosis in rat basophilic leukemia cells by suppressing PI3K/Akt activation and enhancing MEK/ERK activation," *Molecular and Cellular Biochemistry*, vol. 340, no. 1-2, pp. 107–114, 2010.

[92] S. Yan, X. Qu, C. Xu et al., "Down-regulation of Cbl-b by bufalin results in up-regulation of DR4/DR5 and sensitization of TRAIL-induced apoptosis in breast cancer cells," *Journal of Cancer Research and Clinical Oncology*, vol. 138, no. 8, pp. 1279–1289, 2012.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 471076, 8 pages
http://dx.doi.org/10.1155/2015/471076

Research Article

Cell Pluripotency Levels Associated with Imprinted Genes in Human

Liyun Yuan,[1] Xiaoyan Tang,[1] Binyan Zhang,[2] and Guohui Ding[1,3]

[1]*Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, SIBS, CAS, Shanghai 200031, China*
[2]*BasePair BioTechonology Co., Ltd., Shanghai 200235, China*
[3]*Shanghai Center for Bioinformation Technology, Shanghai 200235, China*

Correspondence should be addressed to Guohui Ding; gwding@sibs.ac.cn

Received 20 January 2015; Revised 16 March 2015; Accepted 17 March 2015

Academic Editor: Tao Huang

Pluripotent stem cells are exhibited similarly in the morphology, gene expression, growth properties, and epigenetic modification with embryonic stem cells (ESCs). However, it is still controversial that the pluripotency of induced pluripotent stem cell (iPSC) is much inferior to ESC, and the differentiation capacity of iPSC and ESC can also be separated by transcriptome and epigenetics. miRNAs, which act in posttranscriptional regulation of gene expression and are involved in many basic cellular processes, may reveal the answer. In this paper, we focused on identifying the hidden relationship between miRNAs and imprinted genes in cell pluripotency. Total miRNA expression patterns in iPSC and ES cells were comprehensively analysed and linked with human imprinted genes, which show a global picture of their potential function in pluripotent level. A new CPA4-KLF14 region which locates in chromosomal homologous segments (CHSs) within mammals and include both imprinted genes and significantly expressed miRNAs was first identified. Molecular network analysis showed genes interacted with imprinted genes closely and enriched in modules such as cancer, cell death and survival, and tumor morphology. This imprinted region may provide a new look for those who are interested in cell pluripotency of hiPSCs and hESCs.

1. Background

Undifferentiated embryonic stem cells (ESCs) share the ability to self-renew and differentiate into various different lineages which is fundamental to understanding human development, tissue regeneration, and healthy homeostatic turnover [1, 2]. It has been demonstrated that, only by addition of a few defined factors, pluripotent stem cells can be directly generated from fibroblast cultures [3, 4] and are exhibited similarly in morphology, gene expression, growth properties, and epigenetic modification as ESCs [5, 6]. The reprogramming techniques open eyes for understanding the developmental mechanisms in assigning cells for particular fates and are also bound to provide us with questions about the different pluripotency levels. The pluripotency of induced pluripotent stem cell (iPSC) is much inferior to ESCs, as can be separated by transcriptome and epigenetics. Several recent studies focused on identifying the hidden difference between ES and iPSCs and tried to break the block on the progress of its basic research and clinical application [7, 8].

MicroRNAs (miRNAs) are a class of noncoding RNA genes whose products are 22nt sequences that play significant roles in the regulation of translation and degradation of mRNAs through base pairing to partially complementary sites in the untranslated regions (UTRs) of the message [9]. miRNAs were intensely investigated to identify their mechanisms of action in cell development and progression [10–12]. Studies have reported that several miRNAs expressed diversely in iPSC and hESC have great effect on pluripotency [13, 14]. Interestingly, through analysing the small RNA expression patterns in iPSC, a Dlk1-Dio3 region with a large cluster of miRNAs as well as imprinted genes was identified in fully pluripotent stem cells [8]. This imprinted genomic region was found to be repressed in the cells with partial pluripotency, and aberrant silencing of this region in mouse was found to induce pluripotent stem cells. Mammals genomic imprinting is possibly caused by an interparental genetic conflict to control maternal-dependent growth of the offspring [15] and was found to be linked to a number of human behavioral and developmental disorders as well as

a variety of pediatric and adult malignancies [16]. While imprinted genes were also found near differentially expressed miRNAs in hESCs [8, 17], these phenomena suggested that miRNAs may influence early development together with imprinted genes. In order to scan the association between miRNAs and imprinted genes in regulating cell pluripotency levels, we focus on miRNA expression profiles in both ESC and iPSC. miRNA clusters were found near imprinted genes on most chromosomes and an imprinted region which we think may influence hESCs pluripotency capacity will be discussed in the following paragraph.

2. Results and Discussion

2.1. miRNA Expression Analysis in Cells with Diverse Pluripotent Levels. We analysed a miRNA expression dataset stored in the GEO database which includes 800 miRNAs from normal human dermal fibroblasts (NHDFs), hiPSCs, and hESCs [18] (Section 4). Linear model [19] was implemented in pairwise comparison between hESCs and hiPSCs, NHDFs and hiPSCs, and hESCs and NHDFs, respectively (Figure 1(a)), by Limma package of Bioconductor. After normalization and statistical analysis, 174 miRNAs were identified and differentially expressed between pairwise comparisons (FDR $P \leq 0.05$) (Figures 1(b) and 1(c)) and 131 miRNAs were significantly regulated in both hiPSCs and hESCs compared to NHDFs, among them, 79 with at least 2-fold change were considered to be directly related to cell pluripotency and may dominate during individual development progress (see Supplementary Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/471076). These results included 14q32 microRNA clusters neighbouring the DLK1-DIO3 imprinted regions as well as those miRNAs in pluripotency, reprogramming, and cell fate induction [13, 20]. Besides, 40 miRNAs were found to be significantly regulated in hiPSCs (or hESCs) and slightly in hESCs (or hiPSCs) compared to NHDFs. These miRNAs may influence pluripotency levels between hESCs and hiPSCs.

To see the biological function of these differentially expressed miRNAs, miRNA targets were predicted to examine whether they were associated with cell pluripotency. The miRanda (release 2010) [21] and TargetScan (release 6.2) [22] lists were combined to form an intersection of targets predicted by both algorithms to strengthen the analysis reliability. Selected genes were then collectively subjected to pathway analysis using the R package. The enriched pathways are relevant to neural connectivity and synaptic plasticity, such as axon guidance, long-term potentiation, neurotrophin signaling pathway, and calcium signalling pathway (Supplementary Table 2). The results also show relevance to carcinogenesis including prostate cancer, pancreatic cancer, and pancreatic cancer, which is directly related to cell development and proliferation. The GO enrichment analysis was also performed as shown in Supplementary Table 2.

2.2. Differentially Expressed miRNAs Clusters Overlapped with Imprinted Gene Regions. In order to identify the association between the 131 differentially expressed miRNAs and imprinted genes, PTMCluster algorithm [23] was used to find the statistically significant miRNA clusters (Section 4). 12 differentially expressed clusters including 65 miRNAs were located on chromosomes, the length of which differed from 483bp (chromosome 4) to 111Mb (chromosome 1). Human imprinted genes were collected from Geneimprint website (http://www.geneimprint.com/) and totally 30 of them fall in or near the above miRNA clustering regions. To demonstrate that differentially expressed miRNAs were enriched nearby imprinted genomic regions other than randomly being dispersed, a permutation test was performed (Section 4). 1000 sets of miRNAs with the same quantity as the differentially expressed miRNAs were randomly generated and miRNA enrichment regions of each set were clustered using the same algorithm. The number of imprinted genes that fall in or near the miRNA clusters of each random set was, respectively, calculated and compared to 30 imprinted genes which locate in differentially expressed miRNA clusters. The results show that differentially expressed miRNAs significantly enriched near the imprinted genes ($P = 0.037$).

We then focus on miRNA enrichment region on chromosomes 1, 5, 6, 7, 9, 14, and 16 including most of the miRNAs (blue) and imprinted genes (pink) (Figure 2). Lines represent the pairwise interaction, while dashed lines represent the predicted target gene of miRNA. The genes in white are the predicted target genes other than imprinted genes. miRNA-143, miRNA-145, and miRNA-146a located on chromosome 5 have the same target gene N-ras, which is one of the Ras gene family members located on chromosome 1. N-ras functions as an important factor in a large number of biological process involving cell cycle progression, differentiation, cell proliferation, apoptosis, and cell survival by the level of Ras expression [24–26]. Besides, N-ras has induced tumorigenesis through different signalling pathways [27, 28]. miRNA-7a, miRNA-7f, and miRNA-7d clustering on chromosome 9 have the same target gene CPA4/CASP8, which was originally identified as an initiator caspase and mainly functions in the death receptor pathway of apoptosis [29, 30]. Many studies also explored the significance of the nonapoptotic function of caspase-8 and its mechanism [31, 32]. miRNA-29a, miRNA-29b-1, and miRNA-29b-2 clustering on chromosome 7 have the same target gene MEST/PEG1, loss of which will cause intrauterine growth retardation and abnormal maternal [33]. Both of the CPA4 and MEST are imprinted genes and located close on chromosome. Besides, polyubiquitin gene UBC interacted with many of the target indirectly and directly. Disruption of UBC reduces the absolute number of hematopoietic stem cells in embryonic livers [34] and embryonic lethality with defective fetal liver development [35]. Ubiquitin was also shown to exhibit various functions of phenotypes including cell development and cell cycle progression [36, 37].

2.3. miRNAs Encoded in the Imprinted CPA4-KLF14 Region Associated with Cell Pluripotency Levels. Previous studies have revealed that the Dlk1-Dio3 region in mammals will activate cell pluripotency level [8]. Here, we also find a similar imprinted CPA4-KLF14 region on chromosome 7 encoding both miRNAs and imprinted genes, which we think

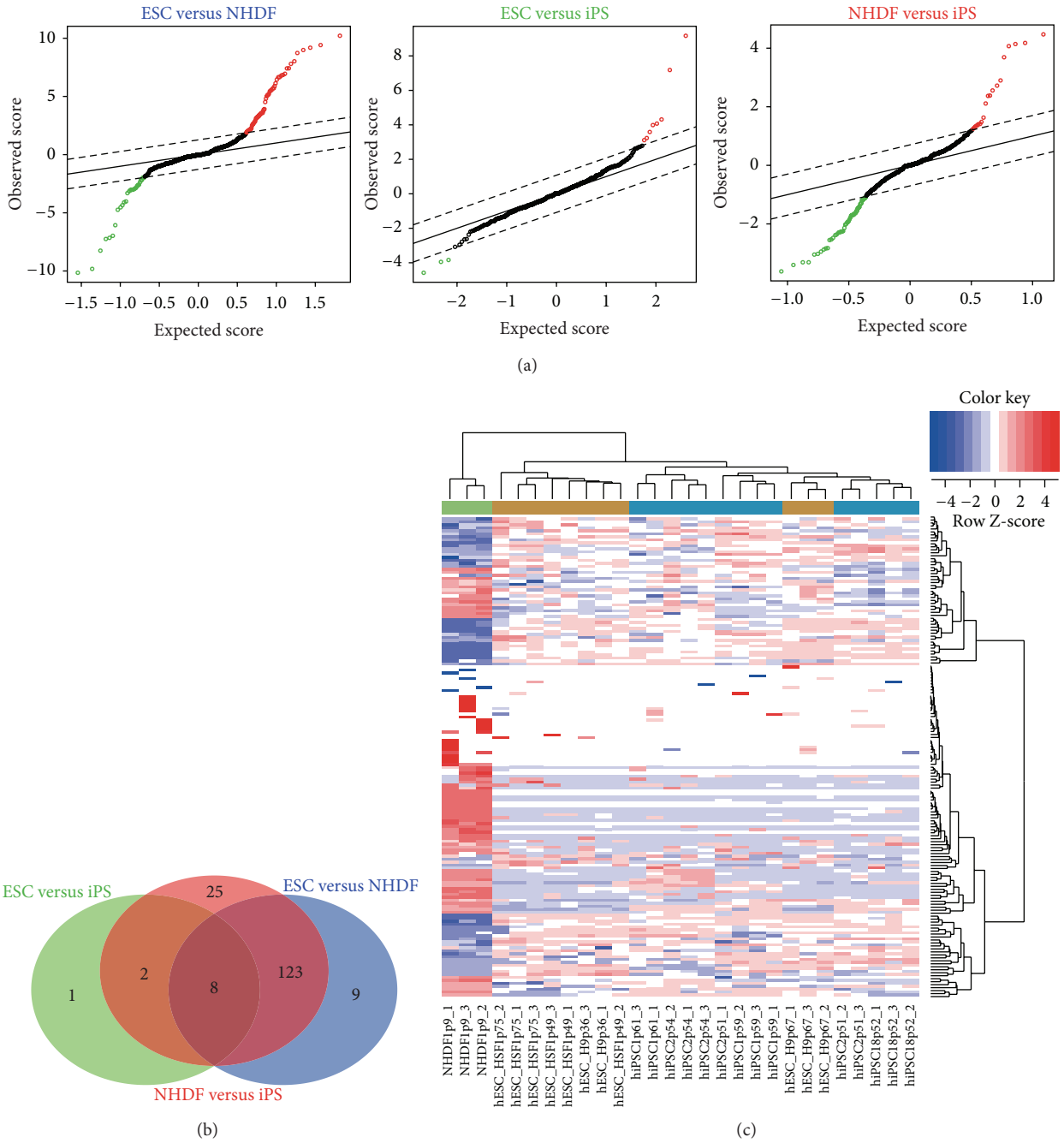


Figure 1: miRNAs expression profiles in different cells. (a) Significance analysis of microarray (SAM) plot of differentially expressed miRNA in hESCs (12), NHDF (3), and hiPSCs (15). The central solid line indicates equal expression and the upper and lower dashed lines indicate levels for significantly altered expression where FDR is 0%. (b) Venn diagram of differentially expressed miRNAs from pairwise comparisons, respectively. miRNA groups are coloured according to different comparison (green = ESC versus iPS; red = NHDF versus iPS; blue = ESC versus NHDF). (c) Heatmap showing hierarchical clustering of significantly regulated microRNA from iPS and ESCs compared with NHDFs. Blue indicates low expression and red indicates high expression.

may have effect on the extent of cell pluripotency. In our result, the abundance of hsa-miR-29a and hsa-miR-29b decreased, while hsa-miR-593 and hsa-miR-182 increased more in iPSCs and hESCs than in NHDFs. Although mir-29a/b is reduced in both hiPSCs and hESCs, they fall further in hESCs than in hiPSCs. The inhibition of mir-29a/b may enhance cell reprogramming efficiency as previous studies have showed [38]. The distance in their expression level will directly affect the extent of cell pluripotency. This region also includes five imprinted genes, CPA4, MEST, MESTIT1, COPG2, and KLF14. Gene MEST is the predicted target gene for nearby mir-29. Furthermore, KLF14 is of the same family

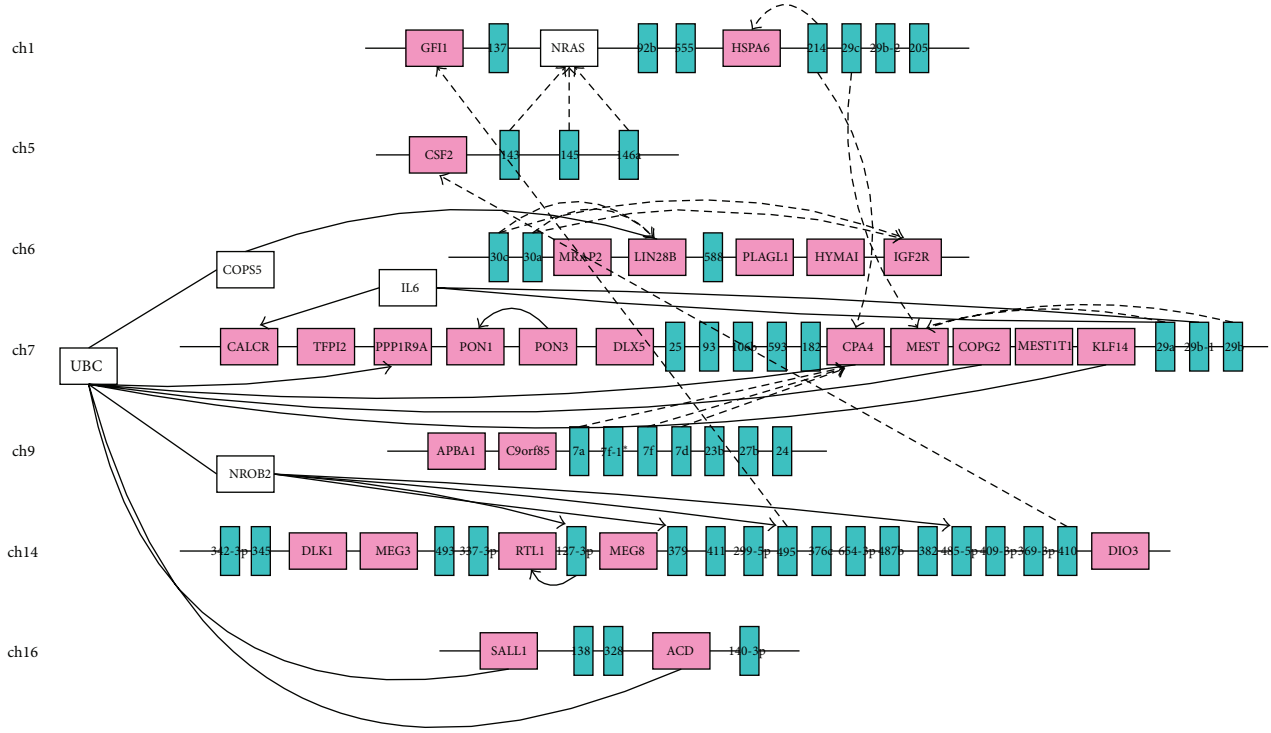


FIGURE 2: miRNA cluster regions and imprinted genes on chromosomes. Location of miRNA (blue) cluster regions and imprinted genes (pink) on chromosome, as well as the interaction between miRNAs and their target genes. The solid lines represent the interaction through IPA database, while the dotted lines represent the predicted target genes of miRNAs by both miRanda and TargetScan.

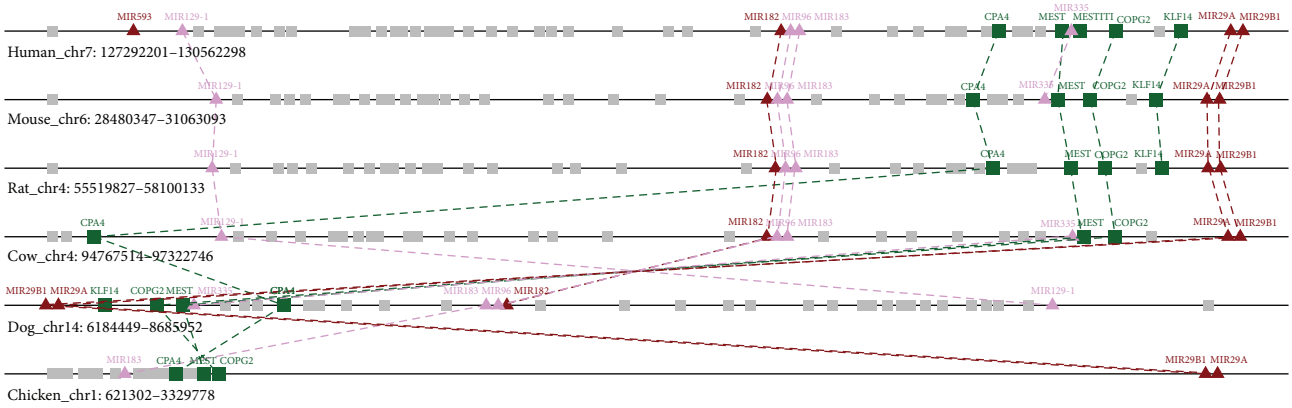


FIGURE 3: Synteny map of CPA4-KLF14 imprinted region. CHSs including CPA4-KLF14 region between human and other mammalians were constructed, respectively, and merged. The imprinted genes and miRNAs with significant regulation in hESCs or hiPSCs are shown as green rectangles and red triangles; the unaltered genes and miRNAs are shown as grey rectangles and pink triangles.

as KLF4, which is a well-known transcription factor and will reset the somatic cell epigenome during induced pluripotent stem cell generation [39]. Synteny map between human and other mammalians was constructed, respectively, and merged together (Figure 3). CPA4-KLF14 region locates in a genome segment which has conserved content chromosomal homologous segments (CHSs) in mouse ($P = 1.32E - 207$), rat ($P = 2.32E - 196$), cow ($P = 4.47E - 193$), dog ($P = 8.69E - 157$), and chicken ($P = 1.15E - 48$) (see Section 4 and Supplementary Figure 1). The gene content and order of this segment region were conserved during evolution which reflected important functional relationships between miRNAs and imprinted genes and may affect the cell fate together.

2.4. Interactions between Differentially Expressed miRNAs and Imprinted Genes. After viewing the global properties of the differentially expressed miRNAs between each pair of cells, we examined details of those 11 miRNAs altered significantly between iPSCs and ESCs to see whether the expression

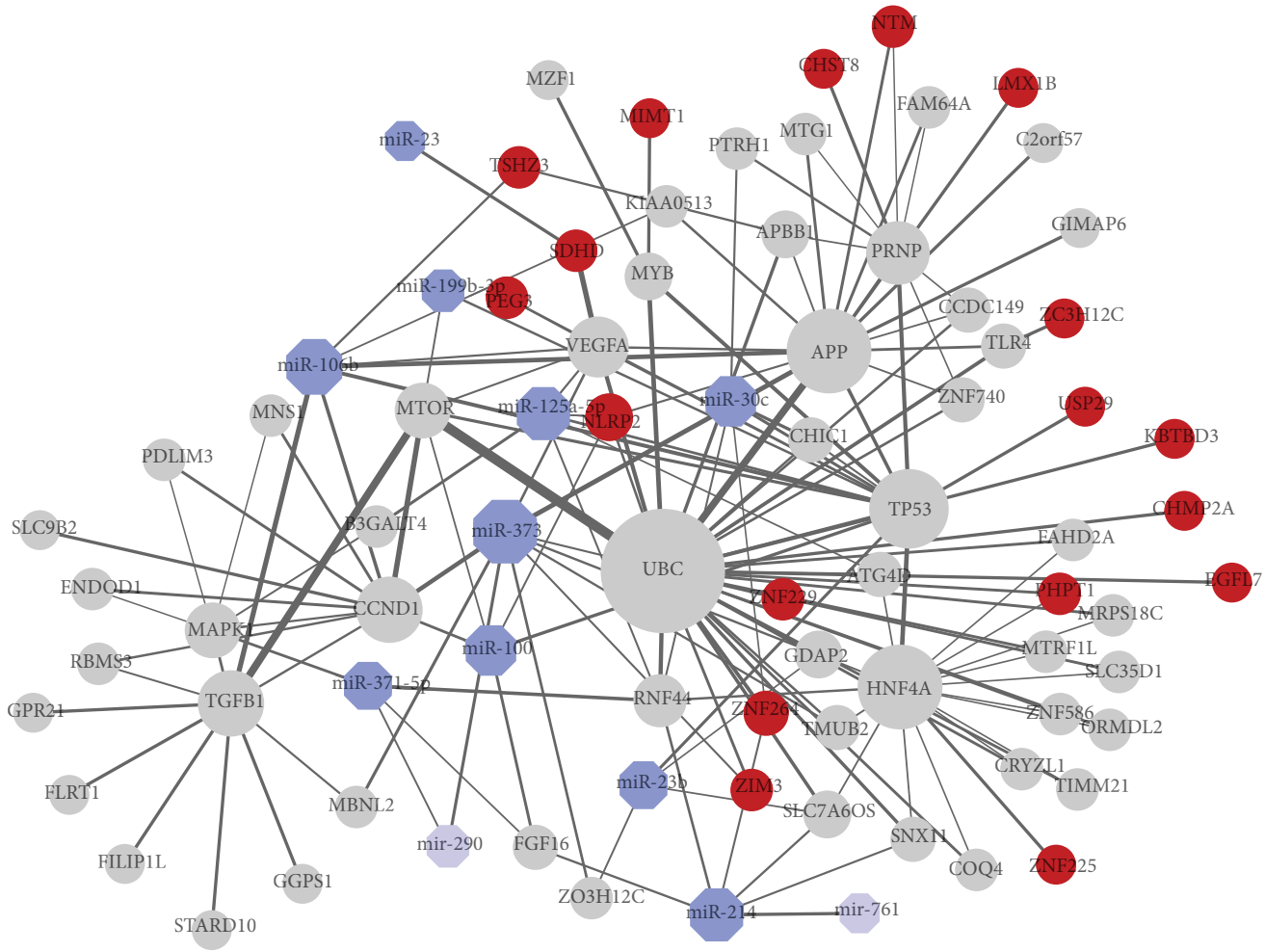


Figure 4: Molecular network of pluripotency related module between iPSC and ESC. Interactions of differentially expressed miRNAs (blue), imprinted genes (red), and other linked miRNAs (light blue) and genes (grey) from IPA database. Node size is proportional to the number of its links.

level will have effect on the pluripotency extent. Molecular network analysis was performed using IPA system to show the relationship between miRNAs and their linked genes (Figure 4). Based on topology, these differentially expressed miRNAs (blue) are connected with imprinted genes (red) closely. Diseases and function analysis showed that this network was primary centred by gene UBC, TP53, and APP and enriched in modules such as cancer, dermatological disease and conditions, organismal development, tissue development, cell death and survival, and tumor morphology (Supplementary Table 3).

Interestingly, UBC was also centred in Figure 2 and strongly interacts with several imprinted genes such as EGFL7, KBTBD3, CHMP2A, and ZC3H12C. Meanwhile, APP and TP53 are connected with UBC as well as other imprinted genes. P53 was first identified as a direct repressor of Nanog in mESCs [40]. Likewise, p53 was further proved to be functioning in apoptosis and differentiation of hESCs [41, 42] and downregulates proliferation and self-renewal of neural stem cells and hematopoietic stem cells [43, 44]. P53 is connected with miR-100, miR-30c, miR-23b, miR-125a-5p, and miR-199b-3p. APP encodes a cell surface receptor and transmembrane precursor protein amyloid beta, which decreases several cell signalling pathways associated with neurogenesis [45] and inhibits the proliferation of neural stem cells by activating the PI3K pathway [46]. These indicated that imprinted genes interact directly with pluripotency related genes.

Here, miR-373, miR-371-5p, and miR-106b show the same expression alteration but different range in hiPSCs and hESCs compared to NHDFs, which indicates that these miRNAs should be upregulated to some extent to maintain the capacity of cell pluripotency in hESCs. Researches have reported that the mmu-mir-290/hsa-mir-371/372/373 cluster expressed in trophoblast stem cells and functioned in cellular self-renewal [47, 48]. For the same reason, other miRNAs may support the cell pluripotency by different expression level. Meanwhile, other miRNAs shown in the network connected mostly with p53 may work in maintaining cell pluripotency.

3. Conclusions

In this study, we examined pluripotency-associated miRNA expression in human ESCs, iPSCs, and NHDFs and identified a significant correlation between miRNAs clusters within imprinted gene region. We then discuss one of the small regions on chromosome 7 which include significantly expressed miRNAs as well as four imprinted genes. The miRNAs encoded in this cluster have been verified involving cell reprogramming and proliferation and target the nearby imprinted gene MEST. Although the meaning of this relationship is not known, this imprinted CPA4-KLF14 region located in content conserved CHSs of mammals and may play important roles during transcriptional regulation and processing. Meanwhile, we construct the interaction network between differentially expressed miRNAs and imprinted genes and found that those miRNAs did interact with imprinted genes directly or indirectly through other genes. Regardless of the molecular function, the imprinted CPA4-KLF14 region observed in our research may provide a new look for those interested in different pluripotent level between hiPSCs and hESCs.

4. Methods

4.1. Transcriptome Profiles Analysis. Microarray dataset GSE16654 for 1 NDF line, 5 iPSC lines, and 3 ESC lines was obtained from GEO (http://www.ncbi.nlm.nih.gov/geo/). Each sample has three biological replicates. As discussed in [49], iPSCs were generated from NDFs by ectopic expression of the defined transcription factors KLF4, OCT4, SOX2, and C-MYC, while the HESCs were obtained from the inner cell mass of a human preimplantation embryo. These various iPSCs resources such as virus-integrating iPSCs, vector-free iPSCs, and protein-directed reprogramming iPSCs show similar biological properties to human ESCs and fall into the same iPSCs category in our study. Routine analysis including normalization and statistical difference was performed using R package. Differentially expressed miRNAs were identified between pairwise cell types by linear model.

4.2. Identifying Genomic Clustered Region. PTMCluster [23] was first developed to find statistically significant PTM site clusters on the same protein using a positional distance tree. Here, in order to define clustered miRNA regions, we adjust the algorithm and search the differentially expressed miRNAs on chromosomes and a p value P is calculated to evaluate whether miRNAs in the cluster are close enough in space than being randomly distributed,

$$P = \left(1 - \left(1 - \frac{M}{L}\right)^{D}\right)^{N}. \quad (1)$$

Here, L is chromosome length. M is the number of miRNAs on the whole chromosome. D is the maximum distance between neighbour miRNAs in the cluster region. N is the number of differentially expressed miRNAs in the cluster region. Significant clustered region will be selected if it satisfies the p value and site number cutoffs as the tool recommended ($P \leq 0.1$; $N \geq 3$).

4.3. Permutation Test. To find the physical location between those miRNAs and imprinted genes, all imprinted genes were collected from Geneimprint (http://www.geneimprint.com/). We scan the total 210 human imprinted genes on chromosome and found that 30 fall in or nearby (at most 6 Mbp) our 12 miRNA clusters. A permutation test was then performed to verify whether the imprinted genes were enriched near those pluripotency related miRNAs other than appear nearby randomly. In this test, 1000 sets of miRNAs with the same quantity as the differentially expressed miRNAs were randomly selected from 800 miRNA in profile. Each set of miRNAs were then clustered by PTMCluster algorithm and the number of imprinted genes that fall in or nearby the miRNA clusters of each random set was, respectively, calculated by the same criteria. Only 37 permutated tests (out of 1000 tests) have more than 30 imprinted genes that fall in or nearby miRNA clusters, which indicated that the imprinted genes were enriched near differentially expressed miRNA clusters in our study ($P = 0.037$).

4.4. Molecular Network Analysis. Target genes for known miRNA were predicted through both miRanda (http://www.microrna.org/microrna/home.do) and TargetScan (http://www.targetscan.org/). Intersection part of the results was considered as the target genes for miRNAs. Ingenuity pathway analysis (IPA) system (Ingenuity Systems, http://www.ingenuity.com/) was used for molecular networks and disease module enrichment. It collects data from a variety of experimental platforms at multiple levels and provides a comprehensive pathway resource including information from literature, gene expression, and gene annotation. Fisher's exact test was implemented to determine whether a disease or function module is enriched with genes of interest. IPA helps us to provide insight into the molecular and chemical interactions, cellular phenotypes, and disease processes of our own interested molecules and construct the related pluripotent network.

4.5. Synteny Map Construction. We used CHSMiner [50] to construct synteny map for human and other mammalians. It detects CHS between two genomes based on shared gene content alone and gives each pair CHS a p value to reflect the probability that a given CHS is observed in two independently and randomly ordered genomes. The CHSs including the CPA4-KLF14 region between human and mouse, rat, cow, dog, and chicken were identified, respectively, and shown in Supplementary Figure 1.

Conflict of Interests

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

Authors' Contribution

Liyun Yuan and Xiaoyan Tang participated in the design of the study and performed the statistical analysis. Binyan Zhang helped to perform the synteny analysis and disease enrichment. Guohui Ding conceived of the study, participated in its design, and helped to draft the paper. All authors read and approved the paper.

Acknowledgments

This research was supported by grants from National HighTech R&D Program (863) (2009AA02Z304, 2012AA020404), State key basic research program (973) (2006CB910705, 2010CB529206, and 2011CBA00801), Research Program of CAS (KSCX2-YW-R-112, KSCX2-YW-R-190, and 2011KIP204), National Natural Science Foundation of China (30900272), and SA-SIBS Scholarship Program.

References

[1] J. Silva and A. Smith, "Capturing pluripotency," *Cell*, vol. 132, no. 4, pp. 532–536, 2008.

[2] J. Silva, O. Barrandon, J. Nichols, J. Kawaguchi, T. W. Theunissen, and A. Smith, "Promotion of reprogramming to ground state pluripotency by signal inhibition," *PLoS Biology*, vol. 6, no. 10, article e253, 2008.

[3] J. Yu, M. A. Vodyanik, K. Smuga-Otto et al., "Induced pluripotent stem cell lines derived from human somatic cells," *Science*, vol. 318, no. 5858, pp. 1917–1920, 2007.

[4] K. Takahashi and S. Yamanaka, "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors," *Cell*, vol. 126, no. 4, pp. 663–676, 2006.

[5] M. Wernig, A. Meissner, R. Foreman et al., "In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state," *Nature*, vol. 448, no. 7151, pp. 318–324, 2007.

[6] K. Okita, T. Ichisaka, and S. Yamanaka, "Generation of germline-competent induced pluripotent stem cells," *Nature*, vol. 448, no. 7151, pp. 313–317, 2007.

[7] M. Stadtfeld, E. Apostolou, H. Akutsu et al., "Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells," *Nature*, vol. 465, no. 7295, pp. 175–181, 2010.

[8] L. Liu, G.-Z. Luo, W. Yang et al., "Activation of the imprinted Dlk1-Dio3 region correlates with pluripotency levels of mouse stem cells," *The Journal of Biological Chemistry*, vol. 285, no. 25, pp. 19483–19490, 2010.

[9] S. Griffiths-Jones, "The microRNA registry," *Nucleic Acids Research*, vol. 32, pp. D109–D111, 2004.

[10] J. F. Palatnik, E. Allen, X. Wu et al., "Control of leaf morphogenesis by microRNAs," *Nature*, vol. 425, no. 6955, pp. 257–263, 2003.

[11] M. F. Mette, W. Aufsatz, T. Kanno et al., "Analysis of double-stranded RNA and small RNAs involved in RNA-mediated transcriptional gene silencing," *Methods in Molecular Biology*, vol. 309, pp. 61–82, 2005.

[12] A. C. Mallory and H. Vaucheret, "Functions of microRNAs and related small RNAs in plants," *Nature Genetics*, vol. 38, supplement 1, pp. S31–S36, 2006.

[13] P. Lüningschrör, S. Hauser, B. Kaltschmidt, and C. Kaltschmidt, "MicroRNAs in pluripotency, reprogramming and cell fate induction," *Biochimica et Biophysica Acta*, vol. 1833, no. 8, pp. 1894–1903, 2013.

[14] U. Lakshmipathy, J. Davila, and R. P. Hart, "MiRNA in pluripotent stem cells," *Regenerative Medicine*, vol. 5, no. 4, pp. 545–555, 2010.

[15] D. Haig and C. Graham, "Genomic imprinting and the strange case of the insulin-like growth factor II receptor," *Cell*, vol. 64, no. 6, pp. 1045–1046, 1991.

[16] W. Reik, M. Constancia, W. Dean et al., "Igf2 imprinting in development and disease," *The International Journal of Developmental Biology*, vol. 44, no. 1, pp. 145–150, 2000.

[17] A. A. Wylie, S. K. Murphy, T. C. Orton, and R. L. Jirtle, "Novel imprinted DLK1/GTL2 domain on human chromosome 14 contains motifs that mimic those implicated in IGF2/H19 regulation," *Genome Research*, vol. 10, no. 11, pp. 1711–1718, 2000.

[18] M. H. Chin, M. J. Mason, W. Xie et al., "Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures," *Cell Stem Cell*, vol. 5, no. 1, pp. 111–123, 2009.

[19] G. K. Smyth, "Linear models for microarray data," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health, pp. 397–420, Springer, New York, NY, USA, 2005.

[20] T. R. Leonardo, H. L. Schultheisz, J. F. Loring, and L. C. Laurent, "The functions of microRNAs in pluripotency and reprogramming," *Nature Cell Biology*, vol. 14, no. 11, pp. 1114–1121, 2012.

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

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

[23] H. Li, X. Xing, G. Ding et al., "SysPTM: a systematic resource for proteomic research on post-translational modifications," *Molecular & Cellular Proteomics*, vol. 8, no. 8, pp. 1839–1849, 2009.

[24] Y. Wang, S. Velho, E. Vakiani et al., "Mutant N-RAS protects colorectal cancer cells from stress-induced apoptosis and contributes to cancer development and progression," *Cancer Discovery*, vol. 3, no. 3, pp. 294–307, 2013.

[25] R. Plattner, S. Gupta, R. Khosravi-Far et al., "Differential contribution of the ERK and JNK mitogen-activated protein kinase cascades to ras transformation of HT1080 fibrosarcoma and DLD-1 colon carcinoma cells," *Oncogene*, vol. 18, no. 10, pp. 1807–1817, 1999.

[26] M. Eskandarpour, S. Kiaii, C. Zhu, J. Castro, A. J. Sakko, and J. Hansson, "Suppression of oncogenic NRAS by RNA interference induces apoptosis of human melanoma cells," *International Journal of Cancer*, vol. 115, no. 1, pp. 65–73, 2005.

[27] Y. Ma, Q. Li, W. Cui et al., "Expression of c-Jun, p73, Casp9, and *N*-ras in thymic epithelial tumors: relationship with the current WHO classification systems," *Diagnostic Pathology*, vol. 7, no. 1, article 120, 2012.

[28] X. Fang, S. Yu, A. Eder et al., "Regulation of BAD phosphorylation at serine 112 by the Ras-mitogen-activated protein kinase pathway," *Oncogene*, vol. 18, no. 48, pp. 6635–6640, 1999.

[29] E. S. Alnemri, "Mammalian cell death proteases: a family of highly conserved aspartate specific cysteine proteases," *Journal of Cellular Biochemistry*, vol. 64, no. 1, pp. 33–42, 1997.

[30] V. Depraetere and P. Golstein, "Dismantling in cell death: molecular mechanisms and relationship to caspase activation," *Scandinavian Journal of Immunology*, vol. 47, no. 6, pp. 523–531, 1998.

[31] A. Apelbaum, G. Yarden, S. Warszawski, D. Harari, and G. Schreiber, "Type I interferons induce apoptosis by balancing cFLIP and caspase-8 independent of death ligands," *Molecular and Cellular Biology*, vol. 33, no. 4, pp. 800–814, 2013.

[32] M. Kikuchi, S. Kuroki, M. Kayama, S. Sakaguchi, K.-K. Lee, and S. Yonehara, "Protease activity of procaspase-8 is essential for cell survival by inhibiting both apoptotic and nonapoptotic cell death dependent on receptor-interacting protein kinase 1 (RIP1) and RIP3," *The Journal of Biological Chemistry*, vol. 287, no. 49, pp. 41165–41173, 2012.

[33] L. Lefebvre, S. Viville, S. C. Barton, F. Ishino, E. B. Keverne, and M. Azim Surani, "Abnormal maternal behaviour and growth retardation associated with loss of the imprinted gene Mest," *Nature Genetics*, vol. 20, no. 2, pp. 163–169, 1998.

[34] K.-Y. Ryu, H. Park, D. J. Rossi, I. L. Weissman, and R. R. Kopito, "Perturbation of the hematopoietic system during embryonic liver development due to disruption of polyubiquitin gene *Ubc* in mice," *PLoS ONE*, vol. 7, no. 2, Article ID e32956, 2012.

[35] K.-Y. Ryu, R. Maehr, C. A. Gilchrist et al., "The mouse polyubiquitin gene UbC is essential for fetal liver development, cell-cycle progression and stress tolerance," *The EMBO Journal*, vol. 26, no. 11, pp. 2693–2706, 2007.

[36] K.-Y. Ryu, S. A. Sinnar, L. G. Reinholdt et al., "The mouse polyubiquitin gene Ubb is essential for meiotic progression," *Molecular and Cellular Biology*, vol. 28, no. 3, pp. 1136–1146, 2008.

[37] K.-Y. Ryu, J. C. Garza, X.-Y. Lu, G. S. Barsh, and R. R. Kopito, "Hypothalamic neurodegeneration and adult-onset obesity in mice lacking the Ubb polyubiquitin gene," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 10, pp. 4016–4021, 2008.

[38] C.-S. Yang, Z. Li, and T. M. Rana, "microRNAs modulate iPS cell generation," *RNA*, vol. 17, no. 8, pp. 1451–1460, 2011.

[39] R. Schmidt and K. Plath, "The roles of the reprogramming factors Oct4, Sox2 and Klf4 in resetting the somatic cell epigenome during induced pluripotent stem cell generation," *Genome Biology*, vol. 13, article 251, 2012.

[40] T. Lin, C. Chao, S. Saito et al., "p53 induces differentiation of mouse embryonic stem cells by suppressing Nanog expression," *Nature Cell Biology*, vol. 7, no. 2, pp. 165–171, 2005.

[41] H. Qin, T. Yu, T. Qing et al., "Regulation of apoptosis and differentiation by p53 in human embryonic stem cells," *The Journal of Biological Chemistry*, vol. 282, no. 8, pp. 5842–5852, 2007.

[42] T. Maimets, I. Neganova, L. Armstrong, and M. Lako, "Activation of p53 by nutlin leads to rapid differentiation of human embryonic stem cells," *Oncogene*, vol. 27, no. 40, pp. 5277–5287, 2008.

[43] K. Meletis, V. Wirta, S.-M. Hede, M. Nistér, J. Lundeberg, and J. Frisén, "p53 suppresses the self-renewal of adult neural stem cells," *Development*, vol. 133, no. 2, pp. 363–369, 2006.

[44] Y. Liu, S. E. Elf, Y. Miyata et al., "p53 regulates hematopoietic stem cell quiescence," *Cell Stem Cell*, vol. 4, no. 1, pp. 37–48, 2009.

[45] N. J. Haughey, D. Liu, A. Nath, A. C. Borchard, and M. P. Mattson, "Disruption of neurogenesis in the subventricular zone of adult mice, and in human cortical neuronal precursor cells in culture, by amyloid β-peptide: implications for the pathogenesis of Alzheimer's disease," *NeuroMolecular Medicine*, vol. 1, no. 2, pp. 125–135, 2002.

[46] H. Choi, H.-H. Park, K.-Y. Lee et al., "Coenzyme Q10 restores amyloid beta-inhibited proliferation of neural stem cells by activating the PI3K pathway," *Stem Cells and Development*, vol. 22, no. 15, pp. 2112–2120, 2013.

[47] H. B. Houbaviy, M. F. Murray, and P. A. Sharp, "Embryonic stem cell-specific microRNAs," *Developmental Cell*, vol. 5, no. 2, pp. 351–358, 2003.

[48] H. B. Houbaviy, L. Dennis, R. Jaenisch, and P. A. Sharp, "Characterization of a highly variable eutherian microRNA gene," *RNA*, vol. 11, no. 8, pp. 1245–1257, 2005.

[49] W. E. Lowry, L. Richter, R. Yachechko et al., "Generation of human induced pluripotent stem cells from dermal fibroblasts," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 8, pp. 2883–2888, 2008.

[50] Z. Wang, G. Ding, Z. Yu, L. Liu, and Y. Li, "CHSMiner: a GUI tool to identify chromosomal homologous segments," *Algorithms for Molecular Biology*, vol. 4, article 2, 2009.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 831790, 12 pages
http://dx.doi.org/10.1155/2015/831790

Research Article

A Model of Regularization Parameter Determination in Low-Dose X-Ray CT Reconstruction Based on Dictionary Learning

Cheng Zhang,[1,2,3] Tao Zhang,[2] Jian Zheng,[1] Ming Li,[1,2,3] Yanfei Lu,[1,2,3] Jiali You,[1,2,3] and Yihui Guan[4]

[1] *Medical Imaging Laboratory, Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou 215163, China*
[2] *Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China*
[3] *University of Chinese Academy of Sciences, Beijing 100049, China*
[4] *PET Center, Huashan Hospital, Fudan University, Shanghai 200235, China*

Correspondence should be addressed to Jian Zheng; zhengj@sibet.ac.cn

Received 13 March 2015; Accepted 11 June 2015

Academic Editor: Lin Lu

In recent years, X-ray computed tomography (CT) is becoming widely used to reveal patient's anatomical information. However, the side effect of radiation, relating to genetic or cancerous diseases, has caused great public concern. The problem is how to minimize radiation dose significantly while maintaining image quality. As a practical application of compressed sensing theory, one category of methods takes total variation (TV) minimization as the sparse constraint, which makes it possible and effective to get a reconstruction image of high quality in the undersampling situation. On the other hand, a preliminary attempt of low-dose CT reconstruction based on dictionary learning seems to be another effective choice. But some critical parameters, such as the regularization parameter, cannot be determined by detecting datasets. In this paper, we propose a reweighted objective function that contributes to a numerical calculation model of the regularization parameter. A number of experiments demonstrate that this strategy performs well with better reconstruction images and saving of a large amount of time.

1. Introduction

Nowadays, X-ray computed tomography (CT) is still an important part of biomedical imaging technologies for the reason that the reconstructed image is of high spatial resolution and quality. Nevertheless, it confirms that an overdose of radiation possibly increases the risk of genetic or cancerous diseases, making it urgent to develop creative and effective reconstruction techniques to fit low-dose CT scanning protocol. Obviously, the X-ray flux cannot be reduced much since the signal-to-noise ratio (SNR) of measured data declines with the reduction of dose. Another approach is to decrease the number of projection angles, which will lead to incomplete few-view data. In this case, analytic-based algorithms like FDK [1], which are derived from a continuous imaging model and in need of dense sampled projections, are sensitive to insufficient projection data and arrive at a terrible result. However, algebraic algorithms like the simultaneous algebraic reconstruction technique (SART) [2] solved the problem better by transforming it to a series of linear equations.

Recently, Candes et al. [3, 4] have made compressed sensing theory popular in information theory field. This theory indicates that a variety of signals can be represented sparsely in a certain transform domain. Therefore, original signal can be recovered accurately by far fewer samples while there is no need to follow the Shannon/Nyquist sampling theorem. A principle called restricted isometry property (RIP) guarantees the perfect recovery of any sparse signal [5]. This novel theory has been applied to many regions, like information technology [6], signal and image processing [7], inverse filtering [8], and so on. It is said that the data acquisition

process with compression is good for enhancing image quality because this method can increase imaging speed and suppress the artifacts caused by patients' movement [9]. For these benefits, many compressed sensing based algorithms are created to deal with few-view CT reconstruction problem. One major group is based on the total variation (TV), which takes the TV of the image as the sparse constraint. The image is determined by minimizing the TV term with the constraints of the linear projection equations. Sidky and Pan presented an improved TV-based algorithm named adaptive steepest descent projection onto convex sets (ASD-POCS) in circular cone-beam framework [10]. Another similar method called gradient projection Barzilari Borwein (GPBB) has a faster convergence speed [11]. Besides the TV minimization algorithms, dictionary learning is also helpful to sparse representation. During the reconstruction process, the image is divided into many overlapped patches, represented sparsely by overcomplete elements of a particular dictionary. Xu et al. combined statistical iterative reconstruction (SIR) with dictionary learning and got a better reconstruction result than TV-based methods in the low-dose CT condition [12]. According to Xu's paper, this method is robust to noise and obtains a better reconstructed image with more details than the TV-based methods do. Naturally, there are some parameters relevant to the final result. Some of them, like the sparse level, the scale of the dictionary, and so on, have less change due to different scanning data and then can be empirically selected. However, there is a special parameter changing according to the phantom, the scanning protocol, the noise level, and other factors. This parameter plays an important role in the reconstruction program to balance the data fidelity term and the regularization term while determining its value is time consuming with many attempts. Hence, there is no doubt that providing a model to select a proper value of this parameter according to the scanning data is essential for the algorithm based on dictionary learning, which leads to better result and time saving.

This paper is organized as follows. In Section 2, the problem of low-dose CT reconstruction is stated and the algorithm based on dictionary learning is reviewed. In Section 3, the model of regularization parameter determination is proposed by function fitting method. In Section 4, a series of experiments are performed and corresponding discussions are given. Finally, there is the conclusion at the end of this paper.

2. Notation and Problem Description

2.1. Background and Notation Interpretation. According to previous work by Xu et al. [12], SIR is united with dictionary learning to derive the algorithm. SIR assumes that the measured data can be regarded as the Poisson distribution

$$y_i \sim \text{Poisson}\left\{b_i e^{-l_i} + r_i\right\}, \quad i = 1, \ldots, I, \tag{1}$$

where $\mathbf{b} = (b_1, b_2, \ldots, b_I)^T \in \mathbf{R}^{I\times 1}$ is the entrance X-ray intensity, $\mathbf{y} = (y_1, y_2, \ldots, y_I)^T \in \mathbf{R}^{I\times 1}$ is the exit X-ray intensity, $\mathbf{l} = (l_1, l_2, \ldots, l_I)^T \in \mathbf{R}^{I\times 1}$ is the integral of the linear attenuation coefficient with $l_i = [\mathbf{A}\boldsymbol{\mu}]_i = \sum_{j=1}^{N^2} a_{ij}\mu_j$, $\mathbf{A} = \{a_{ij}\} \in \mathbf{R}^{I\times N^2}$ is the system matrix, the reconstructed image $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_{N^2})^T$ is a linear attenuation coefficient distribution, which transforms the initial image of $N\times N$ pixels to a vector $\boldsymbol{\mu} \in \mathbf{R}^{N^2\times 1}$, and r_i represents the read-out noise.

The objective function of SIR is as

$$\sum_{i=1}^{I} \frac{\omega_i}{2}\left([\mathbf{A}\boldsymbol{\mu}]_i - \widehat{l}_i\right)^2 + \lambda R(\boldsymbol{\mu}), \tag{2}$$

where $\varphi(\boldsymbol{\mu}) = \sum_{i=1}^{I}(\omega_i/2)([\mathbf{A}\boldsymbol{\mu}]_i - \widehat{l}_i)^2$ is the data fidelity term, $\widehat{\mathbf{l}} = (\widehat{l}_1, \widehat{l}_2, \ldots, \widehat{l}_I)^T \in \mathbf{R}^{I\times 1}$ is the measured data of $\mathbf{l}$ calculated by $\widehat{l}_i = \ln(b_i/(y_i - r_i))$, $\omega_i = (y_i - r_i)^2/y_i$ is the statistical weight, and $R(\boldsymbol{\mu})$ is the regularization term.

The regularization term usually contains prior information of the image, like sparse constraint. When the sparse representation is acquired by dictionary learning theory, we can replace $R(\boldsymbol{\mu}) = \sum_s \|\mathbf{E}_s\boldsymbol{\mu} - \mathbf{D}\boldsymbol{\alpha}_s\|_2^2 + \sum_s \nu_s\|\boldsymbol{\alpha}_s\|_0$ in the objective function. Therefore, the reconstruction problem is equivalent to the following minimization:

$$\begin{aligned}\min_{\boldsymbol{\mu},\boldsymbol{\alpha},\mathbf{D}} &\sum_{i=1}^{I} \frac{\omega_i}{2}\left([\mathbf{A}\boldsymbol{\mu}]_i - \widehat{l}_i\right)^2 \\ &+ \lambda\left(\sum_s \|\mathbf{E}_s\boldsymbol{\mu} - \mathbf{D}\boldsymbol{\alpha}_s\|_2^2 + \sum_s \nu_s\|\boldsymbol{\alpha}_s\|_0\right),\end{aligned} \tag{3}$$

where $\mathbf{E}_s = \{e_{nj}^s\} \in \mathbf{R}^{N_o^2\times N^2}$ is an operator to extract patches with $N_0 \times N_0$ pixels from the image, $\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_K) \in \mathbf{R}^{N_0^2\times K}$ is the training dictionary whose column $\mathbf{d}_k \in \mathbf{R}^{N_0^2\times 1}$ is called an atom of the same size of a patch, $\boldsymbol{\alpha}_s \in \mathbf{R}^{K\times 1}$ has few nonzero entries as a sparse representation of patches by the dictionary basis $\mathbf{D}$, and the variables λ and ν_s are regularization parameters. In this optimization problem, $\boldsymbol{\mu}, \boldsymbol{\alpha}$, and $\mathbf{D}$ are all unknown; hence, a practical plan of minimizing the object function is an alternating minimization scheme. The plan divides the primary problem into two recursive steps: update of the dictionary model and update of the image. The final result is acquired by operating the two steps alternately until reaching a stopping criterion.

2.2. Update of the Dictionary Model. During this procedure, the image $\boldsymbol{\mu}$ is supposed to be fixed, meaning that the data fidelity term is a constant. The optimization problem is simplified to the one as

$$\min_{(\mathbf{D}),\boldsymbol{\alpha}} \sum_s \left\|\mathbf{E}_s\boldsymbol{\mu}^t - \mathbf{D}\boldsymbol{\alpha}_s\right\|_2^2 + \sum_x \nu_s\|\boldsymbol{\alpha}_s\|_0, \tag{4}$$

where $\boldsymbol{\mu}^t$ is an intermediate image of the last updating step. In the adaptive dictionary based statistical iterative reconstruction (ADSIR), the dictionary is defined dynamically based on the unknown image while the dictionary in the global dictionary based statistical iterative reconstruction (GDSIR) is predefined beforehand [12]. Previous researches have proved that the K-SVD algorithm performs well at training the dictionary [13]. Once the dictionary is determined, the OMP algorithm is used to update the sparse coding [14] with a predetermined sparse level, instead of solving the l_0-norm problem as (4) directly.

2.3. Update of the Image. While updating the image, the dictionary and sparse coding remain invariable. In other words, the problem transforms to the form as

$$\min_{\boldsymbol{\mu}} \sum_{i=1}^{I} \frac{\omega_i}{2} \left([\mathbf{A}\boldsymbol{\mu}]_i - \widehat{l}_i \right)^2 + \lambda \sum_s \left\| \mathbf{E}_s \boldsymbol{\mu} - \mathbf{D}\boldsymbol{\alpha}_s \right\|_2^2, \tag{5}$$

where λ is the regularization parameter balancing the data fidelity term $\sum_{i=1}^{I} (\omega_i/2)([\mathbf{A}\boldsymbol{\mu}]_i - \widehat{l}_i)^2$ and the regularization term $\sum_s \|\mathbf{E}_s\boldsymbol{\mu} - \mathbf{D}\boldsymbol{\alpha}_s\|_2^2$. The regularization term is already a separable quadratic function. By replacing the data fidelity term with a separable paraboloid surrogate [15], the optimization can be iteratively solved by

$$\mu_j^{t+1} = \left[\mu_j^t - \frac{\sum_{i=1}^{I} \left(a_{ij}\omega_i \left([\mathbf{A}\boldsymbol{\mu}^t]_i - \widehat{l}_i \right) \right) + 2\lambda \sum_s \sum_{n=1}^{N_0^2} e_{nj}^s \left([\mathbf{E}_s\boldsymbol{\mu}^t]_n - [\mathbf{D}\boldsymbol{\alpha}_s]_n \right)}{\sum_{i=1}^{I} \left(a_{ij}\omega_i \sum_{k=1}^{N^2} a_{ik} \right) + 2\lambda \sum_s \sum_{n=1}^{N_0^2} e_{nj}^s \sum_{k=1}^{N^2} e_{nk}^s} \right]_+ \quad j = 1, 2, \ldots, N^2. \tag{6}$$

3. Materials and Methods

3.1. Effect of the Regularization Parameter. As mentioned above, the regularization parameter λ is of great importance during the update of image (5). We consider the optimizing problem of the form

$$\begin{aligned} \boldsymbol{\mu}(\lambda) &= \arg\min_{\boldsymbol{\mu}} \varphi(\boldsymbol{\mu}) + \lambda R(\boldsymbol{\mu}), \\ \varphi(\boldsymbol{\mu}) &= \sum_{i=1}^{I} \frac{\omega_i}{2} \left([\mathbf{A}\boldsymbol{\mu}]_i - l_i \right)^2, \\ R(\boldsymbol{\mu}) &= \sum_s \left\| \mathbf{E}_s\boldsymbol{\mu} - \mathbf{D}\boldsymbol{\alpha}_s \right\|_2^2. \end{aligned} \tag{7}$$

If there is $\boldsymbol{\mu}(\lambda_1) = \min_{\boldsymbol{\mu}} \varphi(\boldsymbol{\mu}) + \lambda_1 R(\boldsymbol{\mu})$, $\boldsymbol{\mu}(\lambda_2) = \min_{\boldsymbol{\mu}} \varphi(\boldsymbol{\mu}) + \lambda_2 R(\boldsymbol{\mu})$, $0 < \lambda_1 < \lambda_2$, then we can get $\varphi(\boldsymbol{\mu}(\lambda_1)) \le \varphi(\boldsymbol{\mu}(\lambda_2))$ and $R(\boldsymbol{\mu}(\lambda_1)) \ge R(\boldsymbol{\mu}(\lambda_2))$ by an easy derivation of the unequal relations:

$$\begin{aligned} &\varphi(\boldsymbol{\mu}(\lambda_1)) + \lambda_1 R(\boldsymbol{\mu}(\lambda_1)) \\ &\quad \le \varphi(\boldsymbol{\mu}(\lambda_2)) + \lambda_1 R(\boldsymbol{\mu}(\lambda_2)), \\ &\varphi(\boldsymbol{\mu}(\lambda_2)) + \lambda_2 R(\boldsymbol{\mu}(\lambda_2)) \\ &\quad \le \varphi(\boldsymbol{\mu}(\lambda_1)) + \lambda_2 R(\boldsymbol{\mu}(\lambda_1)). \end{aligned} \tag{8}$$

It shows that a smaller λ makes the data fidelity term smaller and the regularization term bigger, which means that the sparse constraint has less effect on the optimizing process and more noise will appear in the final image. On the other hand, a bigger λ weakens the effect of the data fidelity term, generating a loss of some fine details in the image. For example, λ should be increased to suppress the noise increment in the projection domain since the data fidelity term is proportional to the noise standard deviation. In order to get an optimal result, previous work selects a great many values of λ and picks out the best one by comparing the final images. This testing strategy is of great time consuming, making the algorithm based on dictionary learning not friendly to the reconstruction task.

3.2. Morozov's Principle and the Balancing Principle. The research on the choices of regularization parameters in linear inverse problems appears early in 1998 [16]. The original optimizing function of the inverse problem is like

$$J(f, \beta) = \frac{1}{2} \|Tf - z\|_Y^2 + \beta \|f\|_X^2. \tag{9}$$

Take $Y = 2$, $X = 2$ as an example. If the noise level of z is known, one efficient tool for selecting the proper regularization parameter β is the well-known Morozov discrepancy principle [16]. To adaptively determine the regularization parameter, a model function is brought in [17]:

$$m(\beta) = b + \frac{s}{t + \beta}. \tag{10}$$

To find a solution of β, (9) is rewritten as

$$\begin{aligned} F(\beta) &= \frac{1}{2} \left\| Tf_\beta - z \right\|_2^2 + \beta \left\| f_\beta \right\|_2^2 = \varphi(\beta) + \beta F'(\beta) \\ &\text{with } \varphi(\beta) = \frac{1}{2} \left\| Tf_\beta - z \right\|_2^2, \; F'(\beta) = \left\| f_\beta \right\|_2^2. \end{aligned} \tag{11}$$

When $\beta \to \infty$, it obviously shows that the solution of the minimization problem is $f_\beta = \mathbf{0}$, and then $b = (1/2)\|z\|_2^2$ is obtained easily. The other two variables s and t can be determined by the equations $m(\beta^k) = F(\beta^k)$, $m'(\beta^k) = F'(\beta^k)$. To solve the parameter β iteratively, the balancing principle [18] is introduced as

$$(\sigma - 1)\varphi(\beta^*) = \beta^* F'(\beta^*), \tag{12}$$

where $\sigma > 1$ controls the relative weight of the two terms. Equation (12) can be written as

$$F(\beta^*) = \sigma \left(F(\beta^*) - \beta^* F'(\beta^*) \right), \tag{13}$$

which is a fixed point iteration. β^{k+1} is calculated by the formula

$$\begin{aligned} F\left(\beta^{k+1}\right) &= \sigma \left(F\left(\beta^k\right) - \beta^k F'\left(\beta^k\right) \right) \\ &= \sigma \left(m\left(\beta^k\right) - \beta^k m'\left(\beta^k\right) \right). \end{aligned} \tag{14}$$

Although the balancing principle behaves well in the inverse problem model, there is no direct way introducing the method to the dictionary based algorithm. The regularization

term has a minimum value greater than zero, which leads to the derivation as follows:

$$\begin{aligned} &\text{for} \quad R(\boldsymbol{\mu}) = \sum_s \left\| \mathbf{E}_s \boldsymbol{\mu} - \mathbf{D}\boldsymbol{\alpha}_s \right\|_2^2 \geq \delta > 0 \\ &\text{assume} \quad F(\lambda) \\ &\quad = \sum_{i=1}^{I} \frac{\omega_i}{2} \left([\mathbf{A}\boldsymbol{\mu}_\lambda]_i - l_i \right)^2 \\ &\qquad + \lambda \sum_s \left\| \mathbf{E}_s \boldsymbol{\mu}_\lambda - \mathbf{D}\boldsymbol{\alpha}_s \right\|_2^2 = \varphi(\lambda) + \lambda F'(\lambda) \\ &\quad F(\lambda) = m(\lambda) = b + \frac{s}{t+\lambda} \\ &\text{when} \quad \lambda \longrightarrow \infty, \; b \geq \lambda\delta \longrightarrow \infty. \end{aligned} \tag{15}$$

Therefore, the strategy that determining the regularization parameter adaptively accords to the last iterative result is not reasonable. We should look for a selecting strategy which can determine the proper value of the regularization parameter by making an analysis of the projection data.

3.3. Weight Modification of the Objective Function. In order to find out an applicable selecting model of the regularization parameter, we reconsider the minimizing problem (5) and the updating formula (6). According to the former work, the data fidelity term is replaced with a separable surrogate [15]

$$\begin{aligned} &Q\left(\boldsymbol{\mu}; \boldsymbol{\mu}^t\right) \\ &\quad = \sum_{i=1}^{I} \sum_{j=1}^{N^2} \alpha_{ij} \frac{\omega_i}{2} \times \left(\frac{a_{ij}}{\alpha_{ij}} \left(\mu_j - \mu_j^t \right) + \left[\mathbf{A}\boldsymbol{\mu}^t \right]_i - \widehat{l}_i \right)^2, \end{aligned} \tag{16}$$

where

$$\sum_{j=1}^{N^2} \alpha_{ij} = 1 \quad \forall i, \; \alpha_{ij} \geq 0. \tag{17}$$

And one convenient choice is

$$\alpha_{ij} = \frac{a_{ij}}{\sum_{j=1}^{N^2} a_{ij}}. \tag{18}$$

By making use of the surrogate, (5) becomes a separable form

$$\boldsymbol{\mu} = \arg\min_{\boldsymbol{\mu}} Q\left(\boldsymbol{\mu}; \boldsymbol{\mu}^t\right) + \lambda R(\boldsymbol{\mu}), \tag{19}$$

where $Q(\boldsymbol{\mu}; \boldsymbol{\mu}^t)$ and $R(\boldsymbol{\mu})$ both can be represented as the sum of a series of quadratic functions, whose variables are μ_j $(j = 1, 2, \ldots, N^2)$. After some variable exchanges, (19) can be rewritten as

$$\begin{aligned} &Q\left(\boldsymbol{\mu}; \boldsymbol{\mu}^t\right) = \frac{1}{2} \sum_{j=1}^{N^2} p_j \left(\mu_j - c_j^t \right)^2 + C_1, \\ &R(\boldsymbol{\mu}) = \sum_{j=1}^{N^2} q_j \left(\mu_j - d_j^t \right)^2 + C_2 \\ &\text{with } p_j = \sum_{i=1}^{I} \left(a_{ij} \omega_i \sum_{k=1}^{N^2} a_{ik} \right), \; q_j = \sum_s \sum_{n=1}^{N_0^2} e_{nj}^s \sum_{k=1}^{N^2} e_{nk}^s, \; c_j^t = \mu_j^t - \frac{\sum_{i=1}^{I} \left(a_{ij} \omega_i \left([\mathbf{A}\boldsymbol{\mu}^t]_i - \widehat{l}_i \right) \right)}{p_j}, \\ &d_j^t = \mu_j^t - \frac{\sum_s \sum_{n=1}^{N_0^2} e_{nj}^s \left([\mathbf{E}_s \boldsymbol{\mu}^t]_n - [\mathbf{D}\boldsymbol{\alpha}_s]_n \right)}{q_j}, \; \mu_j^{t+1} = \left[\frac{p_j c_j^t + 2\lambda q_j d_j^t}{p_j + 2\lambda q_j} \right]_+. \end{aligned} \tag{20}$$

From the above, the image updating formula is just the same as (6); the quadratic term coefficient p_j only depends on the system matrix $\mathbf{A}$ and the statistical weight $\boldsymbol{\omega}$. We make a weight modification on the regularization term

$$\begin{aligned} &\widetilde{R}(\boldsymbol{\mu}) = \sum_{j=1}^{n} r_j q_j \left(\mu_j - d_j^t \right)^2 + C_2 = \frac{1}{2} \sum_{j=1}^{n} p_j \left(\mu_j - d_j^t \right)^2 \\ &\quad + C_2 \\ &\text{s.t.} \quad r_j = \frac{1}{2} \frac{p_j}{q_j}. \end{aligned} \tag{21}$$

By eliminating the constant term, the image reconstruction process is equivalent to solving the following optimization problem:

$$\begin{aligned} \boldsymbol{\mu} &= \arg\min_{\boldsymbol{\mu}} Q\left(\boldsymbol{\mu}; \boldsymbol{\mu}^t\right) + \lambda \widetilde{R}(\boldsymbol{\mu}) \\ &= \arg\min_{\boldsymbol{\mu}} \frac{1}{2} \sum_{j=1}^{N^2} p_j \left(\mu_j - c_j^t \right)^2 + \frac{1}{2} \lambda \sum_{j=1}^{n} p_j \left(\mu_j - d_j^t \right)^2, \end{aligned} \tag{22}$$

which can be solved iteratively by

$$\mu_j^{t+1} = \left[\frac{c_j^t + \lambda d_j^t}{1 + \lambda} \right]_+. \tag{23}$$

As shown in (23), λ determines the relative impact on the updating image $\boldsymbol{\mu}$ by the data fidelity term and the regularization term, respectively.

3.4. Evaluation Model of Regularization Parameter. Before developing the evaluation model, some discussions about the reconstruction result are displayed firstly. Once a value of λ is selected randomly, by solving (22) iteratively, it infers that the relative error of the data fidelity term is as

$$\delta = \frac{\sum_{i=1}^{I} (\omega_i/2) \left([\mathbf{A}\boldsymbol{\mu}]_i - l_i \right)^2}{\sum_{i=1}^{I} (\omega_i/2) \, l_i^2}. \tag{24}$$

The relative error depends on the phantom image, the noise level, the regularization parameter, and so on. When it comes

Determination of λ
Initialize $\boldsymbol{\mu}^0, \mathbf{D}^0, \boldsymbol{\alpha}_s^0$, and $t = 0$.
While the stopping criterion is not satisfied, do
(1) Implement the OSC algorithm for acceleration;
(2) Extract patches from the intermediate image $\boldsymbol{\mu}^t$;
(3) Update the dictionary $\mathbf{D}^{t+1}$ by K-SVD algorithm;
(4) Update the sparse coding $\boldsymbol{\alpha}_s^{t+1}$ by OMP algorithm;
(5) Update the image $\boldsymbol{\mu}^{t+1}$ by (25) with $\lambda \to \infty, t = t + 1$;
Output the final image $\boldsymbol{\mu}^* = \boldsymbol{\mu}^t$ and relative error $\delta_{\lambda\to\infty}$ calculated by (26).
When the iteration is stopped, determine λ by (27).
Image Reconstruction
Initialize $\boldsymbol{\mu}^0, \mathbf{D}^0, \boldsymbol{\alpha}_s^0$, and $t = 0$, λ is determined in former step.
While the stopping criterion is not satisfied, do
(1) Implement the OSC algorithm for acceleration;
(2) Extract patches from the intermediate image $\boldsymbol{\mu}^t$;
(3) Update the dictionary $\mathbf{D}^{t+1}$ by K-SVD algorithm;
(4) Update the sparse coding $\boldsymbol{\alpha}_s^{t+1}$ by OMP algorithm;
(5) Update the image $\boldsymbol{\mu}^{t+1}$ by (23), $t = t + 1$;
Output the final reconstruction.

Algorithm 1: Workflow of the developed algorithm.

to the situation that the regularization parameter is infinite as $\lambda \to \infty$, (23) and (24) will be like the following form:

$$\mu_j^{t+1} = \left[d_j^t\right]_+, \tag{25}$$

$$\delta_{\lambda\to\infty} = \frac{\sum_{i=1}^{I}(\omega_i/2)\left(\left[\mathbf{A}\boldsymbol{\mu}^*_{\lambda\to\infty}\right]_i - l_i\right)^2}{\sum_{i=1}^{I}(\omega_i/2)\, l_i^2} \quad \text{s.t. } \lambda \longrightarrow \infty. \tag{26}$$

It is naturally derived that the relative error $\delta_{\lambda\to\infty}$ increases with the increment of the noise level in projection domain. In addition, it has been mentioned above (in Section 3.1) that λ should be increased with the noise increment. Therefore, the proper λ has a monotonous relation with the parameter $\delta_{\lambda\to\infty}$. Since $\delta_{\lambda\to\infty}$ can be easily determined by operating the reconstruction algorithm based on dictionary learning once with $\lambda \to \infty$, the proper value of λ can be calculated if a reasonable function as $\lambda^* = f(\delta_{\lambda\to\infty})$ can be found.

With the help of a series of tests, the relationship between λ^* and $\delta_{\lambda\to\infty}$ is fitted by a piecewise quadratic function as follows:

$$\lambda^* = 1.74485\delta_G^2 + 0.58883\delta_G - 6.88253 \quad \text{if } \delta_G > 1.96$$
$$\lambda^* = -0.21545\delta_G^2 + 1.08602\delta_G - 0.32634 \quad \text{if } \delta_G \le 1.96 \tag{27}$$
$$\text{with } \delta_G = 10^6\delta_{\lambda\to\infty}.$$

Finally, taking ADSIR as an example, the workflow of the developed algorithm is exhibited in Algorithm 1. In addition, the ordered subsets convex (OSC) algorithm [19] is utilized as an acceleration of the convergence.

4. Experimental Results and Discussion

To make the evaluation of the regularization parameter possible, the developed algorithm improves ADSIR with a modified weight of the regularization term while the weight is adaptive to the data fidelity term. So the proposed algorithm is named adaptive weight regularized ADSIR (AWR-ADSIR). In this section, a series of reconstruction experiments are exhibited to validate that the regularization selecting principle in AWR-ADSIR is practical. The simulation numerical phantoms are Shepp-Logan phantom, human head slice image, and human abdomen slice image. The Shepp-Logan phantom is a numeric phantom with pixels intensities ranging from 0 to 1. The sample images of human head slice and human abdomen slice are the FBP reconstruction results based on full-sampling scanning data, which are obtained from our collaborator. All of these phantom images are of 256×256 pixels presented in Figure 1. The scanning data are simulated as an undersampling situation with different noise levels. Firstly, different regularization parameters are selected to demonstrate that the one chosen by the algorithm leads to the best reconstruction result. Secondly, by comparing the quality of images reconstructed by diverse algorithms, which are SART, GPBB, ADSIR, and AWR-ADSIR, it can be confirmed that AWR-ADSIR is of remarkable performance among these algorithms. Finally, the selecting principle is used in the GDSIR model, proving that it also works. All the algorithms above are coded in MATLAB and run on a dual-core PC with 3.10 GHz Intel Core i5-2400 and 4 GB RAM.

4.1. Comparison of Different Regularization Parameters. In the following experiments, all the parameters except the regularization parameter λ keep invariant for the same phantom with the same projection noise level. Three values of λ are selected, of which one is calculated by the proposed

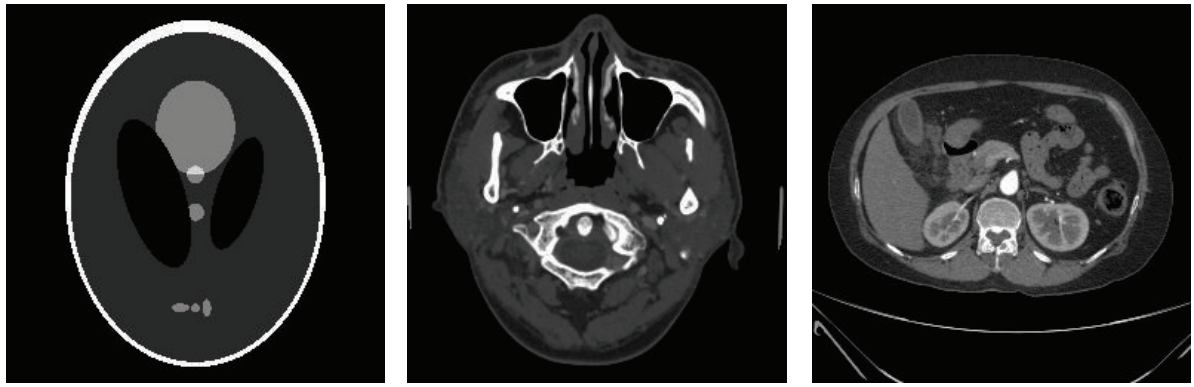


Figure 1: From left to right are Shepp-Logan phantom, human head slice image, and human abdomen slice image. The whole windows are [0, 1], [−1000, 1436] HU, and [−1000, 837] HU while the display windows are [0.15, 0.45], [400, 1000] HU, and [−160, 400] HU, respectively.

algorithm, another one is multiplied by 0.1, and the third one is multiplied by 10. The distance from the X-ray source to the center point of the phantom is twice the length of the image edge. The iteration of the algorithm is stopped when the relative error $\mathrm{err}^{\delta} = |\delta^{t} - \delta^{t-1}|/\delta^{t}$ is less than a stopping value (δ is calculated by (24)).

To compare the difference between different selections of the regularization parameter, the human abdomen slice image is tested as an example. The projection data are simulated by 180 views of 2° step length over a 360° range, and 512 detector elements are distributed in fan-beam geometry covering the phantom. The noise levels added to the projection data are 0.0% and 0.1% Gaussian random noise. The results are displayed in Figures 2 and 3. For the reason that biomedical images are often observed by a proper window to find more details, the images are displayed with a window [−160, 400] HU. The difference between the reconstructed image and the phantom image is displayed by a window [−90, 90] HU.

There are two criterions to evaluate the reconstructed image. One is the normalized mean absolute deviation (NMAD), defined as

$$\mathrm{NMAD}\,(\%) = \frac{\sum_{i,j} \left|\mu_{ij} - \mu_{ij}^{\mathrm{truth}}\right|}{\sum_{i,j} \left|\mu_{ij}^{\mathrm{truth}}\right|} \times 100. \tag{28}$$

The other one is the signal-to-noise ratio (SNR), defined as

$$\mathrm{SNR} = 10 \lg \left(\frac{\sum_{i,j} \left(\mu_{ij}^{\mathrm{truth}}\right)^2}{\sum_{i,j} \left(\mu_{ij} - \mu_{ij}^{\mathrm{truth}}\right)^2} \right). \tag{29}$$

The values of the two criterions are presented in Table 1. Comparing the results with the same noise level of the λ situation and the 10λ situation, the NMAD of the λ situation is smaller and the SNR of the λ situation is larger mostly, which proves that the image reconstructed by choosing the regularization parameter as λ is more close to the sample image. When it comes to the 0.1λ situation, although the values of the two criterions are a little better, there are some artifacts appearing in the reconstructed image, leading to a decline of the image quality. In the middle column of Figures 2 and 3, some horizontal line artifacts appear in the images (regions D, E, and F in Figure 2 and the ellipse regions in Figure 3). It seems that there are more horizontal artifacts in the 0.0% noise level image. The probable reason is the noise added to the projection data since the noise covers the inconspicuous artifact in the 0.1% noise level image. So what is the reason for the fact that the NMAD and SNR of the 0.1λ situation are better? One reasonable explanation might be the smoothing effect of the dictionary learning algorithm. When the iterative image is updated by (6) or (23), the sparse constraint added by dictionary learning method smoothes not only the noise but also the margin details. This effect becomes more significant when the regularization parameter is becoming larger. Therefore, the difference between the reconstructed image and the sample image in the margin area becomes more obvious, which is displayed in the left and right columns of Figure 3. By discovering this effect making the criterion worse, future work should be devoted to improving the reconstruction algorithm based on dictionary learning in order to smooth the noise and preserve the margin details meanwhile.

Table 1: Quantitative evaluation of the results with different regularization parameters.

Noise level	δ_G ($10^6\delta_{\lambda\to\infty}$)	λ		NMAD (%)	SNR (dB)
0.0%	1.7094	0.9005	λ	1.5819	35.5246
			0.1λ	1.2406	37.6730
			10λ	2.0147	32.9517
0.1%	2.5269	5.7462	λ	2.0371	33.1220
			0.1λ	1.8742	34.5110
			10λ	2.1561	32.4527

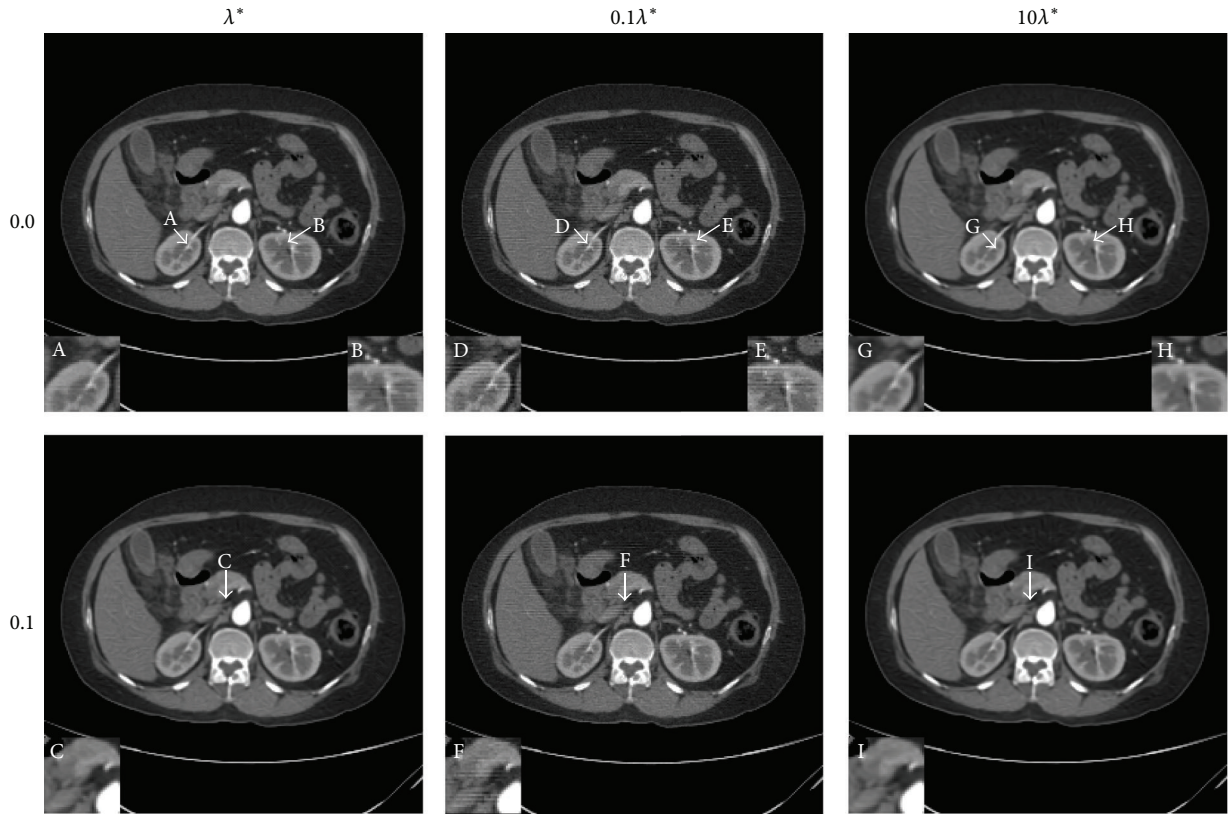


Figure 2: The results of human abdomen slice simulation study. From top to bottom, the noise levels are 0.0% and 0.1% in turn. From left to right, the regularization parameters are λ^*, $0.1\lambda^*$, and $10\lambda^*$. The display window is $[-160, 400]$ HU.

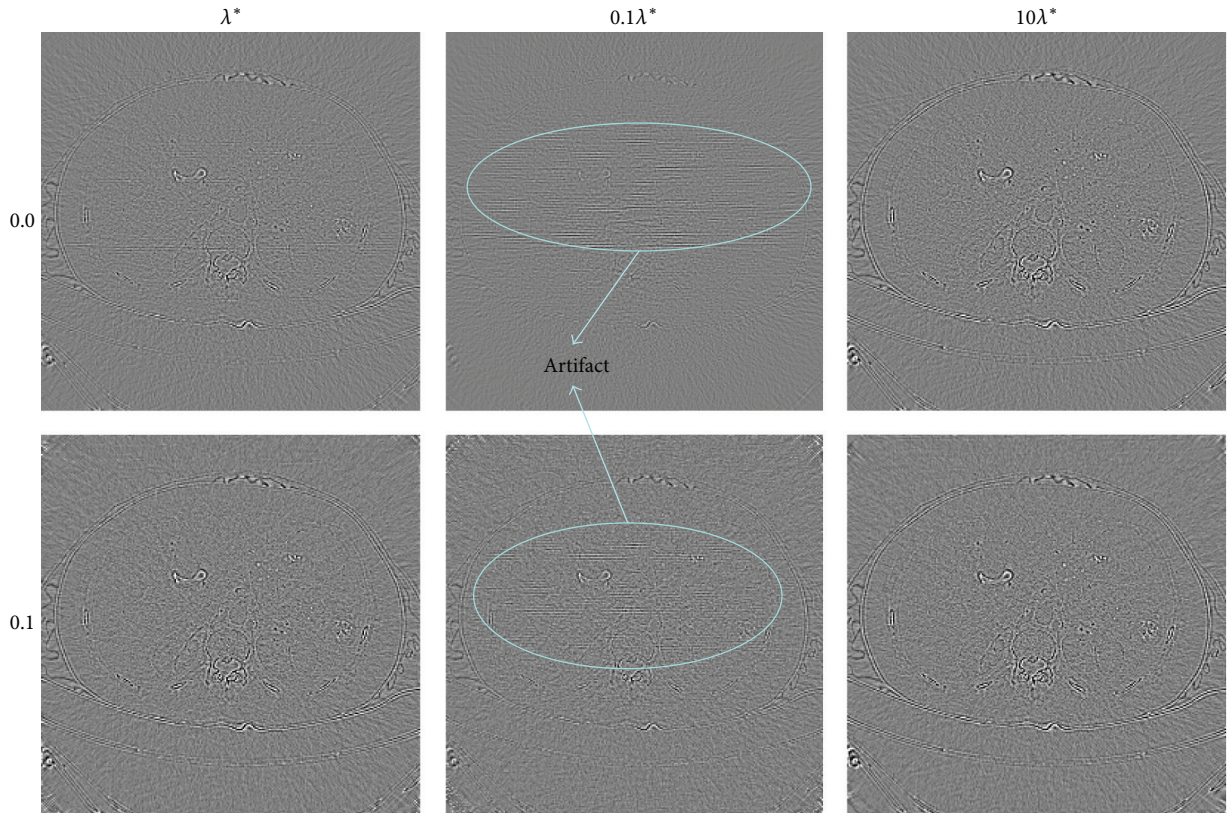


Figure 3: The difference between the reconstructed image and the original image (OI) of the human abdomen slice image. From top to bottom, the noise levels are 0.0% and 0.1% in turn. From left to right, the regularization parameters are λ^*, $0.1\lambda^*$, and $10\lambda^*$. The display window is $[-90, 90]$ HU.

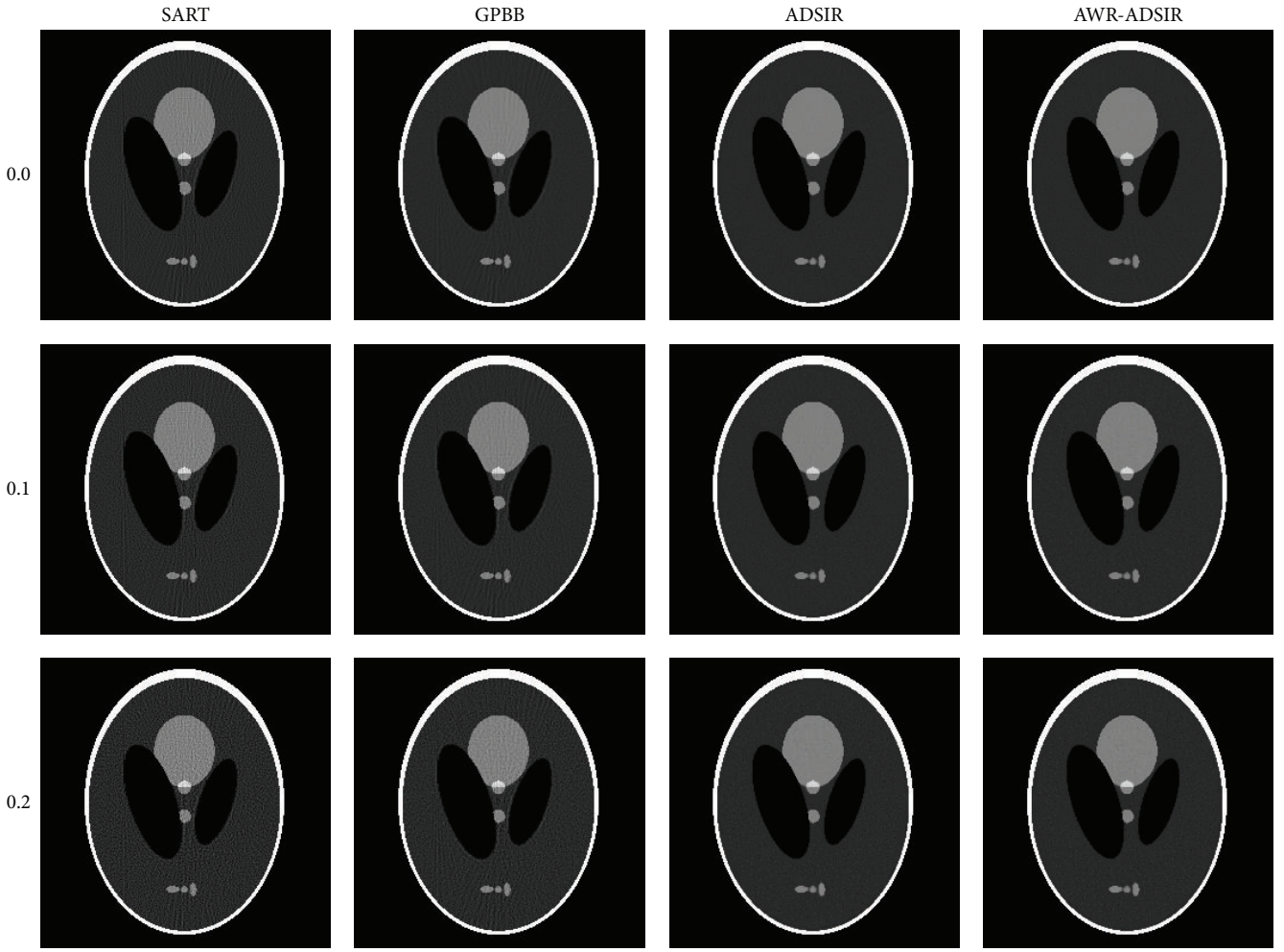


Figure 4: Reconstructed images from low-dose projection data of the Shepp-Logan phantom. From top to bottom, the noise levels are 0.0%, 0.1%, and 0.2% in turn. From left to right, the reconstruction algorithms are SART, GPBB, ADSIR, and AWR-ADSIR, respectively. The display window is [0.15, 0.45].

4.2. Comparison of Different Reconstruction Algorithms. To present the advantage of AWR-ADSIR, the images reconstructed by AWR-ADSIR and some other algorithms (SART, GPBB, and ADSIR) are compared with the same projection data, initial conditions, and stopping criterions. The testing examples are the Shepp-Logan phantom and the human head slice image. The Shepp-Logan phantom is simulated by 120 views of 3° step length over a 360° range, and 512 detector elements are distributed in fan-beam geometry with three different noise levels while the head slice sample is simulated by 180 views of 2° step length over a 360° range. The reconstructed results are displayed in the four figures (Figure 4 to Figure 7).

With the comparative results calculated by (28) and (29) presented in Tables 2 and 3, the quality of all the reconstructed images is decreased with the noise level increasing. Among these four algorithms, SART generates the worst results. GPBB behaves well when the noise level is very low but when the noise level is beyond 0.1%, the quality of reconstructed image degenerates quickly. The regularization parameter in

Table 2: Comparing criterions of the results reconstructed by different algorithms (Shepp-Logan).

Algorithm	Criterion	Noise level		
		0.0%	0.1%	0.2%
SART	NMAD (%)	2.1510	2.9273	4.1032
	SNR (dB)	33.5448	31.5220	28.7997
GPBB	NMAD (%)	1.3472	1.7826	3.0473
	SNR (dB)	36.5443	34.6140	30.5004
ADSIR	NMAD (%)	0.8110	0.9553	1.1823
	SNR (dB)	35.7740	34.8516	33.3413
AWR-ADSIR	NMAD (%)	0.8223	1.0639	1.1549
	SNR (dB)	35.5354	34.5480	33.7801

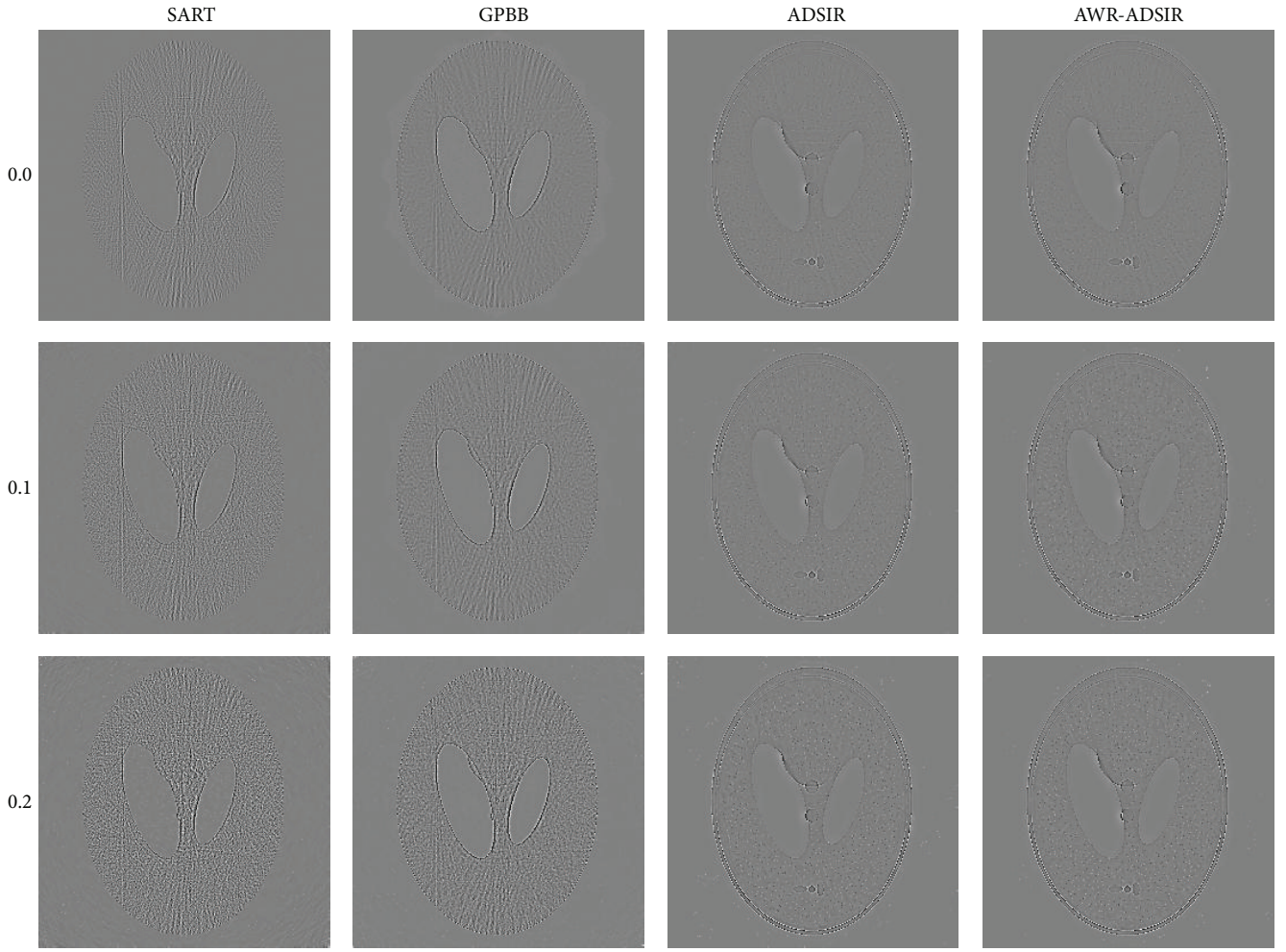


Figure 5: The difference between the reconstructed image and the original image (OI) of the Shepp-Logan phantom. From top to bottom, the noise levels are 0.0%, 0.1%, and 0.2% in turn. From left to right, the reconstruction algorithms are SART, GPBB, ADSIR, and AWR-ADSIR, respectively. The display window is $[-0.05, 0.05]$.

Table 3: Comparing criterions of the results reconstructed by different algorithms (head).

Algorithm	Criterion	Noise level		
		0.0%	0.1%	0.2%
SART	NMAD (%)	1.0904	2.2017	4.0009
	SNR (dB)	36.8554	31.6575	26.4913
GPBB	NMAD (%)	0.4837	1.1631	2.7798
	SNR (dB)	45.0405	36.7565	29.4104
ADSIR	NMAD (%)	0.5981	0.9370	1.1693
	SNR (dB)	39.0515	34.9548	32.8972
AWR-ADSIR	NMAD (%)	0.6765	0.9438	1.1572
	SNR (dB)	37.7225	34.9687	32.9703

ADSIR is empirically selected according to [12]. The fact that the image quality and comparing criterions of ADSIR and AWR-ADSIR are mostly the same proves that the parameter selecting model in AWR-ADSIR is practical and efficient. In addition, the marginal details appearing in Figures 5 and 7 of algorithms ADSIR and AWR-ADSIR indicate the smoothing effect on the margins again. Since the high-contrast edge is smoothed by the algorithm, the structural boundaries appear in the difference image in Figure 7. Obviously, when the empirical regularization parameter is unknown, the ADSIR algorithm costs a large amount of time to determine the proper value by repeatedly operating the iterative process (usually more than ten times) while the AWR-ADSIR algorithm only operates the iterative process twice as Algorithm 1 shows. The model indeed reduces the time consumption to determine the value of the regularization parameter.

4.3. Adaptive Weight Regularized GDSIR. The last experiment is applying the parameter selecting strategy into the GDSIR model. The dictionary displayed in Figure 8 is learned from the overlapping patches of the original image in Figure 1.

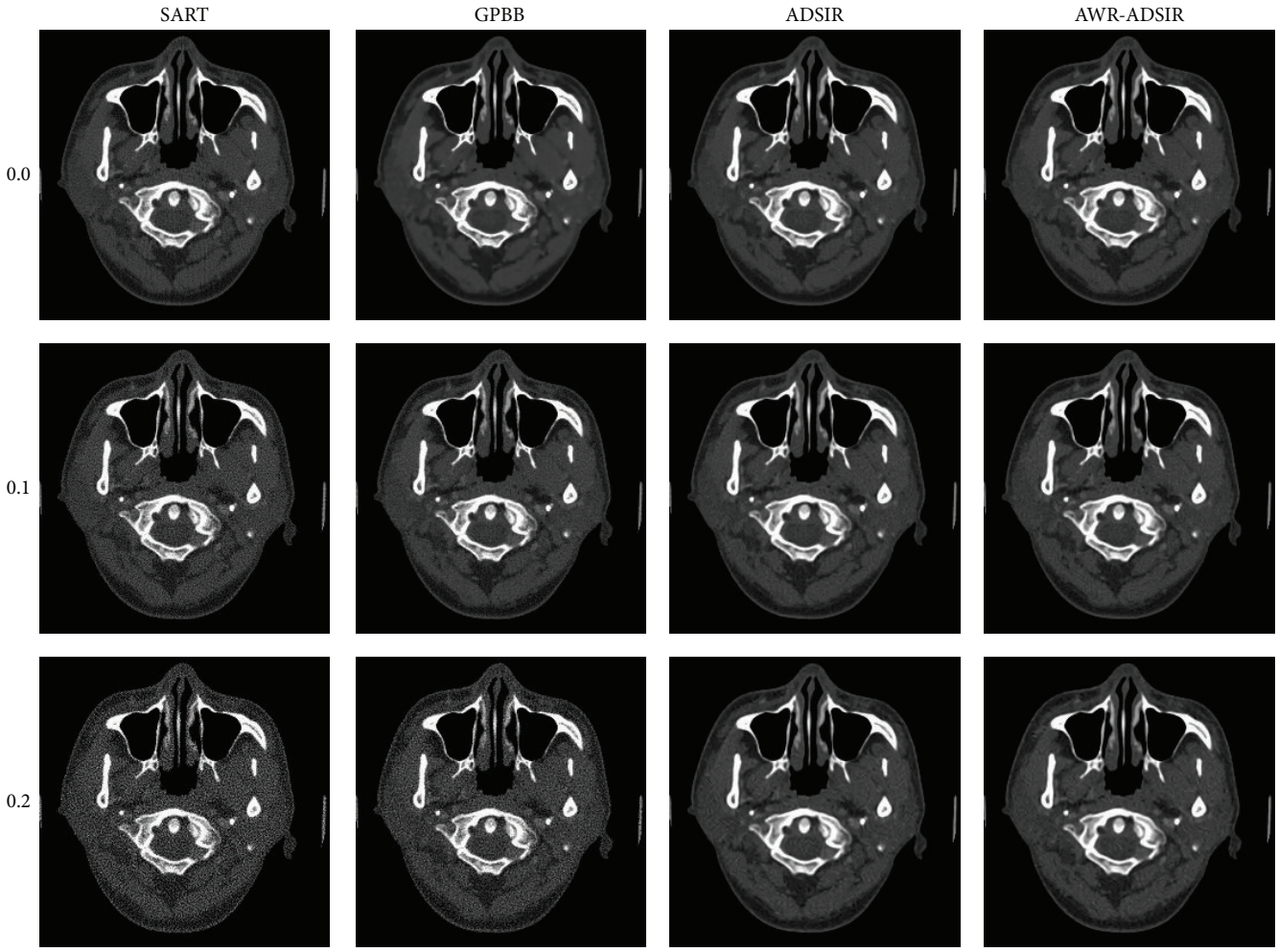


Figure 6: The results of human head slice simulation study. From top to bottom, the noise levels are 0.0%, 0.1%, and 0.2% in turn. From left to right, the reconstruction algorithms are SART, GPBB, ADSIR, and AWR-ADSIR, respectively. The display window is [400, 1000] HU.

The reconstruction process of AWR-GDSIR is almost the same as AWR-ADSIR except that the dictionary has been constructed in advance and dose not change during the reconstruction process. The result shown in Figure 8 indicates that the proposed model is also suitable for the general dictionary condition.

5. Conclusion

In most optimization problems, the determination of the regularization parameter is still a problem. In this paper, aiming to determine the regularization parameter of the algorithm based on dictionary learning, one model function, whose independent variable $\delta_{\lambda\to\infty}$ can be calculated by the known projection data, is proposed depending on some modification on the objective function. When compared to some other algorithms, the images reconstructed by AWR-ADSIR and ADSIR are of similar quality, better than the one reconstructed by SART, and the proposed algorithm is much more robust to noise than GPBB. This indicates that the modification of the objective function does not degrade the performance of ADSIR. What is more, the parameter selection model is demonstrated to be rational by the fact that the image quality of the λ situation is better than the ones of 0.1λ and 10λ situations. However, when some other parameters (like the scale of the dictionary, the scale of the patch, and so on) change, the model function might result in some difference. However, it still works to look for the function relationship between these two parameters since the monotonous relation between $\delta_{\lambda\to\infty}$ and λ remains.

By validating the proposed selecting principle, the smoothing effect on image margins is discovered. Our future work will focus on improving the dictionary learning method, with the expectation to maintain the smoothing

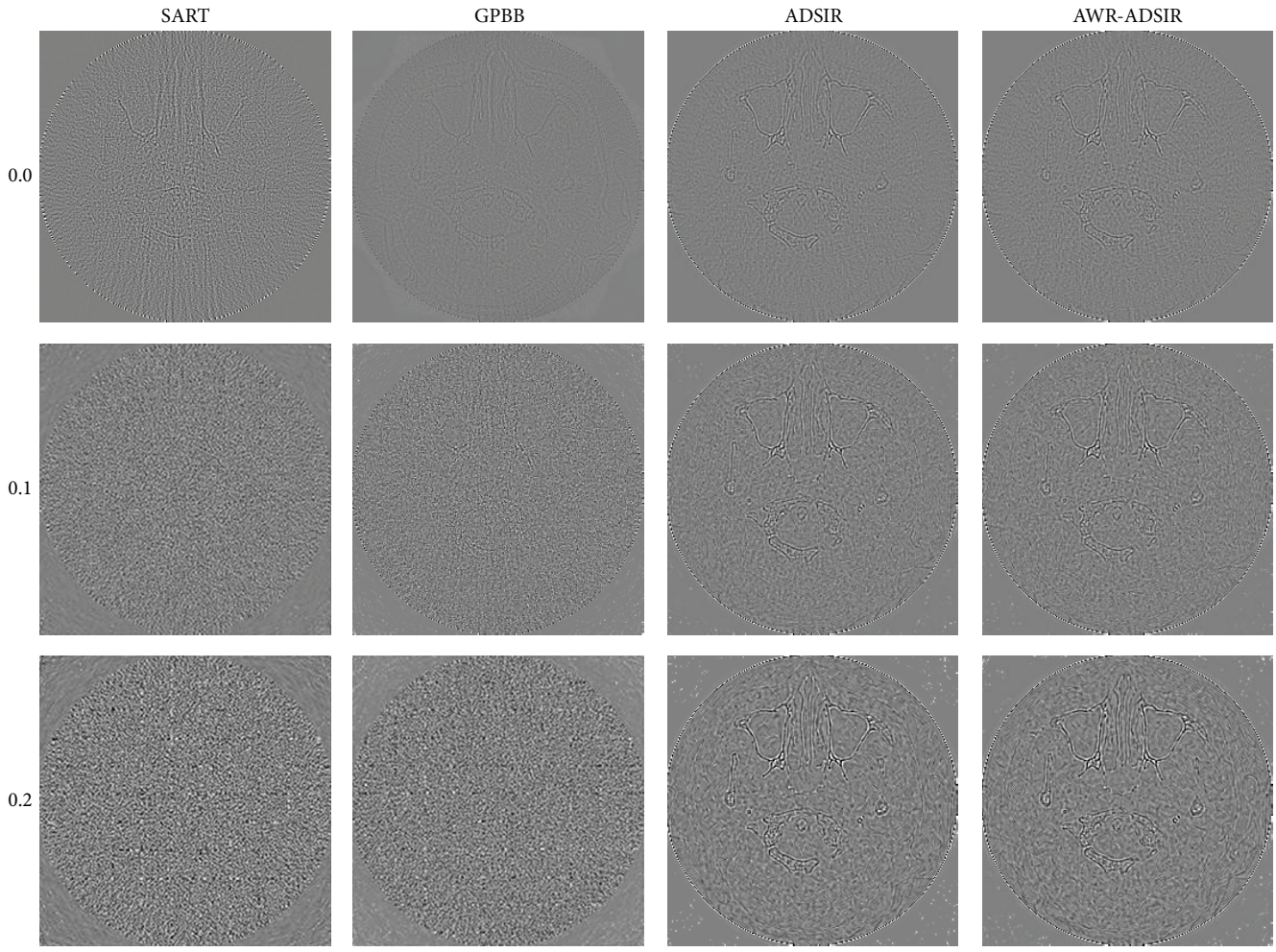


Figure 7: The difference between the reconstructed image and the original image (OI) of the human head slice image. From top to bottom, the noise levels are 0.0%, 0.1%, and 0.2% in turn. From left to right, the reconstruction algorithms are SART, GPBB, ADSIR, and AWR-ADSIR, respectively. The display window is [−100, 100] HU.

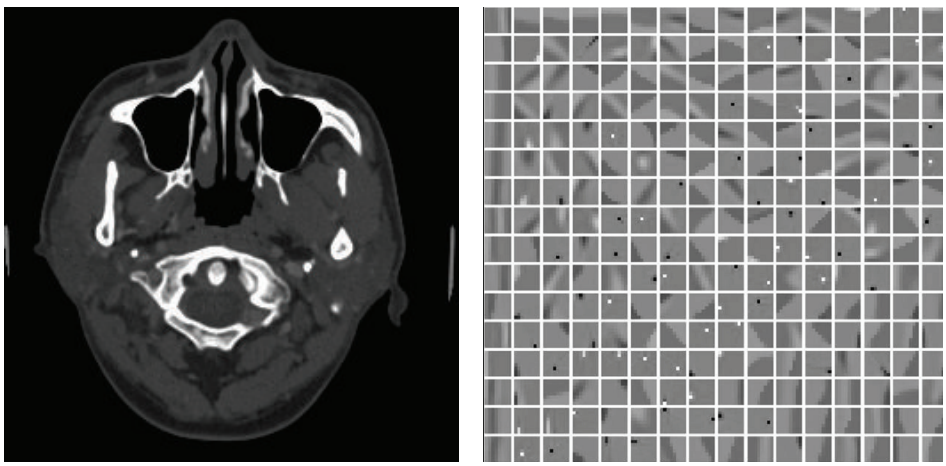


Figure 8: The image reconstructed by AWR-GDSIR and the global dictionary. The dictionary is displayed in window [−1, 1], which is constructed training based on the patches extracted from original image. The image is reconstructed by 0.0% noise level projection data displayed in window [400, 1000] HU.

effect on noise regions and preserve marginal information better.

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 the National Natural Science Foundation of China (NSFC) through Grants no. 61201117 and no. 61301042, the Natural Science Foundation of Jiangsu Province (NSFJ) through Grant no. BK2012189, and Science and Technology Program of Suzhou (no. ZXY2013001). The authors are very grateful for the CT images provided by the PET Center, Huashan Hospital, Fudan University, China.

References

[1] X. Han, J. Bian, E. L. Ritman, E. Y. Sidky, and X. Pan, "Optimization-based reconstruction of sparse images from few-view projections," *Physics in Medicine and Biology*, vol. 57, no. 16, pp. 5245–5273, 2012.

[2] B. Song, J. Park, and W. Song, "SU-E-J-14: a novel, fast, variable step size gradient method for solving simultaneous algebraic reconstruction technique (SART)-type reconstructions: an example application to CBCT," *Medical Physics*, vol. 38, no. 6, article 3444, 2011.

[3] E. J. Candes, 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.

[4] 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.

[5] J. Wang and B. Shim, "On the recovery limit of sparse signals using orthogonal matching pursuit," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4973–4976, 2012.

[6] P. T. Lauzier, J. Tang, and G.-H. Chen, "Prior image constrained compressed sensing: implementation and performance evaluation," *Medical Physics*, vol. 39, no. 1, pp. 66–80, 2012.

[7] W. W. Hager, D. T. Phan, and H. Zhang, "Gradient-based methods for sparse recovery," *SIAM Journal on Imaging Sciences*, vol. 4, no. 1, pp. 146–165, 2011.

[8] R. Volz and S. Close, "Inverse filtering of radar signals using compressed sensing with application to meteors," *Radio Science*, vol. 47, no. 4, Article ID RS0N05, 2012.

[9] G. Wang, Y. Bresler, and V. Ntziachristos, "Guest editorial compressive sensing for biomedical imaging," *IEEE Transactions on Medical Imaging*, vol. 30, no. 5, pp. 1013–1016, 2011.

[10] E. Y. Sidky and X. C. Pan, "Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization," *Physics in Medicine and Biology*, vol. 53, no. 17, pp. 4777–4807, 2008.

[11] J. C. Park, B. Song, J. S. Kim et al., "Fast compressed sensing-based CBCT reconstruction using Barzilai-Borwein formulation for application to on-line IGRT," *Medical Physics*, vol. 39, no. 3, pp. 1207–1217, 2012.

[12] Q. Xu, H. Y. Yu, X. Q. Mou, L. Zhang, J. Hsieh, and G. Wang, "Low-dose X-ray CT reconstruction via dictionary learning," *IEEE Transactions on Medical Imaging*, vol. 31, no. 9, pp. 1682–1697, 2012.

[13] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[14] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[15] I. A. Elbakri and J. A. Fessler, "Statistical image reconstruction for polyenergetic X-ray computed tomography," *IEEE Transactions on Medical Imaging*, vol. 21, no. 2, pp. 89–99, 2002.

[16] K. Karl and Z. Jun, "Iterative choices of regularization parameters in linear inverse problems," *Inverse Problems*, vol. 14, no. 5, pp. 1247–1264, 1998.

[17] J. Feng, C. Qin, K. Jia et al., "An adaptive regularization parameter choice strategy for multispectral bioluminescence tomography," *Medical Physics*, vol. 38, no. 11, pp. 5933–5944, 2011.

[18] C. Clason, B. T. Jin, and K. Kunisch, "A semismooth Newton method for L^1 data fitting with automatic choice of regularization parameters and noise calibration," *SIAM Journal on Imaging Sciences*, vol. 3, no. 2, pp. 199–231, 2010.

[19] C. Kamphuis and F. J. Beekman, "Accelerated iterative transmission CT reconstruction using an ordered subsets convex algorithm," *IEEE Transactions on Medical Imaging*, vol. 17, no. 6, pp. 1101–1105, 1998.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 794141, 10 pages
http://dx.doi.org/10.1155/2015/794141

Research Article

Multivariate Radiological-Based Models for the Prediction of Future Knee Pain: Data from the OAI

Jorge I. Galván-Tejada,[1] José M. Celaya-Padilla,[1] Victor Treviño,[1,2] and José G. Tamez-Peña[2]

[1]*Grupo de Investigación en Bioinformática, Escuela de Medicina, Tecnológico de Monterrey, 64849 Monterrey, NL, Mexico*
[2]*Departamento de Investigación e Innovación, Escuela de Medicina, Tecnológico de Monterrey, 64710 Monterrey, NL, Mexico*

Correspondence should be addressed to Jorge I. Galván-Tejada; gatejo@gmail.com

Received 8 May 2015; Revised 29 July 2015; Accepted 4 August 2015

Academic Editor: Lei Chen

In this work, the potential of X-ray based multivariate prognostic models to predict the onset of chronic knee pain is presented. Using X-rays quantitative image assessments of joint-space-width (JSW) and paired semiquantitative central X-ray scores from the Osteoarthritis Initiative (OAI), a case-control study is presented. The pain assessments of the right knee at the baseline and the 60-month visits were used to screen for case/control subjects. Scores were analyzed at the time of pain incidence (T-0), the year prior incidence (T-1), and two years before pain incidence (T-2). Multivariate models were created by a cross validated elastic-net regularized generalized linear models feature selection tool. Univariate differences between cases and controls were reported by AUC, *C*-statistics, and ODDs ratios. Univariate analysis indicated that the medial osteophytes were significantly more prevalent in cases than controls: *C*-stat 0.62, 0.62, and 0.61, at T-0, T-1, and T-2, respectively. The multivariate JSW models significantly predicted pain: AUC = 0.695, 0.623, and 0.620, at T-0, T-1, and T-2, respectively. Semiquantitative multivariate models predicted paint with *C*-stat = 0.671, 0.648, and 0.645 at T-0, T-1, and T-2, respectively. Multivariate models derived from plain X-ray radiography assessments may be used to predict subjects that are at risk of developing knee pain.

1. Introduction

Knee pain is the most common and disabling symptom of Osteoarthritis (OA) [1, 2]. This disease affects 1 in every 10 adults over 60 years in the United States and the rate of incidence is incrementing due to changes in lifestyle and life expectancy [3–7]. The prevalence and the symptomatic importance of pain in OA subjects make pain prediction a very important task for the management of OA patients. Pain is a late manifestation of a pathological change in joint tissues; therefore, the early detection of pathological process may be used to determine who is at risk of developing OA related pain. This early detection of the underling pathology may be possible with the aid of noninvasive procedures like medical imaging. Medical imaging has proved to be a very important and effective tool in OA diagnosis; it is also the most common first-hand information for physicians and a probed form to obtain a good approach to OA staging [8–13].

Due to its maturity, simplicity and broad base deployment of X-Ray, it is the primary medical imaging modality used in OA diagnosis and staging. Radiological OA has been defined as subjects presenting bone alterations (osteophytes) and reduced joint space [14]. This findings have been correlated to joint symptoms of pain and stiffness [15]; but the bony changes prognosis power have not been properly studied in longitudinal studies [16–19]. The biggest challenge facing radiological correlation to symptomatic OA is the multifactorial source of joint pain and the subjective perception of pain [20]. Other challenge has been the lack of standardized image assessment procedures that allow a proper evaluation and comparisons of OA studies. To overcome these limitations, validated subject questionnaires [21, 22] and standardized image assessments have been developed [23–26].

The Osteoarthritis Initiative (OAI) has been recollecting thousands of clinical data in OA patients, subjects at risk, and control subjects using validated questionnaires and

standardized image assessments procedures. The OAI effort brings very important information that will offer a better understanding of the disease process.

In this work, the OAI X-ray quantitative image assessments of joint space width, the central reads of Kellgren and Lawrence (K&L), and the Osteoarthritis Research Society International (OARSI) scores are explored in their association to concurrent and future knee pain. The number of radiological findings as reported by the OAI central image assessments is large: osteophytes, bone attrition, and reduced joint space at the medial and lateral aspects of the joint. The OAI quantitative image analysis of joint space also provides a set of measurements that makes data exploration a challenge. This large array of radiological features and its association to pain cannot be handled by simple statistical analysis tools. Advanced feature selection and bioinformatics tools provide a proven method to handle this complex issue [27–30]. These advanced methods automatically build simple multivariate models that best describe the association of radiological features with pain.

This work explores the use of this bioinformatics tools to build radiological multivariate models of future joint pain and concurrent joint pain, with the objective of finding what radiological features or models can be used in to determine which set of people with radiological OA findings is at greater risk developing knee pain.

This paper is organized as follows; after introduction, the patients selection and methods of selection are explained, in Data Acquisition, the process of image feature acquisition is presented; in Statistical Analysis, the complete transformations and data analysis are explained; in Results, the tables with the numerical results are presented; finally, Discussion, Conclusion, and the future work are presented.

2. Patients Methods

Study Population. "Data used in the preparation of this article were obtained from the Osteoarthritis Initiative (OAI) database, which is available for public access at http://www.oai.ucsf.edu/datarelease/." All subjects were selected from OAI databases. Based on the available information, this study was designed for right knee only.

Being a pain prediction study, the development of chronic pain in the right knee was used as the variable to look at. Using the five-year screening information, a group of subjects was selected. All subjects should not present chronic pain as a symptom in their baseline visit and should not have been medicated for pain.

Control subjects were selected under the criteria of the following: not presenting pain as a symptom since the baseline visit to 60-month visit, not presenting a symptomatic status since the baseline visit to 60-month visit, and taking no pain medication from the baseline visit to the 60-month visit. Case subjects were selected under the criteria of the following: not presenting pain as a symptom at baseline visit, not presenting a symptomatic status at baseline visit, taking no pain medication at the baseline visit, and developing chronic right knee pain in some time point after baseline and up to 60-month visit.

Only the subjects with a complete quantitative or semiquantitative X-ray assessment screening were included in the final test. Due to this last condition, two different groups were created, one for quantitative study and one for semiquantitative study. All demographic information is presented in Table 1, and the selection process is described in detail in Figure 1.

3. Data Acquisition

In this analysis, right knee assessment from the OAI datasets, "central assessment of longitudinal knee X-rays for quantitative JSW" version 1.6, was right knee assessment from OAI dataset "central reading of knee X-rays for K-L grade and individual radiographic features of knee" version 1.6, and the outcome information was chronic pain, defined by the question in the OAI dataset "right knee symptom status." This information was preanalyzed by two different radiologist groups associated with the OAI; one group evaluated the images using the OARSI quantitative grading scale [25, 31] and the semiquantitative K-L grading scale [26, 32, 33]. In Table 2, a description of the assessed features and their IDs are presented.

All X-ray images were assessed using the OAI method; automated computational software and an external reader delineate the margin of the femoral condyle and the tibial plateau; in Figure 2, an example of the software output is presented.

Using an anatomical coordinate system, an objective x-location is determined. In Figure 3, an example of the reader line is presented. According to OAI information, a study of longitudinal knee radiographs suggested that $x = 0.2$ mm to $x = 0.275$ mm may be the optimal range for measuring medial JSW(x); an example of this measurements is presented in Figure 4.

All semiquantitative variables assessed for this work included the standard OAI protocol. This vendor includes Kellgren and Lawrence (K&L) grades, individual radiographic features (IRFs) such as osteophytes, and joint space narrowing in specific anatomic locations, based on published atlases.

In general, two expert readers independently assessed each film, blinded to each other's reading and to a subject's clinical data. Baseline and follow-up films were scored while being viewed simultaneously and with the readers blinded to chronological order of the images with the baseline film known and follow-up films randomly ordered.

4. Statistical Analysis

For quantitative and semiquantitative data, using the time of pain incidence as a marker, three different groups were built: T-0, using the radiological data of the subject at the moment of chronic pain development; T-1, using the radiological data on the subject a year before chronic pain development; and T-2, using the radiological data of the subject two years before chronic pain development. Seventeen quantitative variables and nineteen semiquantitative variables were explored in this work.

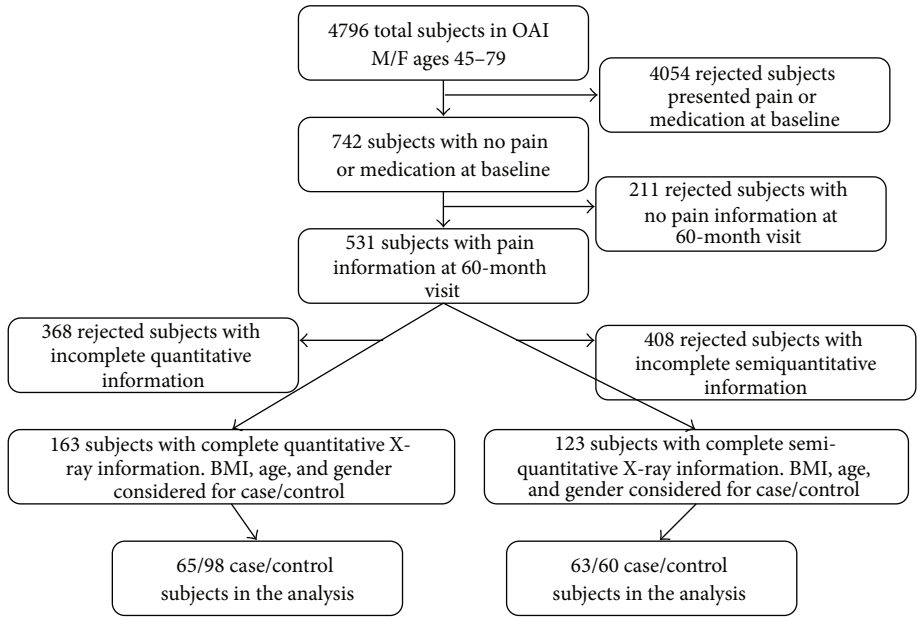


Figure 1: Subject selection.

Table 1: Demographic information.

	Quantitative analysis subjects			Semiquantitative analysis subjects		
	Cases	Controls	All	Cases	Controls	All
Subjects (females)	65 (38)	98 (55)	163 (93)	63 (35)	60 (26)	123 (61)
Average height (σ) [m]	1.69 (.09)	1.68 (.1)	1.68 (.09)	1.66 (.23)	1.69 (.1)	1.67 (.18)
Average BMI (σ)	27.05 (4.3)	26.27 (4)	26.58 (4.1)	27.27 (4.4)	28.48 (4.1)	27.86 (4.3)
Average age (σ)	62.69 (9.6)	61.80 (10.1)	62.15 (9.9)	65.02 (9.6)	66.72 (8.7)	65.85 (9.2)
Age range	46–78	45–79	45–79	46–78	47–79	46–79

Table 2: Features description.

Quantitative features		Semiquantitative features	
Feature ID	Description	Feature ID	Description
MCMJSW	Medial minimum JSW [mm]	XROSFM	Osteophytes (OARSI grades 0–3) femur medial compartment
JSW175	Medial JSW at $x = 0.175$ [mm]	XRSCFM	Sclerosis (OARSI grades 0–3) femur medial compartment
JSW200	Medial JSW at $x = 0.200$ [mm]	XRCYFM	Cysts (grades 0-1) femur medial compartment
JSW250	Medial JSW at $x = 0.250$ [mm]	XRJSM	Joint space narrowing (OARSI grades 0–3) medial compartment
JSW300	Medial JSW at $x = 0.300$ [mm]	XRCHM	Chondrocalcinosis (grades 0-1) medial compartment
JSW225	Medial JSW at $x = 0.225$ [mm]	XROSTM	Osteophytes (OARSI grades 0–3) tibia medial compartment
JSW150	Medial JSW at $x = 0.150$ [mm]	XRSCTM	Sclerosis (OARSI grades 0–3) tibia medial compartment
JSW275	Medial JSW at $x = 0.275$ [mm]	XRCYTM	Cysts (grades 0-1) tibia medial compartment
LJSW850	Lateral JSW at $x = 0.850$ [mm]	XRATTM	Attrition (OARSI grades 0–3) tibia medial compartment
LJSW900	Lateral JSW at $x = 0.900$ [mm]	XRKL	Kellgren and Lawrence (grades 0–4)
LJSW700	Lateral JSW at $x = 0.700$ [mm]	XROSFL	Osteophytes (OARSI grades 0–3) femur lateral compartment
LJSW825	Lateral JSW at $x = 0.825$ [mm]	XRSCFL	Sclerosis (OARSI grades 0–3) femur lateral compartment
LJSW750	Lateral JSW at $x = 0.750$ [mm]	XRCYFL	Cysts (grades 0-1) femur lateral compartment
LJSW875	Lateral JSW at $x = 0.875$ [mm]	XRJSL	Joint space narrowing (OARSI grades 0–3) lateral compartment
LJSW725	Lateral JSW at $x = 0.725$ [mm]	XRCHL	Chondrocalcinosis (grades 0-1) lateral compartment
LJSW800	Lateral JSW at $x = 0.800$ [mm]	XROSTL	Osteophytes (OARSI grades 0–3) tibia lateral compartment
LJSW775	Lateral JSW at $x = 0.775$ [mm]	XRSCTL	Sclerosis (OARSI grades 0–3) tibia lateral compartment
		XRCYTL	Cysts (grades 0-1) tibia lateral compartment
		XRATTL	Attrition (OARSI grades 0–3) tibia lateral compartment

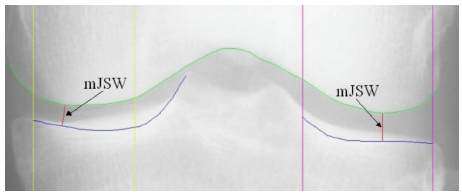


Figure 2: Output of the software on a digital knee radiograph. mJSW is the minimum JSW in a compartment. We provide mJSW for the medial compartment only (image from OAI).

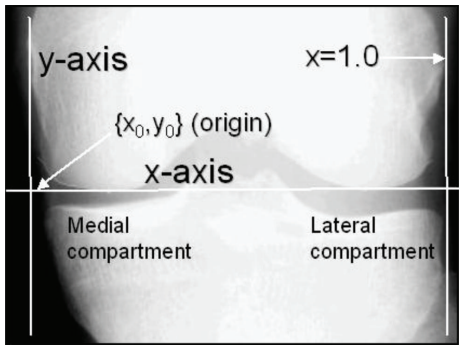


Figure 3: Landmarks and definition of coordinate system (image from OAI).

For the data analysis, the groups were analyzed using univariate and multivariate techniques. In both cases, for quantitative data, allometric association of joint height and gender to joint space width was adjusted using a linear regression [34], a common technique in related literature. All quantitative data was Z normalized using the rank inverse normal transform [35] using the standard levels of normalization reported in literature [36].

Seventeen quantitative features were measured in right knee assessments; the description of the features is shown in Table 2. To avoid the gender bias, all image features from the quantitative datasets went through a height and gender adjustment using a linear regression presented in the following:

$$\text{JSWadj} = \text{JSW} - b_0 - (\text{Height} * b_1) - (\text{Gender} * b_2), \quad (1)$$

where JSWadj represents the adjusted measurement, JSW is the original measurement, and b_0, b_1, and b_2 are the coefficients obtained from the linear regression.

Due to the nature of the distribution of the binary outcome variable, in both cases (quantitative and semiquantitative), the univariate analysis was performed using logistic regression as a cost function using all features presented in Table 2. A general linear model, odds ratios, and the area under the Receiver Operating Characteristic (ROC) curve (AUC) were calculated on each feature; the ROC curve was constructed for each quantitative analysis, and the curve is a graphical representation of the sensitivity against *1-specificity* for a binary classifier system as the discrimination threshold is varied.

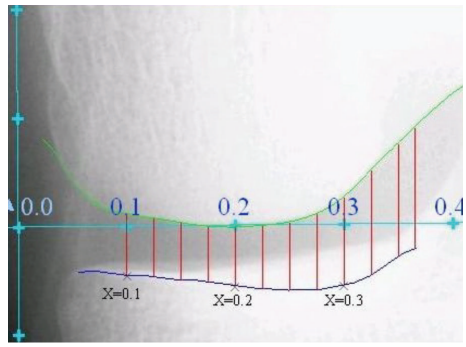


Figure 4: Measurement of JSW(x). Measurements from x = 0.150 mm to x = 0.300 mm are provided in increments of 0.025 mm (image from OAI).

Semiquantitative and quantitative data were analyzed independently. In both analyses, the different groups determined by the time of impact were tested independently to avoid bias.

In multivariate analysis, in order to select the best combination of features for the quantitative and semiquantitative prediction models, a multivariate search strategy was performed using elastic-net regularized generalized linear models as a classifier (LASSO) [37–39], with a *10-fold* cross validation as a feature selection strategy; this method is commonly used in classification works. Accuracy, AUC, C-stat, and confusion matrix were obtained; in order to minimize the residual error of the prediction model, the lambda used in this research was chosen at lambda = lambda·min [37]. Lambda for quantitative was 0.037, 0.029, and 0.069, at T-0, T-1, and T-2, respectively. Lambda for semiquantitative was 0.062, 0.062, and 0.048, at T-0, T-1, and T-2, respectively.

All the statistical analysis was performed using R software and packages [40].

5. Results

The statistical description trough the time points of each quantitative and semi quantitative features are presented in Tables 3 and 4.

In the univariate analysis, all quantitative features showed not to be predictive by it self. In the semi-quantitative features, the "Osteophytes (OARSI grades 0–3) femur medial compartment (XROSFM)" showed to be predictive. In Tables 5 and 6 the complete statistical results of all the individual features are presented.

Using multivariate analysis of quantitative data three predictive models were obtained. For the time of pain incidence, a six features predictive model obtained the best accuracy and AUC. In the one year before pain incidence, a two feature predictive model obtained the best accuracy and AUC. In the two years before the pain incidence a two features predictive model obtained the best accuracy and AUC, the resulting curves are presented in Figure 5. In Table 7, a complete results and statistical analysis of each model is presented.

Using multivariate analysis on semi-quantitative data three predictive models were obtained. For the time of pain

Table 3: Quantitative data statistical information.

	T-0				T-1				T-2			
	Mean	S.D.	Max	Min	Mean	S.D	Max	Min	Mean	S.D.	Max	Min
MCMJSW	0.010	0.931	2.106	−3.389	0.041	0.885	2.468	−3.003	0.071	0.919	2.260	−2.723
JSW175	0.005	0.961	3.652	−3.704	0.052	0.922	3.745	−3.154	0.064	0.920	3.963	−2.484
JSW200	0.011	0.986	3.547	−3.908	0.064	0.957	3.422	−3.657	0.068	0.923	3.660	−2.805
JSW250	0.006	1.021	3.586	−4.014	0.055	0.983	3.359	−3.578	0.092	0.956	3.657	−2.698
JSW300	0.061	1.067	4.123	−2.410	0.054	1.089	4.014	−2.520	0.104	1.022	4.037	−2.022
JSW225	0.008	1.008	3.355	−4.175	0.062	0.966	3.225	−3.818	0.073	0.925	3.451	−2.958
JSW150	0.019	0.958	3.744	−3.730	0.041	0.913	3.709	−3.184	0.098	0.920	3.914	−2.523
JSW275	0.032	1.040	3.834	−3.548	0.049	1.030	3.833	−3.368	0.127	0.970	3.745	−2.131
LJSW850	0.059	1.280	3.393	−4.560	0.081	1.202	3.231	−3.934	0.061	1.117	2.799	−3.958
LJSW900	0.074	1.348	3.477	−4.789	0.093	1.247	3.652	−3.825	0.055	1.206	2.697	−3.960
LJSW700	0.004	1.560	4.453	−4.415	0.094	1.516	4.533	−4.244	0.134	1.451	4.522	−5.466
LJSW825	0.054	1.287	3.332	−4.672	0.091	1.188	3.380	−3.484	0.087	1.121	3.183	−3.393
LJSW750	0.025	1.387	4.000	−5.452	0.100	1.284	3.524	−3.407	0.128	1.258	3.868	−3.921
LJSW875	0.065	1.282	3.324	−4.645	0.066	1.192	3.040	−3.982	0.035	1.135	2.488	−4.237
LJSW725	0.016	1.433	4.446	−5.325	0.104	1.351	3.798	−3.464	0.188	1.426	4.082	−5.838
LJSW775	0.030	1.320	3.695	−4.949	0.121	1.228	3.406	−3.048	0.114	1.189	3.316	−3.001
LJSW800	0.040	1.298	3.555	−4.513	0.120	1.201	3.498	−3.070	0.099	1.136	3.091	−3.077

Table 4: Semiquantitative data statistical information.

	T-0				T-1				T-2			
	0	1	2	3	0	1	2	3	0	1	2	3
XROSFM	80	31	5	7	81	29	5	8	82	29	5	7
XRSCFM	97	17	9	0	98	16	9	0	98	18	7	0
XRCYFM	122	1	0	0	122	1	0	0	122	1	0	0
XRJSM	66	40	16	1	66	42	15	0	68	41	14	0
XRCHM	121	2	0	0	120	3	0	0	121	2	0	0
XROSTM	50	62	9	2	51	61	9	2	51	62	8	2
XRSCTM	97	13	13	0	97	14	12	0	101	10	12	0
XRCYTM	119	4	0	0	119	4	0	0	119	4	0	0
XRATTM	123	0	0	0	123	0	0	0	123	0	0	0
XRKL	18	12	69	21	19	15	69	19	21	16	67	18
XROSFL	85	30	4	4	89	27	3	4	88	29	3	3
XRSCFL	115	3	2	3	116	4	2	1	117	3	2	1
XRCYFL	122	1	0	0	122	1	0	0	122	1	0	0
XRJSL	111	5	3	3	113	5	3	2	114	4	4	1
XRCHL	121	2	0	0	121	2	0	0	121	2	0	0
XROSTL	84	30	7	2	88	27	6	2	91	26	4	2
XRSCTL	115	1	5	2	117	1	4	1	117	2	3	1
XRCYTL	118	5	0	0	120	3	0	0	120	3	0	0
XRATTL	121	2	0	0	121	2	0	0	121	2	0	0

incidence, a four features predictive model obtained the best accuracy and C-stat. In the one year before pain incidence, a two feature predictive model obtained the best accuracy and C-stat. In the two years before the pain incidence a three features predictive model obtained the best accuracy and C-stat. In Table 8, a complete results and statistical analysis of each model is presented.

6. Discussion

This case-control longitudinal analysis of subjects with chronic right knee pain found an association between radiographic evidence of early OA changes and the future onset of chronic pain symptoms. Specifically, it was found that particular radiological changes in knee anatomy are present at least two years in advance of the onset of chronic pain for a selected group of patients. Therefore, these results may indicate that specific changes in joint space and bony structure are risk factors for the future development of OA related pain. These findings reinforce the conclusions of several population-based studies that have reported that persons with radiographic knee OA are at higher risk of pain development compared to persons without radiological OA [11–13, 16, 18, 41, 42]. The reported radiological features may be added to the well-known risk factors of OA severity like varus-valgus mal-alignment [43].

The quantitative driven multivariate predictive models presented in Table 7 may indicate an association between the medial cartilage abnormalities and the chronic pain. The presence of lateral and medial osteophytes in the semiquantitative multivariate models reported in Table 8 was associated with chronic pain development. The changes in the medial JSW (JSW $x = 0.275$ or $x = 0.300$), bony damage, and Chondrocalcinosis were present two years before the pain occurrence. The individual features (Tables 5 and 6) were not as predictive as the multivariate models as expected given

Table 5: Univariate quantitative statistical results.

	T-0					T-1					T-2				
	P	AUC	OR	2.50%	97.50%	*P*	AUC	OR	2.50%	97.50%	*P*	AUC	OR	2.50%	97.50%
MCMJSW	0.84	0.48	1.23	0.87	1.75	0.99	0.50	1.00	0.70	1.43	0.27	0.46	1.23	0.87	1.75
JSW175	0.93	0.50	1.26	0.89	1.80	0.61	0.49	1.09	0.78	1.54	0.20	0.55	1.26	0.89	1.80
JSW200	0.83	0.51	1.23	0.88	1.76	0.54	0.51	1.11	0.80	1.55	0.23	0.54	1.23	0.88	1.76
JSW250	0.89	0.50	1.39	1.00	1.98	0.54	0.52	1.10	0.80	1.53	0.06	0.57	1.39	1.00	1.98
JSW300	0.34	0.53	1.43	1.05	2.00	0.18	0.55	1.22	0.91	1.65	0.03	0.58	1.43	1.05	2.00
JSW225	0.87	0.50	1.28	0.91	1.82	0.55	0.52	1.11	0.80	1.54	0.17	0.55	1.28	0.91	1.82
JSW150	0.74	0.52	1.25	0.89	1.79	0.69	0.51	1.07	0.76	1.52	0.20	0.55	1.25	0.89	1.79
JSW275	0.58	0.49	1.49	1.07	2.12	0.32	0.54	1.17	0.86	1.60	0.02	0.59	1.49	1.07	2.12
LJSW850	0.45	0.58	1.18	0.89	1.59	0.36	0.57	1.13	0.87	1.49	0.27	0.58	1.18	0.89	1.59
LJSW900	0.39	0.60	1.17	0.90	1.54	0.23	0.58	1.17	0.91	1.52	0.25	0.57	1.17	0.90	1.54
LJSW700	0.96	0.52	1.17	0.94	1.48	0.54	0.55	1.07	0.87	1.32	0.18	0.56	1.17	0.94	1.48
LJSW825	0.50	0.58	1.20	0.90	1.61	0.37	0.57	1.13	0.87	1.49	0.23	0.57	1.20	0.90	1.61
LJSW750	0.81	0.54	1.18	0.92	1.54	0.45	0.56	1.10	0.86	1.41	0.21	0.58	1.18	0.92	1.54
LJSW875	0.42	0.59	1.17	0.88	1.56	0.30	0.57	1.15	0.89	1.52	0.30	0.57	1.17	0.88	1.56
LJSW725	0.88	0.53	1.17	0.92	1.50	0.49	0.54	1.09	0.86	1.38	0.22	0.57	1.25	1.00	1.60
LJSW775	0.74	0.55	1.20	0.92	1.58	0.33	0.57	1.14	0.88	1.48	0.21	0.57	1.20	0.92	1.58
LJSW800	0.63	0.56	1.20	0.91	1.60	0.31	0.57	1.15	0.88	1.51	0.22	0.58	1.20	0.91	1.60

Table 6: Univariate semiquantitative statistical results.

	T-0					T-1					T-2				
	P	*C*-stat	OR	2.50%	97.50%	*P*	*C*-stat	OR	2.50%	97.50%	*P*	*C*-stat	OR	2.50%	97.50%
XROSFM	0.01	0.62	2.11	1.28	3.81	0.01	0.62	2.06	1.27	3.64	0.01	0.61	1.97	1.21	3.51
XRSCFM	0.22	0.53	1.48	0.80	2.85	0.28	0.53	1.41	0.77	2.72	0.40	0.52	1.33	0.70	2.63
XRCYFM	0.99	0.51	N/A	N/A	N/A	0.99	0.51	N/A	N/A	N/A	0.99	0.51	N/A	N/A	N/A
XRJSM	0.72	0.50	1.09	0.68	1.76	0.98	0.49	1.01	0.61	1.66	0.69	0.53	0.90	0.54	1.50
XRCHM	0.99	0.52	N/A	N/A	N/A	0.99	0.52	N/A	N/A	N/A	0.99	0.52	N/A	N/A	N/A
XROSTM	0.78	0.53	0.93	0.54	1.57	0.68	0.53	0.90	0.53	1.52	0.78	0.53	0.93	0.54	1.58
XRSCTM	0.78	0.50	1.08	0.63	1.88	0.88	0.50	1.04	0.60	1.84	0.91	0.52	0.97	0.55	1.71
XRCYTM	0.96	0.50	0.95	0.11	8.14	0.96	0.50	0.95	0.11	8.14	0.96	0.50	0.95	0.11	8.14
XRATTM	0.79	0.50	N/A	N/A	N/A	0.79	0.50	N/A	N/A	N/A	0.79	0.50	N/A	N/A	N/A
XRKL	0.07	0.58	1.42	0.98	2.12	0.39	0.53	1.18	0.81	1.75	0.78	0.51	1.05	0.72	1.54
XROSFL	0.02	0.59	2.00	1.14	3.82	0.08	0.56	1.68	0.97	3.14	0.05	0.57	1.89	1.05	3.73
XRSCFL	0.23	0.52	1.62	0.80	4.33	0.55	0.51	1.33	0.54	3.99	0.69	0.50	1.21	0.48	3.53
XRCYFL	0.99	0.51	N/A	N/A	N/A	0.99	0.51	N/A	N/A	N/A	0.99	0.51	N/A	N/A	N/A
XRJSL	0.23	0.53	1.48	0.82	3.10	0.62	0.51	1.20	0.59	2.65	0.62	0.51	1.22	0.57	2.88
XRCHL	0.99	0.52	N/A	N/A	N/A	0.99	0.52	N/A	N/A	N/A	0.99	0.52	N/A	N/A	N/A
XROSTL	0.05	0.57	1.77	1.02	3.27	0.18	0.54	1.48	0.85	2.71	0.20	0.53	1.49	0.83	2.86
XRSCTL	0.18	0.53	1.70	0.85	4.43	0.47	0.52	1.37	0.61	3.70	0.57	0.51	1.29	0.55	3.63
XRCYTL	0.22	0.52	4.00	0.57	79.50	0.59	0.51	1.93	0.18	42.28	0.59	0.51	1.93	0.18	42.28
XRATTL	0.99	0.52	N/A	N/A	N/A	0.99	0.52	N/A	N/A	N/A	0.99	0.52	N/A	N/A	N/A

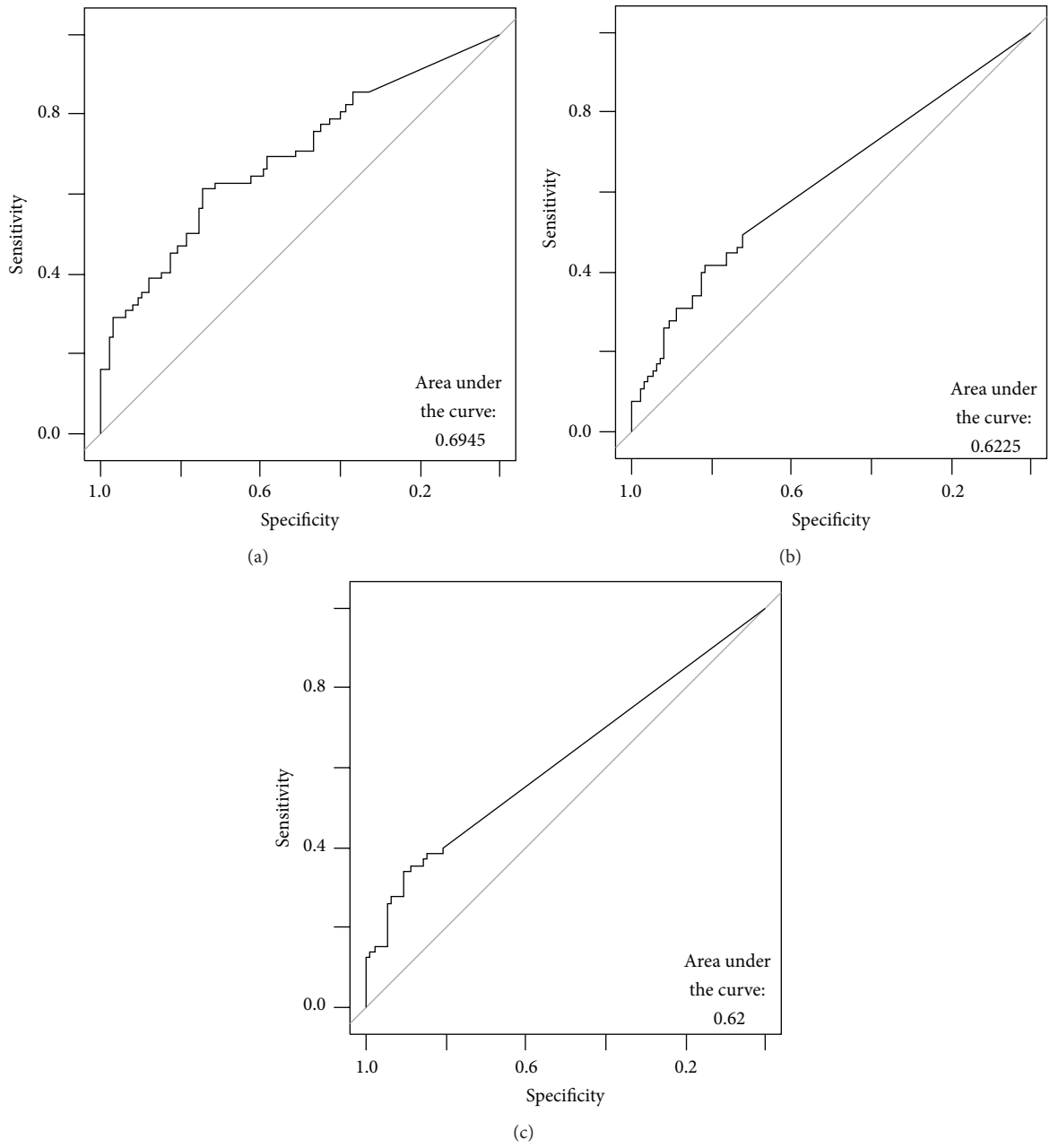


Figure 5: AUC curves on quantitative models: (a) curve for T-0, (b) curve for T-1, and (c) curve for T-2.

the fact that OA is a whole organ disease that affects several tissues at the same time [44].

When comparing these results to our previous efforts [45, 46], we saw an increase in the AUC from 0.652, 0.617, and 0.674 to 0.695, 0.623, and 0.620, at T-0, T-1, and T-2 time points. Furthermore, the process will take less than 2 min of computation time contrasted to the 48 hrs of computation using the same machine. The models obtained using LASSO were more stable since the process is deterministic compared to the stochastic nature of the genetic algorithm of the original work.

There are several limitations to our study. First of all, pain is a subjective outcome that changes from person to person. Second, we limit the inclusion to subjects that were not taking pain medications; therefore, the number of those developed pains during the observation period was small. Third, we limit the exploration to right knee findings, and unilateral pain may be affected by the symptoms of the contra lateral knee. Given these limitations, we cannot generalize the findings and the external validation of the results is required to assess the clinical applicability of the models.

7. Conclusions

Even though pain is a very complex and subjective clinical outcome, the systematic analysis of objective radiological

Table 7: Multivariate models on quantitative X-ray data.

Time point	T-0			T-1			T-2		
Model variables	Medial JSW at $x = 0.250$ [mm] Lateral JSW at $x = 0.700$ [mm] Lateral JSW at $x = 0.750$ [mm] Lateral JSW at $x = 0.800$ [mm] Medial JSW at $x = 0.275$ [mm] Lateral JSW at $x = 0.875$ [mm]			Medial JSW at $x = 0.275$ [mm] Lateral JSW at $x = 0.875$ [mm]			Medial JSW at $x = 0.300$ [mm] Medial JSW at $x = 0.275$ [mm]		
Accuracy	Accuracy: 0.688			Accuracy: 0.626			Accuracy: 0.632		
C.I.	95% CI: (0.6096, 0.7583)			95% CI: (0.5467, 0.7002)			95% CI: (0.5529, 0.706)		
AUC	0.695			0.623			0.620		
Confusion matrix	Pain			Pain			Pain		
	Pred	1	0	Pred	0	1	Pred	0	1
	0	91	43	0	94	57	0	92	54
	1	7	19	1	4	8	1	6	11

Table 8: Multivariate models on semiquantitative X-ray data.

Time point	T-0			T-1			T-2		
Model variables	Osteophytes (OARSI grades 0–3) femur medial compartment Chondrocalcinosis (grades 0-1) medial compartment Osteophytes (OARSI grades 0–3) femur lateral compartment Osteophytes (OARSI grades 0–3) tibia lateral compartment			Osteophytes (OARSI grades 0–3) femur medial compartment Chondrocalcinosis (grades 0-1) medial compartment			Osteophytes (OARSI grades 0–3) femur medial compartment Chondrocalcinosis (grades 0-1) medial compartment Osteophytes (OARSI grades 0–3) femur lateral compartment		
Accuracy	Accuracy: 0.6423			Accuracy: 0.626			Accuracy: 0.618		
C.I.	95% CI: (0.5509, 0.7267)			95% CI: (0.5342, 0.7116)			95% CI: (0.5259, 0.704)		
C-stat	0.671			0.648			0.645		
Confusion matrix	PAIN			PAIN			PAIN		
	Pred	0	1	Pred	0	1	Pred	0	1
	0	45	29	0	46	32	0	46	33
	1	15	34	1	14	31	1	14	30

features was able to find a multivariate model that indicates that there are certain anatomical features that preceded the development of knee pain. A biomarker based on those features may be used to help physicians to choose the best therapy or course of action for patients that present those features. This represents a great area of impact especially in developing countries, where access to the high level of health care system is very restricted.

Based on these results, it is evident that multivariate models obtained by computational methods can make better use of radiological characteristics, increasing the chance for the future development of an effective computer assisted diagnosis and/or treatment selection system.

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 the Consejo Nacional de Ciencia y Tecnología (CONACYT), by Grant 16864 Ciencia Básica from CONACYT, and by Catédra de Bioinformática (CAT220) from Tecnológico de Monterrey. Jorge I. Galván-Tejada gives thanks to PROMEP for partially supporting his doctoral studies. "The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258, N01-AR-2-2259, N01-AR-2-2260, N01-AR-2-2261, and N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories, Novartis Pharmaceuticals Corporation, GlaxoSmithKline, and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This paper was prepared using an OAI public use dataset and does not necessarily

reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners."

References

[1] T. Neogi, "The epidemiology and impact of pain in osteoarthritis," *Osteoarthritis and Cartilage*, vol. 21, no. 9, pp. 1145–1153, 2013.

[2] M. Agaliotis, M. Fransen, L. Bridgett et al., "Risk factors associated with reduced work productivity among people with chronic knee pain," *Osteoarthritis and Cartilage*, vol. 21, no. 9, pp. 1160–1169, 2013.

[3] D. K. White, C. Tudor-Locke, D. T. Felson et al., "Do radiographic disease and pain account for why people with or at high risk of knee osteoarthritis do not meet physical activity guidelines?" *Arthritis and Rheumatism*, vol. 65, no. 1, pp. 139–147, 2013.

[4] T. Neogi, M. A. Bowes, J. Niu et al., "Magnetic resonance imaging-based three-dimensional bone shape of the knee predicts onset of knee osteoarthritis: data from the osteoarthritis initiative," *Arthritis and Rheumatism*, vol. 65, no. 8, pp. 2048–2058, 2013.

[5] C. J. Colbert, O. Almagor, J. S. Chmiel et al., "Excess body weight and four-year function outcomes: comparison of African Americans and whites in a prospective study of osteoarthritis," *Arthritis Care and Research*, vol. 65, no. 1, pp. 5–14, 2013.

[6] D. L. Riddle and P. W. Stratford, "Body weight changes and corresponding changes in pain and function in persons with symptomatic knee osteoarthritis: a cohort study," *Arthritis Care and Research*, vol. 65, no. 1, pp. 15–22, 2013.

[7] S. K. Tanamas, A. E. Wluka, M. Davies-Tuck et al., "Association of weight gain with incident knee pain, stiffness, and functional difficulties: a longitudinal study," *Arthritis Care and Research*, vol. 65, no. 1, pp. 34–43, 2013.

[8] A. Guermazi, J. Niu, D. Hayashi et al., "Prevalence of abnormalities in knees detected by MRI in adults without knee osteoarthritis: population based observational study (Framingham Osteoarthritis study)," *British Medical Journal*, vol. 345, Article ID e5339, 2012.

[9] W. Wirth, J. Duryea, M.-P. H. Le Graverand et al., "Direct comparison of fixed flexion, radiography and MRI in knee osteoarthritis: responsiveness data from the osteoarthritis initiative," *Osteoarthritis and Cartilage*, vol. 21, no. 1, pp. 117–125, 2013.

[10] S. Cotofana, B. T. Wyman, O. Benichou et al., "Relationship between knee pain and the presence, location, size and phenotype of femorotibial denuded areas of subchondral bone as visualized by MRI," *Osteoarthritis and Cartilage*, vol. 21, no. 9, pp. 1214–1222, 2013.

[11] I. A. C. Baert, F. Staes, S. Truijen et al., "Weak associations between structural changes on MRI and symptoms, function and muscle strength in relation to knee osteoarthritis," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 22, no. 9, pp. 2013–2025, 2014.

[12] T. Neogi, D. Felson, J. Niu et al., "Association between radiographic features of knee osteoarthritis and pain: results from two cohort studies," *The British Medical Journal*, vol. 339, no. 7719, pp. 498–501, 2009.

[13] I. K. Haugen, B. Slatkowsky-Christensen, P. Bøyesen, D. van der Heijde, and T. K. Kvien, "Cross-sectional and longitudinal associations between radiographic features and measures of pain and physical function in hand osteoarthritis," *Osteoarthritis and Cartilage*, vol. 21, no. 9, pp. 1191–1198, 2013.

[14] R. Altman, E. Asch, and D. Bloch, "Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis of the knee," *Arthritis & Rheumatism*, vol. 29, no. 8, pp. 1039–1052, 1986.

[15] J. Dekker, B. Boot, L. H. V. van der Woude, and J. W. J. Bijlsma, "Pain and disability in osteoarthritis: a review of biobehavioral mechanisms," *Journal of Behavioral Medicine*, vol. 15, no. 2, pp. 189–214, 1992.

[16] M. B. Kinds, A. C. A. Marijnissen, J. W. J. Bijlsma, M. Boers, F. P. J. G. Lafeber, and P. M. J. Welsing, "Quantitative radiographic features of early knee osteoarthritis: development over 5 years and relationship with symptoms in the CHECK cohort," *The Journal of Rheumatology*, vol. 40, no. 1, pp. 58–65, 2013.

[17] N. A. Glass, J. C. Torner, L. A. Frey Law et al., "The relationship between quadriceps muscle weakness and worsening of knee pain in the MOST cohort: a 5-year longitudinal study," *Osteoarthritis and Cartilage*, vol. 21, no. 9, pp. 1154–1159, 2013.

[18] Y. Shimura, H. Kurosawa, Y. Sugawara et al., "The factors associated with pain severity in patients with knee osteoarthritis vary according to the radiographic disease severity: a cross-sectional study," *Osteoarthritis and Cartilage*, vol. 21, no. 9, pp. 1179–1184, 2013.

[19] J. R. Hochman, A. M. Davis, J. Elkayam, L. Gagliese, and G. A. Hawker, "Neuropathic pain symptoms on the modified painDETECT correlate with signs of central sensitization in knee osteoarthritis," *Osteoarthritis and Cartilage*, vol. 21, no. 9, pp. 1236–1242, 2013.

[20] D. J. Hunter, M.-P. H. Le Graverand, and F. Eckstein, "Radiologic markers of osteoarthritis progression," *Current Opinion in Rheumatology*, vol. 21, no. 2, pp. 110–117, 2009.

[21] N. Bellamy, W. W. Buchanan, C. H. Goldsmith, J. Campbell, and L. W. Stitt, "Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee," *The Journal of Rheumatology*, vol. 15, no. 12, pp. 1833–1840, 1988.

[22] E. M. Roos and L. S. Lohmander, "The Knee injury and Osteoarthritis Outcome Score (KOOS): from joint injury to osteoarthritis," *Health and Quality of Life Outcomes*, vol. 1, article 64, 2003.

[23] J. H. Kellgren and J. S. Lawrence, "Radiological assessment of osteo-arthrosis," *Annals of the Rheumatic Diseases*, vol. 16, no. 4, pp. 494–502, 1957.

[24] R. D. Altman and G. Gold, "Atlas of individual radiographic features in osteoarthritis, revised," *Osteoarthritis and Cartilage*, vol. 15, pp. A1–A56, 2007.

[25] J. Duryea, J. Li, C. G. Peterfy, C. Gordon, and H. K. Genant, "Trainable rule-based algorithm for the measurement of joint space width in digital radiographic images of the knee," *Medical Physics*, vol. 27, no. 3, pp. 580–591, 2000.

[26] J. Kellgren and J. Lawrence, *Atlas of Standard Radiographs: The Epidemiology of Chronic Rheumatism*, vol. 2, Blackwell Scientific Publications, Oxford, UK, 1963.

[27] J. Galván-Tejada, A. Martinez-Torteya, S. Totterman, J. Farber, V. Treviño, and J. Tamez-Pena, "A wide association study of predictors of future knee pain: data from the osteoarthritis initiative," *Osteoarthritis and Cartilage*, vol. 20, supplement 1, p. S85, 2012.

[28] A. Martinez-Torteya, J. Galván-Tejada, S. Totterman, J. Farber, V. Treviño, and J. Tamez-Pena, "Can T2 relaxation be used to

predict koos other symptoms?—data from the osteoarthritis initiative," *Osteoarthritis and Cartilage*, vol. 20, pp. S208–S209, 2012.

[29] A. Martínez-Torteya, V. M. Treviño-Alvarado, and J. G. Tamez-Peña, "Improved multimodal biomarkers for Alzheimer's disease and mild cognitive impairment diagnosis: data from ADNI," in *Medical Imaging 2013: Computer-Aided Diagnosis*, vol. 8670 of *Proceedings of SPIE*, 2013.

[30] A. M. Torteya, J. G. T. Peña, and V. M. T. Alvarado, "Multivariate predictors of clinically relevant cognitive decay: a wide association study using available data from ADNI," *Alzheimer's & Dementia*, vol. 8, no. 4, pp. P285–P286, 2012.

[31] G. Neumann, D. Hunter, M. Nevitt et al., "Location specific radiographic joint space width for osteoarthritis progression," *Osteoarthritis and Cartilage*, vol. 17, no. 6, pp. 761–765, 2009.

[32] D. T. Felson, M. C. Nevitt, M. Yang et al., "A new approach yields high rates of radiographic progression in knee osteoarthritis," *The Journal of Rheumatology*, vol. 35, no. 10, pp. 2047–2054, 2008.

[33] C. B. Hing, M. A. Harris, V. Ejindu, and N. Sofat, "The application of imaging in osteoarthritis," in *Principles of Osteoarthritis—Its Definition, Character, Derivation and Modality-Related Recognition*, chapter 4, InTech, Rijeka, Croatia, 2012.

[34] P. Suri, D. J. Hunter, J. Rainville, A. Guermazi, and J. N. Katz, "Presence and extent of severe facet joint osteoarthritis are associated with back pain in older adults," *Osteoarthritis and Cartilage*, vol. 21, no. 9, pp. 1199–1206, 2013.

[35] T. M. Beasley, S. Erickson, and D. B. Allison, "Rank-based inverse normal transformations are increasingly used, but are they merited?" *Behavior Genetics*, vol. 39, no. 5, pp. 580–595, 2009.

[36] D. Pasta, "Learning when to be discrete: continuous vs. categorical predictors," in *Proceedings of the SAS Global Forum*, Washington, DC, USA, March 2009.

[37] J. Friedman, T. Hastie, and R. Tibshirani, "glmnet: lasso and elastic-net regularized generalized linear models," R Package Version, 2009.

[38] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[39] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 73, no. 3, pp. 273–282, 2011.

[40] The R Project for Statistical Computing, https://www.r-project.org/.

[41] S. Muraki, H. Oka, T. Akune et al., "Prevalence of radiographic knee osteoarthritis and its association with knee pain in the elderly of Japanese population-based cohorts: the ROAD study," *Osteoarthritis and Cartilage*, vol. 17, no. 9, pp. 1137–1143, 2009.

[42] L. Braga, J. B. Renner, T. A. Schwartz et al., "Differences in radiographic features of knee osteoarthritis in African-Americans and Caucasians: the Johnston County Osteoarthritis Project," *Osteoarthritis and Cartilage*, vol. 17, no. 12, pp. 1554–1561, 2009.

[43] C. B. Chang, I. J. Koh, E. S. Seo, Y. G. Kang, S. C. Seong, and T. K. Kim, "The radiographic predictors of symptom severity in advanced knee osteoarthritis with varus deformity," *The Knee*, vol. 18, no. 6, pp. 456–460, 2011.

[44] A. R. Poole, "Osteoarthritis as a whole joint disease," *HSS Journal*, vol. 8, no. 1, pp. 4–6, 2012.

[45] J. I. Galván-Tejada, J. M. Celaya-Padilla, A. Martínez-Torteya, J. Rodriguez-Rojas, V. Treviño, and J. G. Tamez-Peña, "Wide association study of radiological features that predict future knee OA pain: data from the OAI," in *Medical Imaging: Computer-Aided Diagnosis*, vol. 9035 of *Proceedings of SPIE*, International Society for Optics and Photonics, San Diego, Calif, USA, February 2014.

[46] J. Galvan-Tejada, V. Treviño, S. Totterman, and J. Tamez-Pena, "Osteoarthritis pain prediction using X-ray features: data from OAI," *Osteoarthritis and Cartilage*, vol. 22, pp. S275–S276, 2014.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 124630, 12 pages
http://dx.doi.org/10.1155/2015/124630

Research Article

Nonsynonymous Single-Nucleotide Variations on Some Posttranslational Modifications of Human Proteins and the Association with Diseases

Bo Sun,[1,2] Menghuan Zhang,[1,2] Peng Cui,[1,2] Hong Li,[2] Jia Jia,[2] Yixue Li,[1,2] and Lu Xie[2]

[1]*School of Life Science and Biotechnology, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China*
[2]*Shanghai Center for Bioinformation Technology, Shanghai Academy of Science and Technology, 1278 Ke Yuan Road, Shanghai 201203, China*

Correspondence should be addressed to Lu Xie; xielu@scbit.org

Received 25 March 2015; Accepted 12 May 2015

Academic Editor: Lin Lu

Protein posttranslational modifications (PTMs) play key roles in a variety of protein activities and cellular processes. Different PTMs show distinct impacts on protein functions, and normal protein activities are consequences of all kinds of PTMs working together. With the development of high throughput technologies such as tandem mass spectrometry (MS/MS) and next generation sequencing, more and more nonsynonymous single-nucleotide variations (nsSNVs) that cause variation of amino acids have been identified, some of which result in the damage of PTMs. The damaged PTMs could be the reason of the development of some human diseases. In this study, we elucidated the proteome wide relationship of eight damaged PTMs to human inherited diseases and cancers. Some human inherited diseases or cancers may be the consequences of the interactions of damaged PTMs, rather than the result of single damaged PTM site.

1. Introduction

More than 200 different types of protein posttranslational modifications (PTMs) have been detected. PTMs are involved in many protein activities and cellular processes, such as protein folding, stability, conformation, and some significant regulatory mechanisms [1]. For instance, reversible phosphorylation is involved in conformational changes of enzymes, which results in their activation and deactivation in signaling transduction [2]; the proteins with attached single ubiquitin (Ub) or poly-Ub chains are associated with gene transcription, DNA repair and replication, intracellular trafficking, and virus budding [3]; methylation at certain residues of histones can regulate gene expression [4], and glycosylation is responsible for targeting substrates and changing protein half-life [2].

With the development of high-throughput sequencing technology, gene mutation detection has become another important resource to investigate regulatory mechanisms and cellular processes. Some databases such as dbSNP [5] and SNVDis [6] curated such mutation data. Other secondary databases curated mutation data annotated to the phenotype or diseases, such as Clinvar [7], COSMIC [8], and SwissVar [9]. These databases provide resources to analyze the effect of mutations on human health. However protein activities are closer to disease activities. Either at genomic or at proteomic level, mutations have significant impact on normal gene or protein function, and human diseases could be associated with mutations like nonsynonymous single-nucleotide variations (nsSNVs) on amino acids. Yet how gene mutations affect protein activities through posttranslational modification sites have not been widely studied.

A PTM site that bears nsSNVs can be defined as damaged PTM. Recently, large-scale studies have shown that damaged PTMs caused by numerous inherited and somatic amino acid substitutions [10] have profound impact on both gene and protein function [11], and they are associated with human cancer [12]. One instance is that mutation S215R occurring on the PTMs of TP53 could result in breast cancer [13]; another is mutation of T286 in cyclin D1 (CCND1) causing the loss of

phosphorylation of T286 is involved in nuclear accumulation of cyclin D1 in esophageal cancer [14].

However, some of these previous studies concluded the relationship between damaged PTMs and human health based on predications; some focused only on cancers and many focused on only unique type of PTM. Although data of both gene mutations and PTMs are increasing fast, the proteome-wide analysis on the relationship between damaged PTMs and human diseases is not well studied. In this work, we chose eight experimentally demonstrated damaged PTMs to elucidate their association to human diseases including inherited diseases and cancers (somatic diseases). These eight types of damaged PTMs include amino acid variations on Phosphorylation, Ubiquitylation, Acetylation, Glycosylation, Methylation, SUMOylation, Hydroxylation, and Sulfation, which have been well proved to play key roles in important cellular processes and have close relationship with human disease development; moreover, some cross talks among them have been recently revealed in the view of systematic biology [15, 16]. In this study, we focused on the effect of nsSNVs affecting the functions of these eight important normal PTMs and established a new protocol to analyze and view how these damaged PTMs are associated with human diseases.

2. Materials and Methods

2.1. Datasets. The eight human PTM data sets of Phosphorylation, Ubiquitylation, Acetylation, Glycosylation, Methylation, SUMOylation, Hydroxylation, and Sulfation were obtained from SysPTM 2.0 (released in June, 2013) [17], which integrated PTMs from public resources as well as manually curated MS/MS identified PTMs from experimental research articles, and dbPTM 3.0 (released in June, 2012) [18]. In this study, we only collected human-related PTMs, and we chose the most frequently modified residues for each type of PTM, respectively. For Phosphorylation, we chose His, Ser, Thr, and Tyr; for Ubiquitylation, we chose Lys; for Acetylation, we chose Ala, Gly, Lys, Met, Ser, and Thr; for Glycosylation, we chose Lys, Ser, and Thr; for Hydroxylation, we chose Asn, Pro, and Lys; for Methylation, we chose Lys, Arg; for Sulfation, we chose Ser, Thr, and Tyr; for SUMOylation, we chose Lys.

The inherited-diseases-related nsSNVs were obtained from ClinVar (accessed in November, 2013) [7], dbSNP (build 141) [5], and SwissVar [9]. Cancer-related nonsynonymous single-nucleotide variations (nsSNVs) data were retrieved from COSMIC [8], TCGA (https://tcga-data.nci.nih.gov/tcga/), and SNVDis [6]; neutral nsSNVs were extracted based on dbSNP (build 141) [5], excluding cancer-related SNVs that overlapped with those in COSMIC and TCGA, and other deleterious nsSNVs were filtered by UniProtKB/Swiss-Prot (UniProt released in October, 2013) [19] and PolyPhen-2 [20] which curated credible nsSNVs mapped on UniProtKB. Then we mapped all these nsSNVs to UniprotKB according to the accession number.

2.2. Mapping PTM Sites with nsSNV Sites. For phosphorylation mapping, we set three criteria: exact match; ±2 sites around the phosphorylated amino acid; ±7 sites around the phosphorylated amino acid [21]. As for the remaining seven types of PTMs studied, we set two criteria: exact match; ±2 sites around the modified amino acid. For phosphorylation, which is the most widespread type of PTM used in cellular signal transduction [22], in general, protein kinases show a strong selectivity for the primary sequence around the phosphorylation residues such as serine (S), threonine (T), and tyrosine (Y) [12], so we chose the maximum range up to ±7 sites around the phosphorylation sites. However, for ubiquitylation, which is commonly known as a type of PTM that targets proteins for degradation [23], by contrast, little primary sequence selectivity for most E3 ubiquitin ligases surrounding the target Lys was exhibited [15]. For the remaining types of PTMs, such as glycosylation, which is important in protein folding and stability [24] and acetylation, which influences gene regulation in eukaryotic cells [25], in order to unify the range and the numbers of nsSNVs around the modification sites, we all chose the same criteria with ubiquitylation.

2.3. Association between Damaged PTM Sites and Diseases. nsSNV affected PTM sites are defined as damaged PTMs in this work. Annotations of nsSNVs (deleterious or neutral) were based on the information from the databases mentioned above and on Online Mendelian Inheritance in Man (OMIM; http://www.ncbi.nlm.nih.gov/omim) [26] for reference. Moreover, we identified the elaborate annotated information of nsSNV-related diseases from SwissVar [9] and the explicit matching of nsSNVs with PTM sites was performed. We calculated the association between damaged PTMs and human diseases based on proteins carrying damaged PTM (with SNV related disease annotation-inherited diseases (germline diseases) or cancers (somatic disease)) for each type of PTM, respectively, by hypergeometric test. In our hypergeometric test, the diseases-associated nsSNVs mapped on or around PTM sites were taken as the test dataset, the neutral nsSNVs mapped on or around PTM sites mentioned above were used as control dataset, and the total neutral nsSNVs and the total damaging nsSNVs on proteins containing one specific type of PTM were used as the two background datasets, to find the disease-associated damaged PTM proteins (with damaging SNVs on this type of PTM) (with $P < 0.05$).

2.4. Functional Analysis of Diseases Associated Damaged PTM Sites. To further analyze the functions and features of diseases-related damaged PTMs and their proteins, enrichment analyses were performed using DAVID 6.7 (the database for annotation, visualization, and integrated discovery) [27]. Pathways, biomarkers, and related drugs were analyzed by software Ingenuity Pathway Analysis (IPA) (Ingenuity Systems, http://www.ingenuity.com/). In order to find the structure information of the damaged PTMs, we performed domain enrichment analysis for both inherited disease and cancer-related damaged PTMs based on the domain information from Pfam (version 27.0, released in June, 2012); only the domains containing damaged PTMs were chosen. The enrichment results were calculated and chosen based on disease-related PTM-containing proteins using

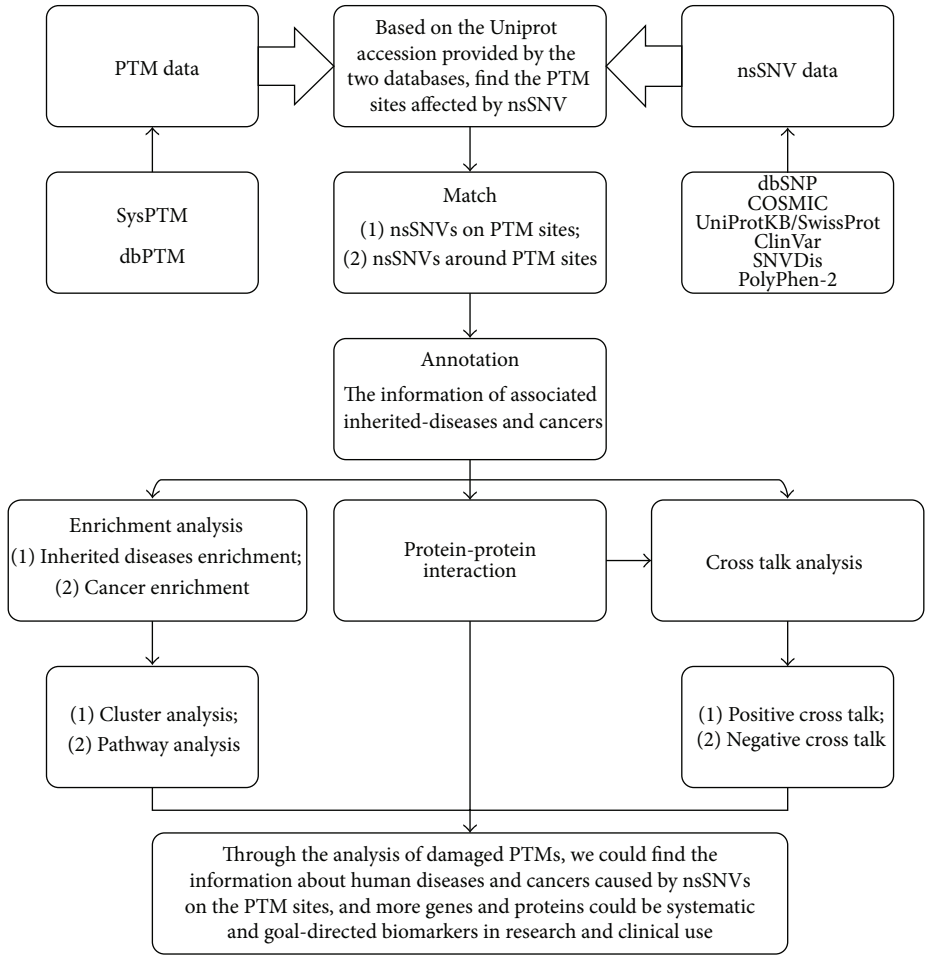


Figure 1: Workflow or protocol for identifying damaged PTMs and associated diseases.

Fisher's exact test and adjusted with Benjamini-Hochberg method (corrected P value < 0.01).

2.5. Cross talks between PTM Types. As for the cross talks between some pairwise types of PTMs, positive and negative cross talks were both considered. Positive cross talk means that one PTM serves as a signal for the addition or removal of a second PTM, or for recognition by a binding protein that carries out a second modification. The negative cross talk could be direct competition for modification of one single residue on a protein, or one modification masks the recognition site of a second PTM [27]. Some positive cross talks can be seen from the pathways or networks they are involved in, based on the physical distance and protein-protein interaction, while negative cross talks can be seen on the same residues where different PTMs compete to occur. Nowadays, more and more information of PTMs have been annotated into protein-protein interaction and associated networks [28], and we mined the cross talks between PTMs based on PTMcode 2 (http://ptmcode.embl.de/) which compiles known and predicated PTM associations [29]. The interaction of the eight damaged PTMs with annotated disease information was illustrated with STRING (http://string-db.org/) [30].

3. Results

The workflow and protocol of this study are shown in Figure 1. We retrieved PTM data and nsSNVs data from the databases mentioned above. Then we matched them to find the PTM sites affected by nsSNVs (the matched results are available in Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/124630); the percentages of the exact matched result out of all eight types of PTMs is shown in Figure 2, and the concrete numbers of nsSNVs on each type of PTM are presented in Table 1.

3.1. The Statistical Relationship between Damaged PTMs and Inherited Diseases and Cancers. We calculated the PTMs affected by inherited disease and cancer-related nsSNVs, respectively, using hypergeometric test and found that phosphorylation affected by nsSNVs was most significantly related to both inherited diseases and cancers. The next is ubiquitylation; however, based on our calculation, it is not significant in inherited diseases, albeit significant in cancers when performing the exact match. The remaining types of PTMs affected by nsSNVs were not significantly associated with inherited diseases. When we expanded to ±2 amino acids around the modified sites, the damaged PTMs significantly

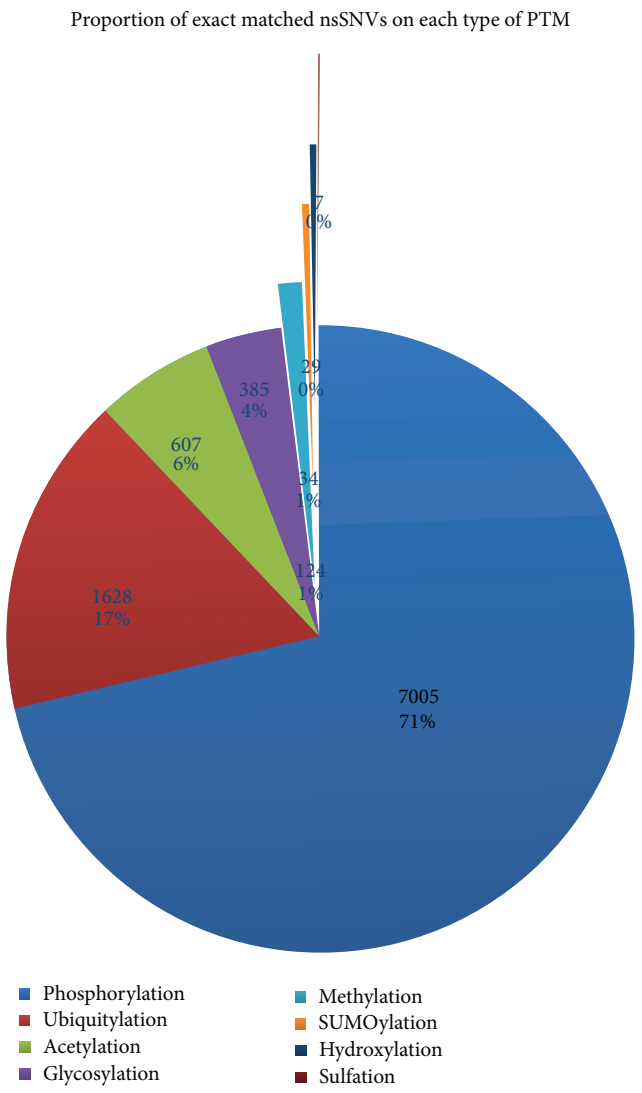


Figure 2: Proportions of exact matched nsSNVs on each PTM out of all sites analyzed. Both the exact number of sites affected and the proportion are shown.

associated with inherited diseases included not only ubiquitylation, but also acetylation and glycosylation. Our results implied that most PTMs affected by nsSNVs were cancer-related, rather than inherited-disease-related (see Tables 1 and 2). This phenomenon might be biased by the data source from big cancer project like The Cancer Genome Atlas (TCGA), Pan-Cancer analysis project [31], and databases like Catalogue of somatic mutations in cancer (COSMIC) [8].

We chose the most frequent modified amino acids, such as Histidine (H), Serine (S), Threonine (T), and Tyrosine (Y) for phosphorylation, Lysine (K) for ubiquitylation, and made a calculation on the frequency of the appearance of nsSNVs on these modified amino acids. We found that the occurring frequency of the modified amino acids affected by nsSNVs was lower compared with their appearance on the whole proteome (data not shown). This demonstrated that the modified amino acids were less affected by mutations.

Table 1: Numbers of nsSNVs on each PTM category.

PTM	# of exact match	# of ±2 match	# of ±7 match
Phosphorylation	7005	33119	78123
Ubiquitylation	1628	10096	—
Acetylation	607	9642	—
Glycosylation	385	2199	—
Methylation	124	427	—
SUMOylation	34	231	—
Hydroxylation	29	328	—
Sulfation	7	26	—

Table 2: Numbers and P values of exact matched nsSNVs related to inherited diseases and cancers on each PTM type.

PTM	Inherited disease	P value*	Cancer	P value
Phosphorylation	313	0.0133	2684	0.0197
Ubiquitylation	59	0.0807	651	0.0172
Acetylation	34	0.0701	233	0.1058
Glycosylation	13	0.2062	57	0.1813
Methylation	15	0.1638	67	0.0912
SUMOylation	1	0.7752	22	0.0152
Hydroxylation	2	0.7423	14	0.3507
Sulfation	4	0.5503	0	0.5503

*P values in this column were calculated using hypergeometric test and all values refer to the left column (genetic disease).

Previous researches showed that PTM sites generally play a key role in normal cellular process like protein-protein interactions and signal transduction and therefore are more stable [15, 32], and our results supported this concept.

Phosphorylation is the best studied and also the most prominent PTM, which has the most abundant data as well [33]. The association between damaged phosphorylation sites and both inherited diseases and cancers is significant, no matter for exact match or for ±2, ±7 amino acids around the phosphorylation sites (Tables 2 and 3). 76736 human phosphorylation sites were obtained in total, out of which only 7005 (9.128%) PTM sites were directly disrupted by nsSNVs. 313 (P value = 0.01331) and 2684 (P value = 0.01974) out of the 7005 damaged phosphorylation sites were inherited-disease-related and cancer-related, respectively. Therefore, phosphorylation affected by nsSNVs was significantly associated with both inherited diseases and cancers (P values < 0.05) (Table 2). For protein kinases, in general, they exhibit a strong selectivity for the primary sequence around the residues they will phosphorylate [33], so ranges of ±2, ±7 residues around the phosphorylated sites were used to find impact by nsSNVs [21] in this study. Ser, Thr, and Tyr can all be phosphorylated; the alterations among these three amino acids can result in diseases, such as S251T in connexin43 (Cx43) protein which is associated with congenital conotruncal anomalies [34] (Table S2, shown in red).

In contrast, ubiquitylation shows little selectivity on primary sequence, such as Lysine, which is highly preferred as the target site of most E3 ubiquitin ligases [15]. So we only chose 2 criteria: exact match and ±2 amino acids

Table 3: Numbers and P values of ±2 AA matched nsSNVs related to inherited diseases and cancer on each PTM.

PTM	Genetic disease	P value*	Cancer	P value
Phosphorylation	1422	$2.51E-04$	12826	0.0111
Ubiquitylation	439	$5.59E-03$	4074	$4.9E-04$
Acetylation	552	$3.93E-07$	4019	$8.21E-03$
Glycosylation	214	0.0261	795	$1.26E-37$
Methylation	44	0.1036	231	$1.57E-02$
SUMOylation	11	0.1526	115	$7.93E-06$
Hydroxylation	22	0.1997	63	$2.82E-03$
Sulfation	7	0.2446	9	0.2526

*P values in this column were calculated using hypergeometric test and all values refer to the left column (genetic disease).

around Lysine. Compared to phosphorylation, the ratio of ubiquitylation sites affected by nsSNVs over total ubiquitination sites (7.22%) found on ubiquitylation was lower (22542 ubiquitylation sites, 5988 proteins). There were 1628 exactly matched nsSNVs found on ubiquitylation proteins, only 59 (3.624%, P value = 0.08067) were inherited disease-associated and 651 (39.98%, P value = 0.01722) were cancer-related sites. For acetylation and glycosylation, both were not found closely related with inherited diseases and cancers (Table 1).

Then, for the remaining four types of PTMs, the numbers of both exact match and ±2 range match were much less than those of the PTMs above, albeit these four types of PTMs are involved in a lot of important cellular processes, and recent works also discovered their related functions and diseases. For instance, SUMOylation proteins are implicated in human diseases including cancers and "Huntington's, Alzheimer's, and Parkinson's diseases"; hydroxylation in Asp110Asn is related with "hemophilia b"; methylation in Arg75Trp is associated with "deafness" [35]; as for sulfation, however, we only identified four mutations in one protein FA8_HUMAN and those were associated with "hemophilia."

Although we found that a lot of damaged PTMs were related with human inherited diseases and cancers, however, almost half of the data remain to be elucidated on their relationships with human diseases. With more damaged PTMs being annotated and analyzed, their impact over health or disease development may become clearer.

3.2. The Damaged PTMs Annotated with Information of Inherited Diseases and Cancers. For all of the eight PTM types studied, we annotated some curated information of diseases based on SwissVar, some annotation information were obtained from the source databases. Although the disease information is up-to-date, the limitation of different databases makes it hard to acquire all the information of known diseases. For instance, inherited-disease-related phosphorylation, "congenital, hereditary, and neonatal diseases and abnormalities," is the most associated disease based on the analysis of SwissVar on exact matched inherited-diseases-related nsSNVs. The next is "skin and connective tissue diseases" and "nervous system diseases." However, "neoplasms" account for the most part of the known diseases in ubiquitylation and acetylation.

In order to acquire more information on related diseases, we performed enrichment analysis of diseases using IPA (Figures 3(a) and 3(b)). We performed both inherited-diseases and cancers enrichment analysis on web tool IPA based on the proteins that carried the damaged PTMs, which were caused by the nsSNVs on or around the modification sites. Through enrichment analysis, we could see that in the exact matched phosphorylation related inherited diseases, "autosomal dominant disease" ($n = 50$, corrected P value = $5.23E-30$,) ranked the first with 50 proteins. For example, PSN1_HUMAN, TNR1A_HUMAN, VHL_HUMA, and PSN1_HUMAN were well studied and associated with "autosomal dominant early-onset Alzheimer's disease" in human [36]. The most significant cancer for the exact matched phosphorylation is "Adenocarcinoma" ($n = 1074$, corrected P value = $4.36E-45$), which ranked the top with 1074 proteins; RASK_HUMAN, P53_HUMAN, EGFR_HUMAN, and so forth were the representative ones. RASK_HUMAN is associated with adenocarcinoma in human large intestine and lung and other tissues. P53_HUMAN is well known for its associations with human colon and rectal and other cancers [37, 38]; for instance, mutation on Ser376 results in the loss of phosphorylation sites, which creates a consensus binding site for 14-3-3 proteins and increases the affinity of p53 for sequence-specific binding sites on DNA [39]. As to ubiquitylation, "Skin abnormality" was the most significant inherited disease ($n = 11$, corrected P value = $3.36E-10$), and two proteins were closely related to it: TSC2_HUMAN and TSC1_HUMAN. They were reported to be associated with tuberous sclerosis syndrome in human [40]. Non-small-cell lung cancer was found significant ($n = 54$, corrected P value = $4.28E-6$) in Ubiquitylation. For acetylation and glycosylation, we also examined both associated inherited diseases and cancers. As to acetylation, we observed disorders of cellular development and cellular growth and proliferation besides cancers that were led by mutations on P53_HUMAN. With regard to glycosylation, the diseases were closely related to lipid metabolism and molecular transport.

We then expanded our search range to the nsSNVs that could affect the PTMs: ±2, ±7 around phosphorylation sites and ±2 for the remaining types of PTMs. First, we chose ±2 range for all the 8 types of PTMs to analyze the associated diseases. For inherited diseases, "autosomal dominant disease" and "autosomal recessive disease" ranked top three in phosphorylation, Ubiquitylation, Acetylation, Glycosylation, Methylation, Hydroxylation, and Sulfation. This was clearly different from the exact matched results. Both autosomal diseases and X-linked hereditary diseases became significant when more nsSNVs were accumulated around PTM sites. The comparison between exact-matched and ±2 range-matched results indicates that (a) mutations on PTMs are rare and, only some certain kinds of inherited diseases were indicated to be caused by them, while more kinds of diseases were indicated to be caused by nsSNVs surrounding PTM sites; (b) human inherited diseases are closely associated with disturbances on and surrounding PTM sites.

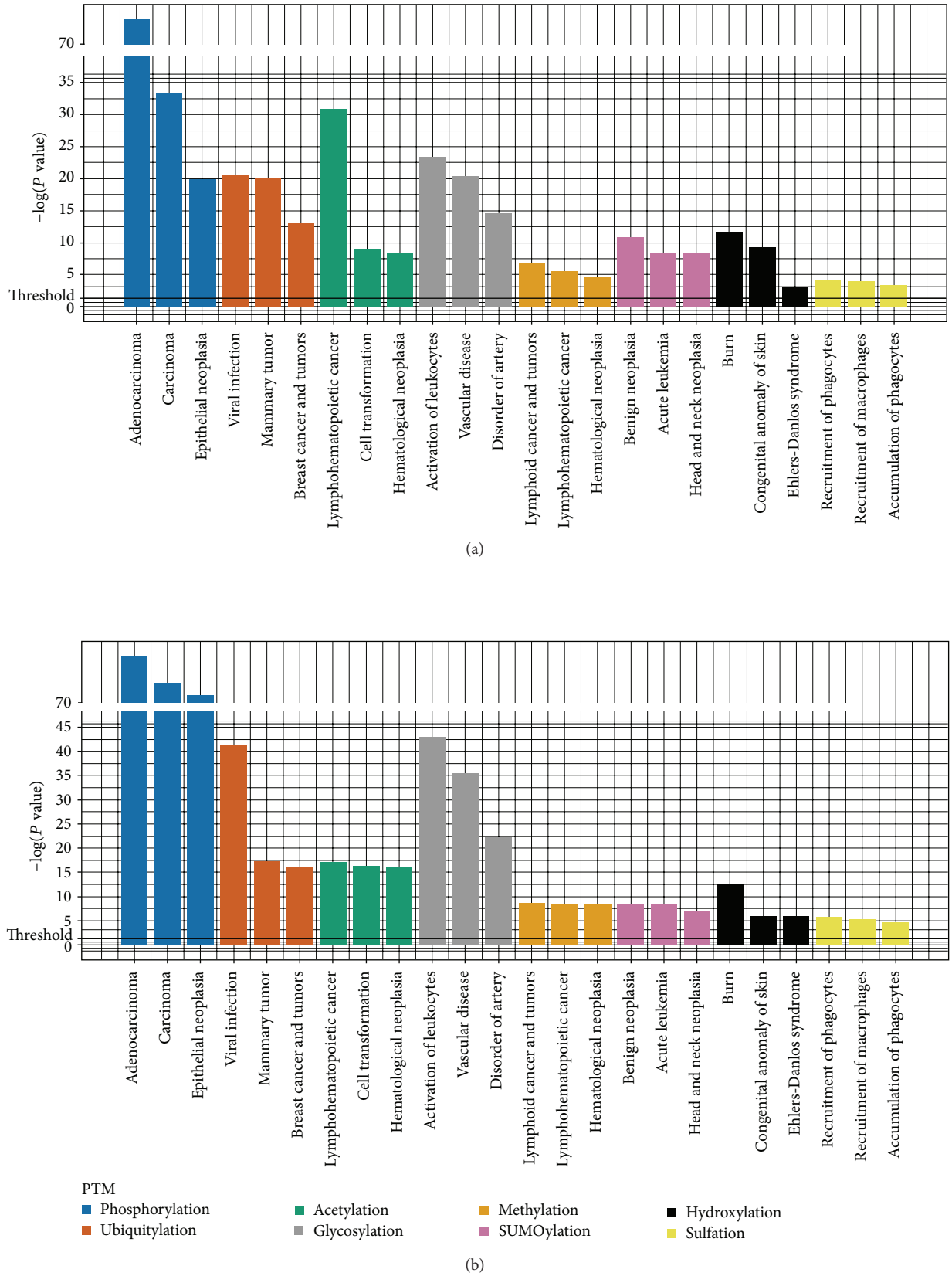


Figure 3: Diseases for each type of damaged PTM affected by nsSNVs in IPA. Threshold was chosen as $P < 0.05$ for all the PTMs and data presented in the charts against −log of P values. Different PTMs are shown in different colors. Both (a) and (b) present nsSNVs on the range of ±2 amino acids around modified residues. (a) Diseases for each PTM affected by inherited-disease-related nsSNVs; (b) diseases for each PTM affected by cancer-related nsSNVs.

Next, we analyzed the ±2 sites range-matched on cancers; the results did not introduce as many changes as exact-matched results. We also compared the data between ±2 and ±7 range around phosphorylation sites; however, their difference was not significant. The differences of human inherited diseases and cancers could be related with the damages of nsSNVs on PTM sites and phenotype: cancers are mostly caused by somatic mutations and present in the current generation; however, the damages of nsSNVs on PTM sites are not easily inherited to the next generation, so the numbers and types of inherited diseases are less compared with damaged-PTM related cancers.

3.3. Functional and Structural Analysis

3.3.1. Enrichment Analysis of Keywords, GO, and Domains. We performed functional enrichment analysis using DAVID. First, we performed keywords and GO association analysis (FDR < 0.01). We still divided data into two parts: exact match and ±2 amino acids (AA) match. "Disease mutation" was the most significant keyword based on the inherited-disease-related nsSNVs that appeared in all the four types of PTMs: Phosphorylation, Ubiquitylation, Acetylation, and Glycosylation. The enrichment analyses showed that the proteins we chose were more likely related to diseases when they encountered mutations. GO enrichment analysis was also performed for the four types of PTMs mentioned above. For each PTM category, the differences of functions among them are obvious (see Table S3). For example, the proteins with phosphorylation mainly involve cell activities like cell death, apoptosis, and signal transduction. Coagulation and wound healing were the GO tags for glycosylation. Through the analyses, we found that the diseases led by the damaged PTMs were closely associated with the role of these proteins played in the regulation of normal cellular processes, which indicated that the damage caused by damaged PTMs was serious.

When we moved to cancer-related nsSNVs on PTMs, the keywords about them had less information about mutations, but rather directing to the function of the proteins. What interested us the most was ubiquitylation; the keywords did not show much about themselves, but other modifications on them. This indicates that ubiquitylation is more likely coexisting with other types of PTMs. Then we examined the GO terms on cancers, besides the functions of the proteins performed, also the chemical characters of them showed up. Like phosphorylation, the most significant GO term about phosphorylation was "protein amino acid phosphorylation" on both exact match and ±2 range match. For the remaining types of PTMs, GO terms more revealed protein roles on different processes; for example, "modification-dependent protein catabolic process" ranked in the top two on both range criteria of ubiquitylation.

Then we examined the damaged PTMs associated domains based on the data from Pfam to analyze the impact of damaged PTMs on protein structures. For damaged phosphorylation, "protein tyrosine kinase" ($n = 13$, corrected P value $= 2.66E-8$) and "protein kinase domain" ($n = 81$, corrected P value $= 2.03E-14$) ranked the first in human inherited diseases and cancers, respectively. The damaged phosphorylation on the kinases could result in damage to another phosphorylation and thus nsSNVs do not affect only one phosphorylation site. Then, in terms of ubiquitylation, "P53 DNA-binding domain" ($n = 8$, corrected P value $= 5.24E-11$) and "Histone" ($n = 11$, corrected P value $= 5.73E-4$) were the most significant domains. On P53_HUMAN, lots of phosphorylation and ubiquitylation sites coexisted and some of them affected the same domains, such as "P53 DNA-binding domain." "Connexin" ($n = 6$, corrected P value $= 7.06E-7$) and "HMG14 and HMG17" ($n = 18$, corrected P value $= 6.59E-8$) were the domains damaged acetylation was enriched in. Glycosylation was involved in wound healing, cell-adhesion, and cellular proliferation and we found that "immunoglobulin domain" ($n = 8$, corrected P value $= 0.042$) and "class I histocompatibility antigen, domains alpha 1 and 2" ($n = 25$, corrected P value $= 3.97E-4$) were enriched in glycosylation domains. Also for Hydroxylation, "collagen triple helix repeat (20 copies)" ($n = 6$, corrected P value $= 1.06E-9$) was found in cancer-related dataset. For other types of PTMs, the domains were scattered compared with PTMs mentioned above. From the data of associated domains, we found that the damaged PTMs associated domains were closely related to molecular binding and protein-protein interactions, which was a major function of PTMs [15].

3.3.2. Pathway Analysis. In order to investigate the function of damaged PTMs in proteome-wide scale, we performed pathway analysis by IPA (details available in Table S4). In IPA analysis for inherited-disease associated damaged PTMs of the exact matched data, some pathways are significant: "ovarian cancer signaling" in Phosphorylation (corrected P value $= 2.17E-12$, ratio $= 0.131$), Ubiquitylation (corrected P value $= 3.21E-5$, ratio $= 0.046$), and Acetylation (corrected P value $= 2.63E-3$, ratio $= 0.031$); "hereditary breast cancer signaling" in Phosphorylation (corrected P value $= 4.8E-9$, ratio $= 0.116$), Ubiquitylation (corrected P value $= 6.94E-7$, ratio $= 0.062$), Acetylation (corrected P value $= 2.36E-3$, ratio $= 0.036$), and Methylation (corrected P value $= 7.69E-4$, ratio $= 0.027$); "Role of BRAC1 in DNA damage response" in Phosphorylation (corrected P value $= 1.34E-9$, ratio $= 0.18$), Ubiquitylation (corrected P value $= 4.72E-4$, ratio $= 0.066$), Acetylation (corrected P value $= 4.4E-3$, ratio $= 0.049$), and Methylation (corrected P value $= 6.43E-3$, ratio $= 0.033$). In these pathways, some are associated with their functions like "Coagulation system" (corrected P value $= 7.75E-10$, ratio $= 0.171$) in glycosylation. As for cancers, we examined each type of PTM category and found that the pathways were more associated with their functions of the proteins, for instance, "protein kinase A signaling" (corrected P value $= 2.52E-16$, ratio $= 0.269$) in Phosphorylation, "protein ubiquitylation pathway" (corrected P value $= 9.03E-11$, ratio $= 0.134$) in Ubiquitylation; we found that more cancer-related damaged PTMs were associated with signaling pathways and this indicated that somatic mutations could affect normal cellular processes more often and may thus result in human cancers.

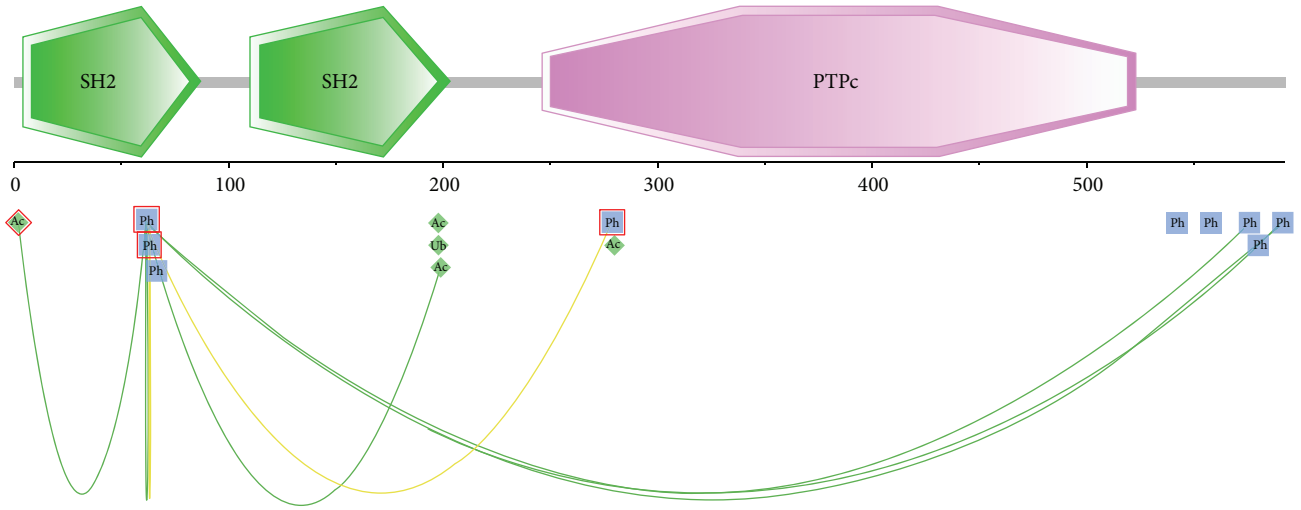


Figure 4: The cross talk of disease-related phosphorylation site Y62 with other PTM sites in protein PTN11_HUMAN. The two "SH2" and one "PTPc" boxed in green and pink are domains in the protein; green lines and yellow lines show the association between PTM sites based on evidence of coevolution and physical distance, respectively. Disease-related PTM sites are boxed in red.

3.3.3. Protein-Protein Interaction Analysis. On the proteome-wide range, the associations among these proteins were close, and we illustrated the interactions using networks of protein-protein interactions with STRING (Figure 5). With a total of 159 proteins which carried identified damaged PTM sites with SwissVar annotated information, we manually divided the associated proteins of different types of PTMs into six major parts, while Sulfation and SUMOylation were not shown for the limited number of data. Not only did some proteins carry one kind of PTMs, such as KRAS, MRE11A, but also phosphorylation, ubiquitylation, and acetylation coexisted on these proteins. From this network, we found that, except for phosphorylation, the interactions among one kind of PTMs were less compared with their interactions with phosphorylation. This result showed us that phosphorylation which was the hub of signal transduction with a strong relationship with other types of PTMs played a key role in the association between damaged PTMs and human inherited diseases and cancers. For example, PTPN11, which was found carrying damaged acetylation caused by (T2I) associated with "noonan syndrome 1" [41], was involved in downstream effectors of cytoplasmic protein tyrosine kinases.

3.3.4. Cross talk Analysis. Cross talk between some paired PTMs of different types such as phosphorylation and ubiquitylation and ubiquitylation and acetylation, has become a study theme on proteomics [15, 16]. It shows that the extensive use of PTMs to generate multiple distinct protein states from a single gene product could compensate for the relative paucity of genes in vertebrate genomes [15]. In this work, we investigated the impact of nsSNVs on cross talks between some pairwise PTMs. Cross talks of PTMs can be defined as positive and negative; both mean one PTM has an impact on the other PTM [15]. In this study, we mined the information of cross talks based on PTMcode [29]. Most of the PTM sites have cross talks with other PTM sites based on some evidences such as coevolution and physical distance. Here, we took PTN11_HUMAN as an example for the cross talk within one protein, which totally carried 23 PTMs with 55 functional associations. In our inherited-disease-related dataset, 4 nsSNVs occurred on phosphorylation sites (T2I, Y62D, Y63C, ad Y279C) and 1 on acetylation site of PTN11_HUMAN(Y279S) (Figure 4). The mutations on Y279 are associated with "human LEOPARD syndrome 1" [42], and the mutations on the remaining sites are associated with "human Noonan syndrome 1" [41, 43]; also, within this protein, T2 is associated with both Y62 and Y63, which are all found changed in "Noonan syndrome 1" [41]. Thus, the association of the damaged PTMs could play a key role in the development of human inherited diseases.

On the proteome-wide range, the associations were more prevalent. Then we took P53_HUMAN and TOP1_HUMAN as examples for the cross talks between different PTM sites on distinct proteins: on P53_HUMAN, we found 21 phosphorylation sites, 14 ubiquitylation sites, and 9 acetylation sites; among them, the associations were prevalent within the protein, and the damaged PTMs mostly resulted in the deficiency in the role it played in significant cellular functions [44]; K326R on TOP1_HUMAN is related to human breast cancer [45], and the protein-protein interaction between them is among 159 proteins (Figure 5, boxed in brown); we found that the ubiquitylation on K326 was associated with 33 PTMs in protein P53 (Figure 6); 18 phosphorylation sites were among our inherited disease-related dataset. From the cross talks among these PTMs, we could infer that not only the nsSNVs on one PTM site affect that site, but also other associated sites could be affected. For instance, O-GlcNacylation of S149 in p53 reduces phosphorylation of T155 [15]. Not only human inherited diseases, but also cancers are related to these damaged PTMs.

For the negative cross talk, where more than one kind of PTMs could happen on the same residue, could be occurred

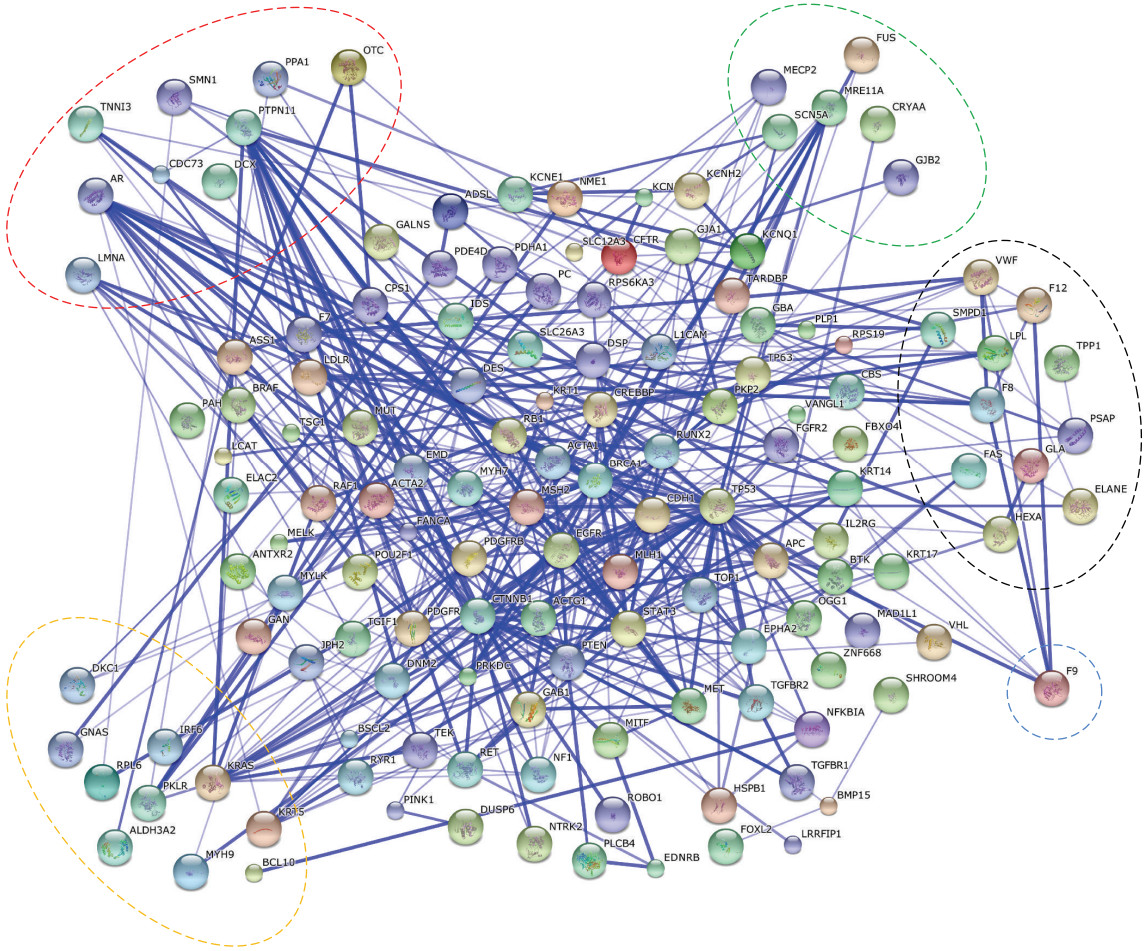


FIGURE 5: Network of protein-protein interactions among the proteins carrying inherited-disease or cancers related damaged PTMs identified by SwissVar. The proteins were divided into six parts; each category was circled by different colors except for phosphorylation in the center: red represented acetylation, green represented methylation, black represented glycosylation, blue represented hydroxylation and yellow represented ubiquitylation. Stronger associations were represented by thicker lines.

in different stage of cellular processes or on different positions. We chose three pairwise PTMs to perform the analysis: phosphorylation and ubiquitylation, phosphorylation and acetylation, and ubiquitylation and acetylation. For the first and second group, phosphorylation and ubiquitylation, and phosphorylation and acetylation, the exact match sites were not overlapped, but when we used damaged ubiquitylation and acetylation sites to match with ±7 sites around phosphorylational sites, we obtained 12 overlapping sites and 10 overlapping sites, respectively, for ubiquitylation and acetylation, and, among them, 7 and 5 sites were on P53_HUMAN, respectively. For example, K320 on TP53 could be ubiquitylated or acetylated (Figure 6). Then we examined the group concerning ubiquitylation and acetylation; we matched their exact sites and obtained 13 overlapping sites. For example, both ubiquitylation and acetylation were detected on K97; nsSNVs on this site could result in "cardiomyopathy, dilated 1a" [46]. Positive cross talk, in which one PTM promotes or prevents another PTM directly on the same site or indirectly on other sites, extends the impact of nsSNVs on PTMs, thus increasing the chance of development of human inherited diseases and cancers in wider ranges. Negative crosstalk with distinct PTMs competing the same site could render nsSNVs on these sites damages to the normal function of all these PTMs, to result in the damages to the related protein functions.

3.4. Potential of Damaged PTMs as Biomarkers in Inherited Diseases and Cancers. The damaged PTMs may cause protein functions to be out of control in canonical pathways [47]. For research and medical use, some of them might be very good biomarker candidates [48], which could be used as the drug targets for intervention. We found some proteins with damaged PTMs among the canonical pathways that could be most likely regarded as biomarker candidates using information from IPA. For the exact matched phosphorylation sites with nsSNVs, we filtered 481 gene/proteins; several of them had already been used as the targets of some drugs, but plenty of them still remained to be explored as targets of new drugs (more details available in Table S5). We further identified 169 filtered proteins for ubiquitylation and 90 filtered proteins for acetylation (Table S5). Proteins

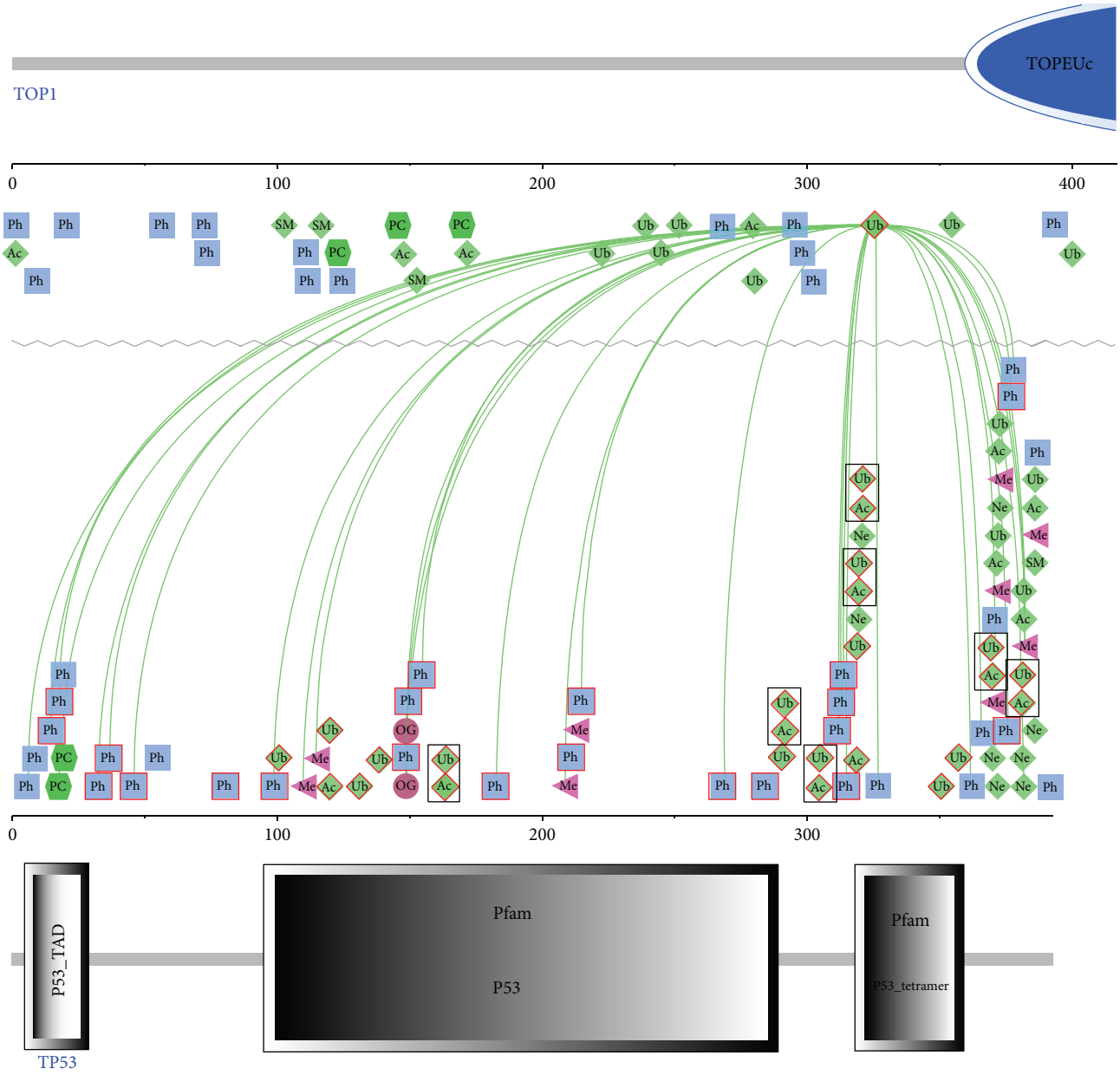


Figure 6: The cross talks between the ubiquitylation site K326 of protein TOP1 with other PTM sites on TP53. Green lines show the association of K326 with other PTM sites based on the evidence of coevolution. Some domains on the two proteins are also given, largely boxed in blue and grey. The different PTMs boxed in red show disease-related PTM sites and those with more than one kind of PTM on the same residue were boxed in black.

carrying damaged PTMs are usually associated with lots of critical signaling pathways during the development of diseases [49], such as VHL, which were von Hippel-Lindau tumor suppressor, E3 ubiquitin protein ligase, which was involved in cardiovascular disease, hematological disease, and other diseases. Some of the candidate biomarkers are functionally similar to the known proteins in clinical use. MRP1_HUMAN, which belonged to the family of ABCC1, has been recognized as a biomarker in breast cancer and other cellular disorders [49], with drugs like "sulfinpyrazone." For each PTM, we provided some most likely biomarkers as candidates (Table S5).

4. Conclusions

In summary, through this work, we investigated the associations between PTMs affected by nsSNVs and human inherited diseases and cancers from diverse perspectives such as functions, pathways, and cross talks. These provided us a proteome-wide view of how the proteins, which carry modifications and nsSNVs, play roles in the development of diseases and cancers. Not only do PTMs play key roles in almost every important cellular process, but also their dysfunction could result in human diseases. We provided a practical protocol to analyze disease-related proteins that

carry damaged PTMs; some valuable proteins were listed out as the candidate biomarkers for potential research and clinical use. However, still almost half of damaged PTMs did not demonstrate associations with human health based on our current analysis, and their functions need to be revealed. Moreover, what we need to do in the future is to identify the causative relationships between the damaged PTMs and human diseases, by discovering key nsSNVs on protein modifications.

Abbreviations

PTM: Protein posttranslational modification
nsSNVs: Nonsynonymous single-nucleotide variations
GO: Gene Ontology
TCGA: The Cancer Genome Atlas
CCND1: Cyclin D1
AA: Amino acid.

Conflict of Interests

The authors confirm that this paper's content has no conflict of interests.

Acknowledgments

This work was funded by National Hi-Tech Program (2012AA020201); Key Infectious Disease Project (2012ZX10002012-014); National Key Basic Research Program (2010CB912702, 2011CB910204).

References

[1] J. G. Tooley and C. E. Schaner Tooley, "New roles for old modifications: emerging roles of N-terminal post-translational modifications in development and disease," *Protein Science*, vol. 23, no. 12, pp. 1641–1649, 2014.

[2] J. Seo and K.-J. Lee, "Post-translational modifications and their biological functions: proteomic analysis and systematic approaches," *Journal of Biochemistry and Molecular Biology*, vol. 37, no. 1, pp. 35–44, 2004.

[3] K. Haglund and I. Dikic, "Ubiquitylation and cell signaling," *The EMBO Journal*, vol. 24, no. 19, pp. 3353–3359, 2005.

[4] J. Nakayama, J. C. Rice, B. D. Strahl, C. D. Allis, and S. I. S. Grewal, "Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly," *Science*, vol. 292, no. 5514, pp. 110–113, 2001.

[5] S. T. Sherry, M. Ward, and K. Sirotkin, "dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation," *Genome Research*, vol. 9, no. 8, pp. 677–679, 1999.

[6] K. Karagiannis, V. Simonyan, and R. Mazumder, "SNVDis: a proteome-wide analysis service for evaluating nsSNVs in protein functional sites and pathways," *Genomics, Proteomics and Bioinformatics*, vol. 11, no. 2, pp. 122–126, 2013.

[7] M. J. Landrum, J. M. Lee, G. R. Riley et al., "ClinVar: public archive of relationships among sequence variation and human phenotype," *Nucleic Acids Research*, vol. 42, no. 1, pp. D980–D985, 2014.

[8] S. Bamford, E. Dawson, S. Forbes et al., "The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website," *British Journal of Cancer*, vol. 91, no. 2, pp. 355–358, 2004.

[9] A. Mottaz, F. P. A. David, A.-L. Veuthey, and Y. L. Yip, "Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar," *Bioinformatics*, vol. 26, no. 6, pp. 851–852, 2010.

[10] C. Greenman, P. Stephens, R. Smith et al., "Patterns of somatic mutation in human cancer genomes," *Nature*, vol. 446, no. 7132, pp. 153–158, 2007.

[11] C. Cole, K. Krampis, K. Karagiannis et al., "Non-synonymous variations in cancer and their effects on the human proteome: workflow for NGS data biocuration and proteome-wide analysis of TCGA data," *BMC Bioinformatics*, vol. 15, no. 1, article 28, 2014.

[12] P. Radivojac, P. H. Baenziger, M. G. Kann, M. E. Mort, M. W. Hahn, and S. D. Mooney, "Gain and loss of phosphorylation sites in human cancer," *Bioinformatics*, vol. 24, no. 16, pp. i241–i247, 2008.

[13] E. Manié, A. Vincent-Salomon, J. Lehmann-Che et al., "High frequency of TP53 mutation in BRCA1 and sporadic basal-like carcinomas but not in BRCA1 luminal breast tumors," *Cancer Research*, vol. 69, no. 2, pp. 663–671, 2009.

[14] S. Benzeno, F. Lu, M. Guo et al., "Identification of mutations that disrupt phosphorylation-dependent nuclear export of cyclin D1," *Oncogene*, vol. 25, no. 47, pp. 6291–6303, 2006.

[15] T. Hunter, "The age of crosstalk: phosphorylation, ubiquitination, and beyond," *Molecular Cell*, vol. 28, no. 5, pp. 730–738, 2007.

[16] J.-S. Lee, E. Smith, and A. Shilatifard, "The language of histone crosstalk," *Cell*, vol. 142, no. 5, pp. 682–685, 2010.

[17] J. Li, J. Jia, H. Li et al., "SysPTM 2.0: an updated systematic resource for post-translational modification," *Database*, vol. 2014, p. bau025, 2014.

[18] C. T. Lu, K. Y. Huang, M. G. Su et al., "DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications," *Nucleic Acids Research*, vol. 41, no. 1, pp. D295–D305, 2013.

[19] M. Magrane and U. P. Consortium, "UniProt Knowledgebase: a hub of integrated protein data," *Database*, vol. 2011, Article ID bar009, 2011.

[20] I. A. Adzhubei, S. Schmidt, L. Peshkin et al., "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[21] J. Reimand, O. Wagih, and G. D. Bader, "The mutational landscape of phosphorylation signaling in cancer," *Scientific Reports*, vol. 3, article 2651, 2013.

[22] J. D. Graves and E. G. Krebs, "Protein phosphorylation and signal transduction," *Pharmacology and Therapeutics*, vol. 82, no. 2-3, pp. 111–121, 1999.

[23] P. Beltrao, P. Bork, N. J. Krogan, and V. van Noort, "Evolution and functional cross-talk of protein post-translational modifications," *Molecular Systems Biology*, vol. 9, article 714, 2013.

[24] M. M. Chen, A. I. Bartlett, P. S. Nerenberg et al., "Perturbing the folding energy landscape of the bacterial immunity protein Im7 by site-specific N-linked glycosylation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 52, pp. 22528–22533, 2010.

[25] L. Verdone, E. Agricola, M. Caserta, and E. di Mauro, "Histone acetylation in gene regulation," *Briefings in Functional Genomics & Proteomics*, vol. 5, no. 3, pp. 209–221, 2006.

[26] J. Amberger, C. Bocchini, and A. Hamosh, "A new face and new challenges for Online Mendelian Inheritance in Man (OMIM)," *Human Mutation*, vol. 32, no. 5, pp. 564–567, 2011.

[27] X. Jiao, B. T. Sherman, D. W. Huang et al., "DAVID-WS: a stateful web service to facilitate gene/protein list analysis," *Bioinformatics*, vol. 28, no. 13, pp. 1805–1806, 2012.

[28] G. Duan and D. Walther, "The roles of post-translational modifications in the context of protein interaction networks," *PLoS Computational Biology*, vol. 11, no. 2, Article ID e1004049, 2015.

[29] P. Minguez, I. Letunic, L. Parca, and P. Bork, "PTMcode: a database of known and predicted functional associations between post-translational modifications in proteins," *Nucleic Acids Research*, vol. 41, no. 1, pp. D306–D311, 2013.

[30] D. Szklarczyk, A. Franceschini, S. Wyder et al., "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, pp. D447–D452, 2015.

[31] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson et al., "The Cancer Genome Atlas Pan-Cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[32] P. Radivojac, P. H. Baenziger, M. G. Kann, M. E. Mort, M. W. Hahn, and S. D. Mooney, "Gain and loss of phosphorylation sites in human cancer," *Bioinformatics*, vol. 24, no. 16, pp. I241–I247, 2008.

[33] G. A. Khoury, R. C. Baliban, and C. A. Floudas, "Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database," *Scientific Reports*, vol. 1, article 90, 2011.

[34] P. Chen, L.-J. Xie, G.-Y. Huang, X.-Q. Zhao, and C. Chang, "Mutations of connexin43 in fetuses with congenital heart malformations," *Chinese Medical Journal*, vol. 118, no. 12, pp. 971–976, 2005.

[35] G. Richard, T. W. White, L. E. Smith et al., "Functional defects of Cx26 resulting from a heterozygous missense mutation in a family with dominant deaf-mutism and palmoplantar keratoderma," *Human Genetics*, vol. 103, no. 4, pp. 393–399, 1998.

[36] D. Campion, C. Dumanchin, D. Hannequin et al., "Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum," *The American Journal of Human Genetics*, vol. 65, no. 3, pp. 664–670, 1999.

[37] B. Dix, P. Robbins, S. Carrello, A. House, and B. Iacopetta, "Comparison of p53 gene mutation and protein overexpression in colorectal carcinomas," *British Journal of Cancer*, vol. 70, no. 4, pp. 585–590, 1994.

[38] The Cancer Genome Atlas Network, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, vol. 487, no. 7407, pp. 330–337, 2012.

[39] M. F. Lavin and N. Gueven, "The complexity of p53 stabilization and activation," *Cell Death and Differentiation*, vol. 13, no. 6, pp. 941–950, 2006.

[40] M. van Slegtenhorst, R. de Hoogt, C. Hermans et al., "Identification of the tuberous sclerosis gene TSC1 on chromosome 9q34," *Science*, vol. 277, no. 5327, pp. 805–808, 1997.

[41] A. Sarkozy, E. Conti, D. Seripa et al., "Correlation between PTPN11 gene mutations and congenital heart defects in Noonan and LEOPARD syndromes," *Journal of Medical Genetics*, vol. 40, no. 9, pp. 704–708, 2003.

[42] B. Keren, A. Hadchouel, S. Saba et al., "PTPN11 mutations in patients with LEOPARD syndrome: a French multicentric experience," *Journal of medical genetics*, vol. 41, no. 11, article e117, 2004.

[43] M. Tartaglia, K. Kalidas, A. Shaw et al., "PTPN11 mutations in noonan syndrome: molecular spectrum, genotype-phenotype correlation, and phenotypic heterogeneity," *The American Journal of Human Genetics*, vol. 70, no. 6, pp. 1555–1563, 2002.

[44] J. Rutherford, C. E. Chu, P. M. Duddy et al., "Investigations on a clinically and functionally unusual and novel germline p53 mutation," *British Journal of Cancer*, vol. 86, no. 10, pp. 1592–1596, 2002.

[45] T. Sjöblom, S. Jones, L. D. Wood et al., "The consensus coding sequences of human breast and colorectal cancers," *Science*, vol. 314, no. 5797, pp. 268–274, 2006.

[46] E. Arbustini, A. Pilotto, A. Repetto et al., "Autosomal dominant dilated cardiomyopathy with atrioventricular block: a lamin A/C defect-related disease," *Journal of the American College of Cardiology*, vol. 39, no. 6, pp. 981–990, 2002.

[47] J. V. Olsen, B. Blagoev, F. Gnad et al., "Global, in vivo, and site-specific phosphorylation dynamics in signaling networks," *Cell*, vol. 127, no. 3, pp. 635–648, 2006.

[48] N. Rifai, M. A. Gillette, and S. A. Carr, "Protein biomarker discovery and validation: the long and uncertain path to clinical utility," *Nature Biotechnology*, vol. 24, no. 8, pp. 971–983, 2006.

[49] J. Zhang, M. J. Guy, H. S. Norman et al., "Top-down quantitative proteomics identified phosphorylation of cardiac troponin I as a candidate biomarker for chronic heart failure," *Journal of Proteome Research*, vol. 10, no. 9, pp. 4054–4065, 2011.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 141363, 11 pages
http://dx.doi.org/10.1155/2015/141363

Research Article

KIR Genes and Patterns Given by the A Priori Algorithm: Immunity for Haematological Malignancies

J. Gilberto Rodríguez-Escobedo,[1] Christian A. García-Sepúlveda,[2] and Juan C. Cuevas-Tello[1]

[1]*Facultad de Ingeniería, Universidad Autónoma de San Luis Potosí, Avenida Dr. Manuel Nava No. 8, Zona Universitaria, 78290 San Luis Potosí, ZC, Mexico*
[2]*Laboratorio de Genómica Viral y Humana, Facultad de Medicina, Universidad Autónoma de San Luis Potosí, Avenida Venustiano Carranza No. 2405, Colonia Filtros las Lomas, 78210 San Luis Potosí, CP, Mexico*

Correspondence should be addressed to Juan C. Cuevas-Tello; cuevastello@gmail.com

Received 27 May 2015; Revised 5 August 2015; Accepted 9 August 2015

Academic Editor: Lei Chen

Killer-cell immunoglobulin-like receptors (KIRs) are membrane proteins expressed by cells of innate and adaptive immunity. The KIR system consists of 17 genes and 614 alleles arranged into different haplotypes. KIR genes modulate susceptibility to haematological malignancies, viral infections, and autoimmune diseases. Molecular epidemiology studies rely on traditional statistical methods to identify associations between KIR genes and disease. We have previously described our results by applying support vector machines to identify associations between KIR genes and disease. However, rules specifying which haplotypes are associated with greater susceptibility to malignancies are lacking. Here we present the results of our investigation into the rules governing haematological malignancy susceptibility. We have studied the different haplotypic combinations of 17 KIR genes in 300 healthy individuals and 43 patients with haematological malignancies (25 with leukaemia and 18 with lymphomas). We compare two machine learning algorithms against traditional statistical analysis and show that the "a priori" algorithm is capable of discovering patterns unrevealed by previous algorithms and statistical approaches.

1. Introduction

One goal in systems biology, along with functional genomic (Human Genome Project) analysis and physiology (Human Physiome Project), is to provide personalized medicine in a practical, clinically useful way. The digital genome and environmental signals are two fundamental types of biological information that dictate whether an individual adopts a normal or diseased phenotype. Therefore, functional genomics data can help diagnose disease and guide therapy [1].

Several cancer research initiatives employing genomic information focus mainly on DNA microarray data in the search for biomarkers using tens of thousands of genetic polymorphisms [2]. However, after recent discoveries relating to KRAS gene mutations in cancer patients, novel research strategies are focusing on circulating tumour DNA (ctDNA) and to the way that it might allow for a closer surveillance of the clinical evolution of cancer in certain types of patients [3]. Several diseases have been studied in systems biology; this paper focuses on haematological malignancies (leukaemia and lymphomas). Contrary to DNA microarray data and ctDNA, this paper studies the impact of specific innate immunity genes with disease occurrence or protection.

Traditionally, hypothesis driven approaches based on current knowledge have been used to uncover associations between a small number of genetic traits and disease occurrence or disease progression. Genome-wide analysis studies (GWAS) have rapidly become powerful tools for the analysis of tens of thousands and sometimes millions of genetic markers and of their association with complex diseases. In the last 15 years, several GWAS have demonstrated the importance that immune and nonimmune gene polymorphisms have at determining an individual's capability to mount an immune response against infectious pathogens, residual leukaemia,

antileukaemia drug metabolism, and haemopoietic stem cell transplantation (HSCT) outcome. However, only a few studies have addressed the importance of analysing the full context of innate immunity genes and of their interplay with the adaptive immune system with regards to leukaemias and lymphomas. In more recent years network-assisted analysis (NAA) of GWAS data has demonstrated enormous power for the study of various human diseases or traits [4–7].

A small subset of CD8 lymphocytes and Natural Killer (NK) cells are represented by the Killer-Cell Immunoglobulin-like receptors (KIR), and they are key participants of immune responses to tumours. KIR genes, in comparison to genes of the adaptive immune system, are genetically predetermined and remain unchanged throughout life [8, 9]. Nowadays, 17 KIR genes have been discovered, which exhibit allelic polymorphism [10], forming a cluster in the locus 19q13.4. The KIR genes are physically contiguous strings, known as haplotypes [11, 12]. The variability in KIR genotype is such that most pairs of unrelated human individuals have different KIR genotypes, so the unique feature of the human KIR system is the representation of two distinctive groups of haplotypes (A and B), and many haplotypes having presence and absence of genes and variants are known [13]. A KIR haplotype is composed of two motifs, centromeric and telomeric. The KIR haplotype motifs are cA01, cB01, cB02, cB03, tA01, and tB01 [11, 12]. The KIR haplotypes of the great majority of individuals contain the four framework genes KIR3DL3, KIR3DP1, KIR2DL4, and KIR3DL2 [11, 14].

KIR genes encode for two (2D) or three (3D) extracellular domain membrane bound proteins capable of transducing activating (S) or inhibitory (L) signals on binding of their cognate ligands. It is the balance and integration of these signals that modulates NK cell cytotoxicity and cytokine release. The haplotypes of group A are more important because they have simple and constant gene content, dominated by inhibitory genes (L). On the other hand, haplotypes of group B have variable and greater gene content, involving both inhibitory and activating receptors [11]. NK cells were initially identified by their ability to spontaneously kill tumour cells without prior sensitisation [15–17]. Historical studies of the immunogenetic factors that determine clinical outcome in patients subjected to HSCT for haematological malignancies were the first to highlight the clinical relevance of KIR genes in antitumour responses [18].

The first study to suggest such an association described a potent graft-versus-leukaemia effect arising from predicted NK cell alloreactivity in the Graft-versus-Host direction amongst patients subjected to HSCT for leukaemias [19]. Many other studies published since then have described KIR gene associations with antitumour effects and posttransplant clinical endpoints [20–26]. In addition, NK cell antitumour activity has been demonstrated *in vitro* against a wide variety of haematological malignancies [18, 27]. In all, these findings support the notion that KIRs allow NK cells to play an important role at determining susceptibility to certain haematological tumours [28–30].

Previous findings based on our data employing multivariate analysis of KIR carrier frequencies with a traditional statistical comparison (contingency tables using Pearson's or Fishers' exact test [31]) revealed only that KIR2DL2 was more frequent amongst patients with haematological malignancy in comparison to the healthy donors ($p \leq 0.0001$). Decision trees (ID3 algorithm [32]) generated at 50% and 75% training data also provided support the importance of KIR2DL2 [33]. Other findings produced with the ID3 algorithm on our similar data suggest a protective effect for (i) cB03 motif (KIR2DL3, KIR2DL5, KIR2DS5, KIR2DP1, and KIR2DL1 genes) in agreement with KIR3DS1-2DL5-2DS5-2DS1 genotype with protection from Hodgkin's lymphoma [34]; (ii) KIR3DS1 gene (only provided a protective effect when observed in the absence of KIR2DL2 or KIR2DL5 genes) as suggested previously [25, 34, 35]; and (iii) KIR2DS1 when present together with KIR2DL2, KIR2DS2, and KIR2DL3 but in the absence of KIR3DL1 [33].

Nevertheless, the ID3 algorithm failed to find associations related to the KIR2DS3, as described previously by others researchers [35–37]. Neither KIR2DL1 nor KIR2DL3 are on their own important factors in the ID3 decision processes [33]. One reason is that the ID3 algorithm is based only on entropy of information, which could not identify other patterns with this measure of information. Genes KIR2DL1, KIR2DL3, KIR2DL5, and KIR2DS3/S5 were also present in our patients in haplotype motifs other than the classic cA01 (or KIR2DL1 and -2DL3) and cB01 (for the KIR2DL1, KIR2DL5, KIR2DS3, and KIR2DS5), as suggested for certain Hodgkin's lymphomas [38]. Differences in patient demographics, clinical management, KIR typing method, and the preferred transplant modality have largely contributed to the heterogeneity of the KIR gene associations that have been described across the literature.

In this paper, we further study the *a priori* algorithm on the same dataset in an effort to discover novel associations not identified by the ID3 algorithm. The *a priori* algorithm is an algorithm that belongs to the family of data mining algorithms in the field of machine learning and artificial intelligence [39–41]. Regarding classification algorithms, previous research has already described the potential that support vector machines (SVM) have [33], as well as that of other state-of-the-art classification algorithms including Deep Neural Networks and Convolutional Neural Networks [42]. Moreover, research on classification algorithms is also focusing on creating an ensemble of classifiers such as LibD3C [43]. However, these algorithms are deficient at finding association rules and defining them, so more research is needed. As our work with KIR and haematological malignancies represents an imbalanced classification problem [44], the a priori algorithm was considered as an interesting and informative approach for work with this dataset. The main contributions of this paper are (i) we follow a data mining methodology to study associations between KIR genes and disease; (ii) the novel application of the a priori algorithm to identify associations between KIR genes and haematological malignancies; (iii) we found novel associations not detected before by the ID3 algorithm (see Section 3) (iv) we apply an improved version of the ID3 algorithm, known as J48, so one can validate that the results of the a priori algorithm are novel.

Table 1: Clinical data for the haematological cohort.

	n	%
Gender		
Male	23	53
Female	20	46
Diagnosis		
Chronic myeloid leukaemia	25	58
Hodgkin's lymphoma	18	42
B symptoms		
Present	30	70
Absent	13	30
ECOG[a]		
0	3	7
1	16	37
2	20	46
3	3	7
4	1	2

[a]Eastern Cooperative Oncology Group (ECOG).

2. Materials and Methods

2.1. Study Population. Samples belonging to the Mexican Reference Genomic DNA Collection (MGDC-REF), which includes 300 unrelated blood donors, were used as healthy controls for this study. This Mexican mestizo reference population included 135 (45%) males and 165 (55%) females aged between 19 and 38 years (median of 24) of which 75% were residents of the city of San Luis Potosí and 25% were residents of rural areas of this Mexican state. These DNA samples were extracted from blood-bank discarded leukocyte concentrates referred to us by Hospital Central "Dr. Ignacio Morones Prieto" according to previously published protocols [45]. A more detailed description of the KIR features present in this reference population is given in the original publication [46]. In addition, 43 DNA samples obtained from patients with haematological malignancies (25 with leukaemia and 18 with lymphomas) referred to us by the Haematology Department of Hospital Central "Dr. Ignacio Morones Prieto" were included as representatives of a diseased study group. More information for the haematological cohort is given in Table 1. All samples were provided to us in accordance with state and national ethics regulations and lacking personal identifying information so as to ensure patient/donor confidentiality.

2.2. KIR Genotyping and Encoding. KIR gene content was determined using a locally developed sequence specific priming polymerase chain reaction (SSP-PCR) genotyping technique capable of detecting the presence or absence of each of the 17 genes [46]. This SSP-PCR approach did not enable us to distinguish between KIR2DL5A and KIR2DL5B nor the centromeric/telomeric localisation of genes. PCR amplicons were resolved in 1.5% agarose gels and digitally documented after ethidium bromide staining. Genotypes having KIR2DL2, KIR2DL5, KIR2DS1, KIR2DS2, KIR2DS3, KIR2DS5, or KIR3DS1 were considered to have at least one group B haplotype. Genotypes having KIR2DL3, KIR2DP1, KIR2DL1, KIR3DL1, and KIR2DS4 in the absence of any group B haplotype gene were classified as homozygous for group A haplotypes. Genotypes having all group A haplotype genes with at least one group B defining gene were considered heterozygous for groups A and B haplotypes. Centromeric and telomeric KIR haplotype motifs were deterministically inferred for the 300 samples after manually comparing their genotyping profile to that of the previously described KIR haplotype motifs based on criteria published previously by Pyo et al. [11]; see also Table 1 [46]. Similarly, KIR gene content haplotypes were inferred for the eleven most frequent genotypes observed in our population (present in >1% of our study population) based on Pyo's criteria [11]. As our genotyping approach does not resolve cis and trans relationships between genes, other haplotype motifs and/or haplotype combinations cannot be ruled out. Figure 1 provides overall classical KIR haplotype, haplotype motif, and extended haplotype frequencies for both study cohorts as provided by our online tool KIRHAT (KIR gene Haplotype Analysis Tool (KIRHAT) available through http://www.genomica.uaslp.mx).

Since KIR haplotype motifs can be inferred from genotyping results as described with greater detail in the original publication [11] and with the fact that the KIR haplotypes of the great majority of individuals contain the four framework genes KIR3DL3, KIR3DP1, KIR2DL4, and KIR3DL2 [11, 14]. Then, we only focus on the following 12 KIR genes: KIR2DL1, KIR2DL2, KIR2DL3, KIR2DL5, KIR2DS1, KIR2DS2, KIR2DS3, KIR2DS4, KIR2DS5, KIR2DP1, KIR3DL1, and KIR3DS1.

KIR gene encoding strings included information for the 12 genes for each of the 343 samples, stored in rows; see Table 2. Additionally, we have included a health status variable (C, known as class), which was =1 in samples obtained from individuals having a haematological malignancy and 0 in healthy donors, as shown in the last column of Table 2.

2.3. Traditional Statistical Tests. KIR gene carrier frequencies were calculated by direct counting of the number of individuals bearing a genetic trait. KIR gene and haplotype carrier frequency comparisons between healthy controls and diseased patients employed a two-sided Pearson's χ^2 or Fisher's exact test, significance being established at $p < 0.05$. This test is also known as 2-way contingency table analysis [31].

2.4. J48 Algorithm. The ID3 algorithm was originally introduced by Quinlan in 1983, and it is used for automatic rule generation in expert systems [32]. ID3 is also employed as a data mining tool to generate decision trees by using information entropy. Improved versions of ID3 include C4.5 and C5 algorithms. The J48 algorithm belongs to this class of algorithms for generating C4.5 decision trees [47].

2.5. A Priori Algorithm. This algorithm is used to find association rules given a dataset [39, 48]. A rule has two main components: the *if* and *then* part and the antecedent and the consequent part, respectively. We are going to use the symbols

Table 2: Study population; for visualization purposes, we only show the first five rows (disease, $C = 1$) and the last three rows (healthy, $C = 0$). Note that the last column corresponds to the class. Boxes with the mark √ indicate the presence of the gen (1), otherwise the absence (0).

Id	2DL1	2DL2	2DL3	2DL5	2DS1	2DS2	2DS3	2DS4	2DS5	2DP1	3DL1	3DS1	Disease (class—C)
1	√	√	√			√		√		√			1
2	√	√						√		√	√		1
3	√		√					√		√	√		1
4	√		√					√		√	√		1
5	√		√	√	√			√	√	√	√	√	1
⋮													⋮
341	√		√					√		√	√		0
342	√		√					√		√	√		0
343	√	√		√	√	√			√			√	0

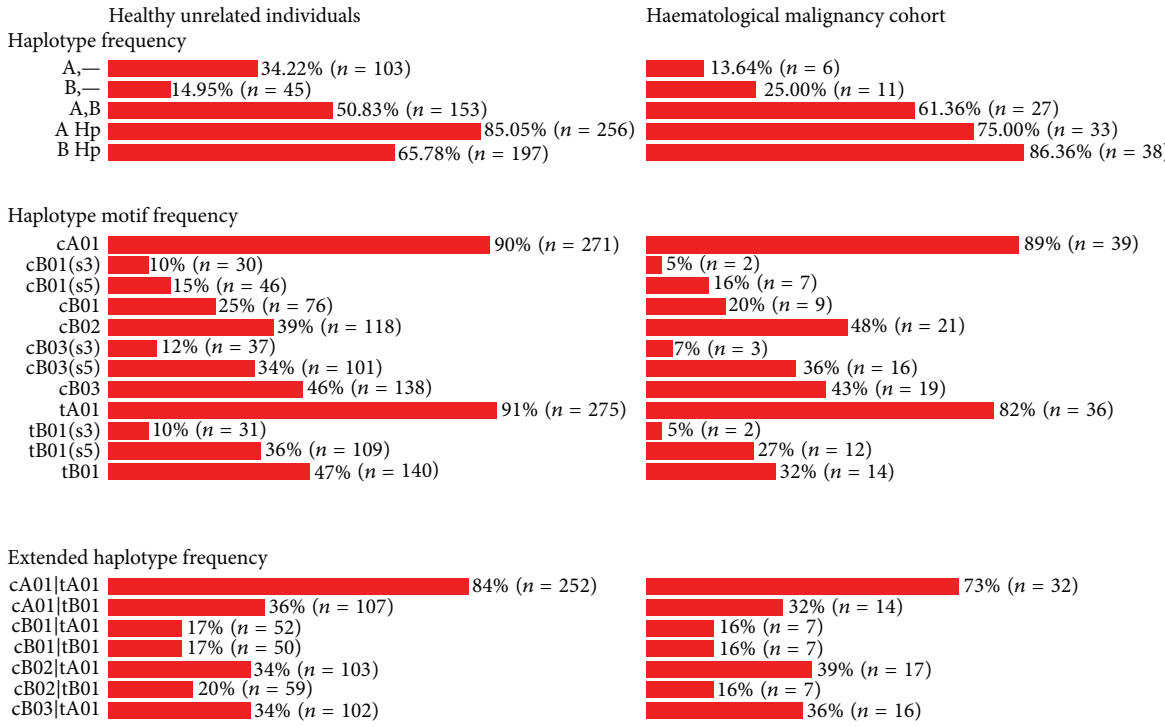


Figure 1: KIR gene features present in the healthy unrelated donor and haematological malignancy cohorts. KIR haplotype. A,—corresponds to group A homozygous haplotypes, whereas A Hp includes both homozygous and heterozygous group A haplotypes (vice versa for B). cB01 haplotypes having KIR2DS3 but not KIR2DS5 are indicated as "cB01(s3)," vice versa for those containing KIR2DS5 instead of KIR2DS3. The same applies to cB03 and tB01 categories. Combinations of centromeric and telomeric motifs that are thought to be very likely occurring based on Pyo's 2010 criteria [11] have been included at the bottom of the figure as extended haplotypes.

==> or ⇒ to separate those components of a rule. When several variables are involved within the *if* part, we consider the logical operator *and* (inclusive).

2.5.1. A Toy Example for the A Priori Algorithm. Before a formal explanation of the algorithm is given, a toy example with two genes (variables) is given. Let us consider only two genes ($g1$, $g2$, 0 indicates absence of gene while 1 indicates presence) and the clinical outcome (class 0 for healthy subjects and 1 for diseased); see Table 3. One can clearly see that only the cooccurrence of both genes leads to a diseased phenotype in this example while other combinations of the genes lead to a normal phenotype. In this specific case, the underlying behavior is best described by the AND operator (∧), in logic, where the performance is given by a truth table; see Table 3.

Based on this simple example, we then proceed to create an artificial dataset; see Table 4.

The dataset in Table 4 simulates 20 individuals, with only two genes ($g1$ and $g2$), and one class (C). If we apply a statistical analysis, we obtain the statistically significant p values of cross-tabulation comparison (shown in Figure 2(a)). Likewise, by applying the J48 algorithm a pruned tree (given in Figure 2(b)) is generated detailing associations rules along

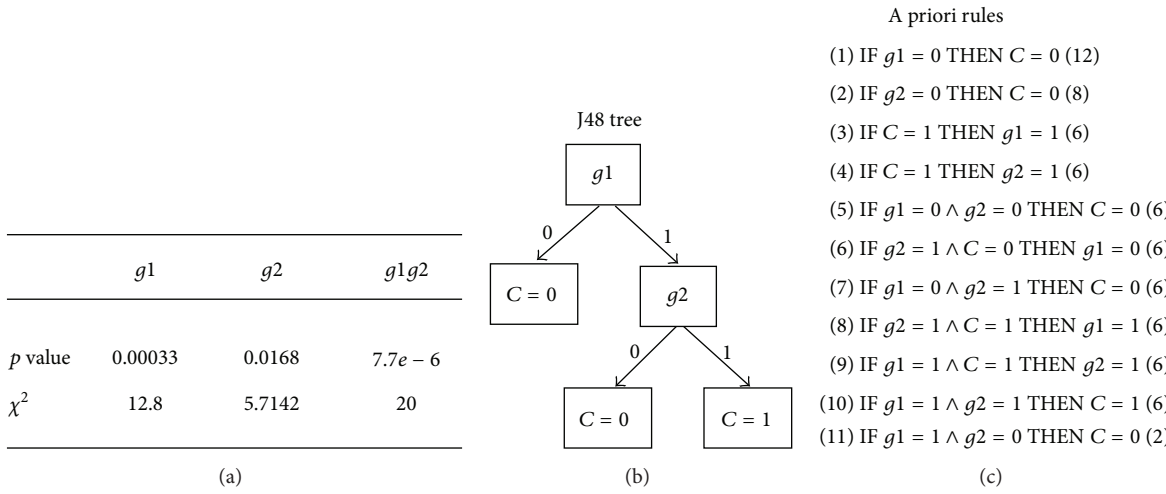


Figure 2: Results from the example. (a) Statistical test. (b) J48 pruned tree. (c) Rules given by the a priori algorithm.

Table 3: Truth table, AND operator (∧).

$g1$	$g2$	C (class)
0	0	0
0	1	0
1	0	0
1	1	1

Table 4: This table contains 20 records; there are two variables ($g1$ and $g2$); the class C also represents 0 when the donor is healthy and 1 diseased.

#	$g1$	$g2$	C (Class)
1	1	1	1
2	0	0	0
3	0	1	0
4	1	1	1
5	0	0	0
6	1	0	0
7	0	1	0
8	1	1	1
9	0	0	0
10	0	1	0
11	1	1	1
12	0	0	0
13	1	0	0
14	0	1	0
15	0	1	0
16	1	1	1
17	1	1	1
18	0	0	0
19	0	1	0
20	0	0	0

with a summary of those rules generated by the a priori algorithm (given in Figure 2(c)).

From this example, we can observe the following.

(1) Statistical Analysis. Here we show both univariate and multivariate statistical analyses. The column $g1g2$ combines the two variables $g1$ and $g2$. Since we apply the AND operator for combining variables, the data of the column $g1g2$ and C (Class) are the same. Therefore, the smallest p value is for the combined variable $g1g2$. However all p values are lower than our threshold ($p < 0.05$), so the results for all variables are statistically significant (or correlated). This is all we can infer from this simple statistical analysis.

(2) J48. The decision tree generated by the J48 algorithm agrees with the statistical analysis; the most important variable is $g1$, because it is at the first level of the tree. Moreover, it tells us that if the variable $g1$ is 0, then variable C is also 0. Still, it tells us that if variable $g1$ is 1, then we need to look at variable $g2$ to decide the value for C.

(3) A Priori Algorithm. This algorithm gives us the total of rules that can be inferred from the dataset in Table 3, which is all possible combinations among variables including the class variable (C). Besides, it also gives the most important rules, the first ones; that is, $g1 = 0 \Rightarrow C = 0$. This rule agrees with the statistical analysis and the J48 decision tree. The number (12), that is, the frequency, within the first rule, indicates how many times this rule applies in the whole dataset. Moreover, we can ask the algorithm to mine for class association rules, as we are only interested in rules where the class (C) appears as the consequent part of the rule:

(1) IF $g1 = 0$ THEN $C = 0$ (12)
(2) IF $g2 = 0$ THEN $C = 0$ (8)
(3) IF $g1 = 0 \wedge g2 = 0$ THEN $C = 0$ (6)
(4) IF $g1 = 0 \wedge g2 = 1$ THEN $C = 0$ (6)
(5) IF $g1 = 1 \wedge g2 = 1$ THEN $C = 1$ (6)
(6) IF $g1 = 1 \wedge g2 = 0$ THEN $C = 0$ (2).

If one observes these 6 rules, apart from the two first rules, they show the full performance of the AND operator, as shown in the truth table; see Table 3. It also tells us that if $g1 = 0$ then $C = 0$, regardless of the value of $g2$, and the same happens when $g2 = 0$. Finally, this result captures the

```
L_1 = {large1_itemsets} count item frequency
for (k = 2; L_{k-1} ≠ {}; k++) do begin
    C_k = apriori_gen(L_{k-1}); this function generate new candidates
    ∀transaction t ∈ D do begin
        C_t = subset(C_k, t); this function generate candidates in transaction t
        ∀candidates c ∈ C_t do
            c. count++; determine support
    end
    L_k = {c ∈ C_k | c. count ≥ min sup} create new set
end
Answer = ∪_k L_k
```

Pseudocode 1

main rule, which establishes the only case when $C = 1$; that is, $g1 = 1$ and $g2 = 1$.

2.5.2. Formal Definition of the A Priori Algorithm. Let us define formally the a priori algorithm, so $I = \{i_1, i_2, i_3, \ldots, i_m\}$ is a set of binary attributes called items. $D \subseteq \mathbb{P}(I)$ is a set of transactions, where $\mathbb{P}$ denotes the power set of I, that is, all subsets of I. For example, the power set of $S = \{a, b\}$ is $\mathbb{P}(S) = \{\{\}, \{a\}, \{b\}, \{a, b\}, \{b, a\}\}$. We are looking for implications, rules, of the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \phi$. We measure the quality of the rule by the following: (i) the support is the number of transactions where the antecedent of the rule is present, that is, $\text{supp}(X) = |X|/|D|$; (ii) the confidence measures the strength of the rule, and this measure is based on the support, where $\text{confidence}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X) = |X \cup Y|/|X|$; (iii) the correlation of a rules is based on probabilities, where $\text{correlation}(X \Rightarrow Y) = P(X \cup Y)/P(X)P(Y)$ [39, 48].

The pseudocode of the a priori algorithm [39, 48] is shown in Pseudocode 1.

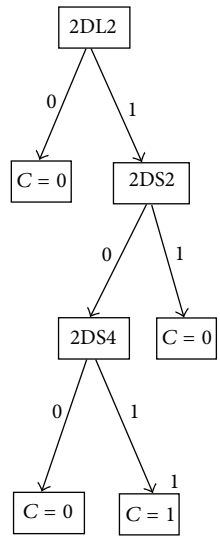


Figure 3: J48 decision tree.

2.5.3. Our Model. For our dataset of 12 KIR genes with the information of the 343 donors, as illustrated in Table 2, we use a set $I = \{i_1, i_2, i_3, \ldots, i_{13}\}$ with 13 items. The first twelve items represent the KIR genes, where $i_j = 1$ if the gene is present, and $i_j = 0$ if it is not. The item i_{13} corresponds to the class (C), where 0 indicates when the donor is healthy and 1 when the donor has some hematological malignancy (disease). The set D corresponds to the 343 donors; we are interested in association rules of the form

$$\left(i_j = v_j\right) \wedge \left(i_k = v_k\right) \wedge \cdots \wedge \left(i_l = v_l\right) \Longrightarrow C, \tag{1}$$

where $v_j, v_k, \ldots, v_l$ are the values of each item (0 or 1) and C denotes the class. Also the set $\{i_j, i_k, \ldots, i_l\} \subseteq I$, where $j \neq k \neq \cdots \neq l$.

2.6. Weka. The software that we use for our experiments is called Weka (http://www.cs.waikato.ac.nz/~ml/weka/index.html) [49]. It is open source software under the GNU general public license. The motivation of this software project is the invention and application of machine learning methods, so computer programs can automatically analyze large datasets. The results of these machine learning algorithms, in particular data mining algorithms, can be used to automatically make predictions or help people make decisions faster and accurately [49]. Weka contains a collection of machine learning algorithms for data mining tasks, in our case the J48 and a priori algorithm; see Figures 2 and 3. The algorithms can either be applied directly to a dataset through a graphical user interface (known as GUI) or called from your own Java code [49].

3. Results and Discussion

We use the programming language GNU Octave for performing both univariate and multivariate statistical analysis [50]; we employ a 2-way contingency table analysis [31]. We also use the Weka software to perform our experiments with J48 and the a priori algorithm. We then feed the J48 and the a priori algorithms with the dataset shown in Table 2, that having 12 KIR genes along with the class variable (healthy and disease donors) for 343 patients (samples).

Table 5: Univariate statistical analysis.

	2DL1	2DL2	2DL3	2DL5	2DS1	2DS2	2DS3	2DS4	2DS5	2DP1	3DL1	3DS1
p value	0.752	0.0000087	0.467	0.214	0.421	0.271	0.131	0.199	0.946	0.921	0.042	0.888
χ^2	0.100	19.764	0.530	1.547	0.649	1.213	2.281	1.649	0.005	0.010	4.128	0.020

Table 6: Multivariate statistical analysis; here we show only the variable combinations associated to the haplotype cA01|tA01. Boxes with the mark √ indicate that the variable is part of the variable combination; otherwise it is not taken in account.

#	2DL1	2DL2	2DL3	2DL5	2DS1	2DS2	2DS3	2DS4	2DS5	2DP1	3DL1	3DS1	p value	χ^2
1	√	√	√					√	√		√		0.00036	12.7
2	√	√	√					√		√	√	√	0.01918	5.4
3	√	√	√					√	√	√	√		0.00053	11.9
4	√	√	√		√			√		√	√		0.00022	13.5
5	√	√	√	√				√		√	√		**0.00002**	**17.4**
6	√	√	√					√	√	√	√	√	0.04289	4.09
7	√	√	√		√			√		√	√	√	0.01918	5.4
8	√	√	√		√			√	√	√	√		0.00574	7.6
9	√	√	√	√				√		√	√	√	0.01918	5.4
10	√	√	√	√				√	√	√	√		0.00213	9.4
11	√	√	√	√	√			√		√	√		0.00246	9.1
12	√	√	√		√			√	√	√	√	√	0.04289	4.09
13	√	√	√	√				√	√	√	√	√	0.04289	4.09
14	√	√	√	√	√			√		√	√	√	0.01918	5.4
15	√	√	√	√	√			√	√	√	√		0.00574	7.6
16	√	√	√	√	√			√	√	√	√	√	0.04289	4.09

3.1. Statistical Analysis Results. The results of the univariate statistical analysis are in Table 5. The significant results ($p < 0.05$) are only for KIR2DL2 and KIR3DL1. There are neither motifs nor haplotypes associated to these two genes.

Traditional statistical comparison of KIR gene carrier frequencies (2×2 tables using Fishers' exact test) showed that KIR2DL2 was more frequent amongst the haematological malignancy cohort in comparison to the healthy individuals (77.8% versus 40.3%, resp.; $p < 0.0001$), Group A homozygosity was less frequent (11.1% versus 32%, resp.; $p < 0.0044$) and A,B heterozygous haplotypes were more frequent (86.7% versus 58.7%, resp.; $p < 0.0002$). This finding is interesting as KIR2DL2 is in tight linkage disequilibria (LD) with another gene, KIR2DS2. Both KIR2DL2 and KIR2DS2 are thought to bind HLA-C allotypes having C1 group specificity. Nevertheless, KIR2DL2 is an inhibitory protein whereas 2DS2 is activating. All genotyping reactions were carried out in triple, with further confirmatory runs if required. In addition, all genotyping was done at the same lab. As such, we are certain that this lack of LD is not related to technical issues. However, we cannot rule out that this might be the result of genotyping allele-dropout (failure to amplify a KIR2DS2 allele particularly common in the leukaemia cohort) or of cross-hybridization of 2DL2 oligonucleotides with other genes. This last possibility is unlikely as this genotyping approach has been previously validated and this finding does not occur in the healthy donor cohort.

For the multivariate statistical analysis, we take into account all 12 KIR genes variables. Therefore, we have $\sum_{i=2}^{12} \binom{12}{i} = 4083$ combinations. From these set of combinations, if we set our threshold to $p \leq 0.05$ then we obtain 336 significant variable combinations. There are only 16 variable combinations associated to the haplotype cA01|tA01; see Table 6. If we set $p \leq 0.0001$, then we obtain only 35 significant variable combinations and only one variable combination is associated to the haplotype cA01|tA01, that is, the variable combination #5 in Table 6. From the multivariate statistical analysis, the best variable combination is for KIR2DL1, KIR2DL2, KIR2DL3, KIR2DL5, KIR2DS4, KIR2DP1, and KIR3DL1; p value = 0.00002.

3.2. J48 Algorithm Results. In Figure 3, we show the results of the J48 algorithm. The only case when the donor is associated to a hematological malignancy (disease; $C = 1$) is when the gen KIR2DL2 is present (=1), KIR2DS2 is absent (=0), and KIR2DS4 is present (=1). There are not any motifs and haplotypes associated with this decision tree.

3.3. A Priori Algorithm Results. The a priori algorithm generates a total of 71,006 rules, taking in account only the rules where the class (C) appears at the consequent part of the rule, and there are only 12,052 rules associated to $C = 1$ (disease). In Table 7, we show only the first rules as generated by the a priori algorithm (where $C = 1$). The first 24 rules (out of 12,052) are more important because they are more frequent than the others. In Table 7, the frequency means that these rules are satisfied for 10 donors out of 43; that is, this pattern is present in 23% of the disease donors.

Because the variability in KIR genotype is such that most pairs of unrelated human individuals have different KIR genotypes, the unique feature of the human KIR system is the representation of two distinctive groups of haplotypes (A and B) [11]. Therefore, the more relevant rule given by the a priori algorithm, in Table 7, is the rule Id = 1870 (2DL1

Table 7: Rules generated by the a priori algorithm represented in tabular form. This figure contains only 24 rules with frequency 10, where the class = 1 (C).

#	Id	KIR2DL1	KIR2DL2	KIR2DL3	KIR2DL5	KIR2DS2	KIR2DS4	KIR2DP1	KIR3DL1	Frequency
1	1476		2DL2 = 1		2DL5 = 1	2DS2 = 0	2DS4 = 1			10
2	1477		2DL2 = 1		2DL5 = 1	2DS2 = 0			3DL1 = 1	10
3	1528	2DL1 = 1	2DL2 = 1		2DL5 = 1	2DS2 = 0	2DS4 = 1			10
4	1529	2DL1 = 1	2DL2 = 1		2DL5 = 1	2DS2 = 0			3DL1 = 1	10
5	1558		2DL2 = 1	2DL3 = 1	2DL5 = 1	2DS2 = 0	2DS4 = 1			10
6	1559		2DL2 = 1	2DL3 = 1	2DL5 = 1	2DS2 = 0			3DL1 = 1	10
7	1560		2DL2 = 1		2DL5 = 1	2DS2 = 0	2DS4 = 1	2DP1 = 1		10
8	1561		2DL2 = 1		2DL5 = 1	2DS2 = 0	2DS4 = 1		3DL1 = 1	10
9	1562		2DL2 = 1		2DL5 = 1	2DS2 = 0		2DP1 = 1	3DL1 = 1	10
10	1651	2DL1 = 1	2DL2 = 1	2DL3 = 1	2DL5 = 1	2DS2 = 0	2DS4 = 1			10
11	1652	2DL1 = 1	2DL2 = 1	2DL3 = 1	2DL5 = 1	2DS2 = 0			3DL1 = 1	10
12	1653	2DL1 = 1	2DL2 = 1		2DL5 = 1	2DS2 = 0	2DS4 = 1	2DP1 = 1		10
13	1654	2DL1 = 1	2DL2 = 1		2DL5 = 1	2DS2 = 0	2DS4 = 1		3DL1 = 1	10
14	1655	2DL1 = 1	2DL2 = 1		2DL5 = 1	2DS2 = 0		2DP1 = 1	3DL1 = 1	10
15	1681		2DL2 = 1	2DL3 = 1	2DL5 = 1	2DS2 = 0	2DS4 = 1	2DP1 = 1		10
16	1682		2DL2 = 1	2DL3 = 1	2DL5 = 1	2DS2 = 0	2DS4 = 1		3DL1 = 1	10
17	1683		2DL2 = 1	2DL3 = 1	2DL5 = 1	2DS2 = 0		2DP1 = 1	3DL1 = 1	10
18	1684		2DL2 = 1		2DL5 = 1	2DS2 = 0	2DS4 = 1	2DP1 = 1	3DL1 = 1	10
19	1784	2DL1 = 1	2DL2 = 1	2DL3 = 1	2DL5 = 1	2DS2 = 0	2DS4 = 1	2DP1 = 1		10
20	1785	2DL1 = 1	2DL2 = 1	2DL3 = 1	2DL5 = 1	2DS2 = 0	2DS4 = 1		3DL1 = 1	10
21	1786	2DL1 = 1	2DL2 = 1	2DL3 = 1	2DL5 = 1	2DS2 = 0		2DP1 = 1	3DL1 = 1	10
22	1787	2DL1 = 1	2DL2 = 1		2DL5 = 1	2DS2 = 0	2DS4 = 1	2DP1 = 1	3DL1 = 1	10
23	1806		2DL2 = 1	2DL3 = 1	2DL5 = 1	2DS2 = 0	2DS4 = 1	2DP1 = 1	3DL1 = 1	10
24	**1870**	**2DL1 = 1**	**2DL2 = 1**	**2DL3 = 1**	**2DL5 = 1**	**2DS2 = 0**	**2DS4 = 1**	**2DP1 = 1**	**3DL1 = 1**	**10**

= 1, 2DL2 = 1, 2DL3 = 1, 2DL5 = 1, 2DS2 = 0, 2DS4 = 1, 2DP1 = 1, 3DL1 = 1 ==> Class = 1). This rule refers to the haplotype cA01|tA01 [11], which is strongly inhibitory and then tolerates the tumors. In addition to this haplotype, two more inhibitory genes 2DL2 and 2DL5 are also present in this rule (which are part of the haplotype cB03), and the activating gen KIR2DS2 is absent. This association has been suggested for certain Hodgkin's lymphomas [38].

Moreover, it is clear, from Table 7, that the first 23 rules are a subset of the main rule (Id = 1870), the new discovered pattern. In fact, all of them have the same frequency. In other words, the first 23 rules are derivations from the rule #24 (Id = 1870); for example, the toy example shown above for the AND operator has the rule IF $g1 = 0 \wedge g2 = 0$ THEN $C = 0$, so the rules IF $g1 = 0$ THEN $C = 0$ and IF $g2 = 0$ THEN $C = 0$ are a subset of the previous rule. From Table 7, we can also infer that the genes KIR2DS1, KIR2DS3, KIR2DS5, and KIR3DS1 are somehow irrelevant, since they do not appear in any of these 24 rules.

Some researchers have reported some associations related to KIR2DS3 [35–37]. The rules shown in Table 7 (Class = 1) are only associated to disease ($C = 1$) with the absence of KIR2DS3. However, neither the J48 decision tree (Figure 3) nor the main rules generated by the a priori algorithm (Table 7) found some association between KIR2DS3 and disease.

3.4. Statistical Analysis versus the A Priori Algorithm. The unique feature of the human KIR system, which is not mirrored in other higher primates, is the representation of haplotypes (A and B). The haplotypes are present in all the >150 human populations studied [4]. Therefore, the association between haplotypes and disease is more important than only KIR genotype and disease.

In Table 8, we show the comparison between the multivariate statistical analysis and the a priori algorithm results. Table 8(a) shows the contingency table for the statistical analysis, and we can observe that this variable combination is associated to 18 disease donors (41%) of our study populations, although it is also associated to 46 healthy donors (15%). On the other hand, in Table 8(b), the contingency table for the rule found by the a priori algorithm shows that the rule is associated to 10 disease donors (23%), but it is not associated to any healthy donor. In other words, this rules is unique since it is only associated to disease donors. In fact, the p value and the χ^2 value show that the result is more statistically significant for the rule found by the a priori algorithm.

4. Conclusions

We studied a population of 300 healthy donors and 43 donors with haematological malignancies. The J48 algorithm and the univariate statistical analysis did not find any associations between haplotypes and disease. The multivariate analysis found 336 statistically significant variable combinations associated with the haplotype cA01|tA01 ($p \leq 0.05$). From these set of combinations there is only one variable combination associated to this haplotype with $p \leq 0.0001$ (see #5 in Table 6). This variable combination is associated to both disease and healthy donors (see Table 8). On the other hand,

Table 8: Statistical analysis of 2-way contingency tables. (a) This table corresponds to the variable combination #5 in Table 6. (b) This table corresponds to the rule Id = 1870 in Table 7.

(a) Multivariate statistical analysis

	Disease	Healthy
Disease	18	25
Healthy	46	254

p value = 0.00002; χ^2 = 17.4.

(b) A priori algorithm

	Disease	Healthy
Disease	10	33
Healthy	0	300

p value = 0.0; χ^2 = 71.86.

the a priori algorithm was able to discover a unique pattern through the rule Id = 1870. This pattern is more statistically significant than the variable combinations found by the multivariate statistical analysis (see Table 8). Moreover, the rule Id = 1870 is only associated to disease donors. In contrast, the variable combination found by the multivariate analysis is associated to both healthy and diseases donors. The rule Id = 1870 not only refers to the haplotype cA01|tA01, which is a predominantly inhibitory haplotype. This rule also refers to the genes KIR2DL2 and KIR2DL5, which are also inhibitory but not present in this haplotype which can be thought of more likely to tolerate tumours in our study population (with strict absence of KIR2DS2), that is, Mexican mestizos of San Luis Potosi State. This pattern was not discovered with previous studies on the same study population [33]. The methodology proposed in this paper provides a new insight into the analysis of datasets that allow researchers to find biomarkers for cancer and other diseases. Although the size and heterogeneity of our study cohort together with the lack of HLA typing data limits the clinical inferences that can be made from our results, it sets an example for a different way of analysing the clinical and functional relevance of complex genetic systems. Despite this, our methodology is able to discover patterns unseen for statistical analysis and decision trees generated by ID3 or J48 algorithms. The huge amount of rules generated by the a priori algorithm involves a data mining work to obtain the relevant rules. We found that the best performance is when a lower bound support is set to zero in combinations with a configuration that allows us to select rules only when the class is equal to one. The disadvantage of the a priori algorithm is that it requires huge computational resources (memory and processing). More research is needed to speed this algorithm up, and this may be the reason that this algorithm is not used in bioinformatics. A dataset with 23 variables is intractable for the Weka software with a personal computer. However, the dataset studied in this paper is able to run in the Weka software using a personal computer with a processor Intel Core i7 with 2.3 Ghz speed and 3 Gb memory. Undergoing investigations by our research group include the study of a dataset with KIR and HLA information of 413 HIV donors against our reference population of 300 healthy donors. We found that datasets with less 13 variables can be analysed on a personal computer regardless of the number of donors. Alternatively, there is commercial software to execute the a priori algorithm on a given dataset such as STATISTICA [51]; this software can manage more than 13 variables, but it also demands high computational resources.

Disclaimer

The study sponsor had no role in study design, collection, analysis, and interpretation of data.

Conflict of Interests

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

Acknowledgments

The authors wish to thank Dr. Oscar Pérez Ramírez and Dr. Arturo Sánchez Arriaga of the Haematology Service and Blood Bank of Hospital Central "Dr. Ignacio Morones Prieto" for providing the patient samples that made this work possible. Special thanks to Dr. Daniel E. Noyola of the Virology Laboratory, Facultad de Medicina, Universidad Autónoma de San Luis Potosí, for proofreading this paper. The authors also thank Dr. Victor Trevino of the ITESM campus Monterrey for reviewing and helping to improve this paper. This work was funded by grants provided from Universidad Autónoma de San Luis Potosí (P/PIFI2009-24MSU0011E-12), Convocatoria CONACYT de Investigación Científica Básica 2006 (CONACYT no. 55360), and PRODEP (Apoyo para gastos de publicación SEP-23-007-B).

References

[1] C. Auffray and L. Hood, "Systems biology and personalized medicine—the future is now," *Biotechnology Journal*, vol. 7, no. 8, pp. 938–939, 2012.

[2] V. Trevino, F. Falciani, and H. A. Barrera-Saldaña, "DNA microarrays: a powerful genomic tool for biomedical and clinical research," *Molecular Medicine*, vol. 13, no. 9-10, pp. 527–541, 2007.

[3] E. Yong, "Cancer biomarkers: written in blood," *Nature*, vol. 511, no. 7511, pp. 524–526, 2014.

[4] K. A. McAulay and R. F. Jarrett, "Human leukocyte antigens and genetic susceptibility to lymphoma," *Tissue Antigens*, vol. 86, no. 2, pp. 98–113, 2015.

[5] A. M. Dickinson and J. Norden, "Non-HLA genomics: does it have a role in predicting haematopoietic stem cell transplantation outcome?" *International Journal of Immunogenetics*, vol. 42, no. 4, pp. 229–238, 2015.

[6] P. Jia and Z. Zhao, "Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives," *Human Genetics*, vol. 133, no. 2, pp. 125–138, 2014.

[7] E. Gelmann, C. Sawyers, and F. Rauscher II, *Molecular Oncology: Causes of Cancer and Targets for Treatment*, Cambridge University Press, 2014.

[8] V. Litwin, J. Gumperz, P. Parham, J. H. Phillips, and L. L. Lanier, "Specificity of HLA class I antigen recognition by human NK clones: evidence for clonal heterogeneity, protection by self and non-self alleles, and influence of the target cell type," *Journal of Experimental Medicine*, vol. 178, no. 4, pp. 1321–1336, 1993.

[9] A. Moretta, C. Bottino, D. Pende et al., "Identification of four subsets of human CD3-CD16+ natural killer (NK) cells by the expression of clonally distributed functional surface molecules: correlation between subset assignment of NK clones and ability to mediate specific alloantigen recognition," *Journal of Experimental Medicine*, vol. 172, no. 6, pp. 1589–1598, 1990.

[10] J. Robinson, K. Mistry, H. Mcwilliam, R. Lopez, and S. G. E. Marsh, "Ipd-the immuno polymorphism database," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D863–D869, 2009.

[11] C.-W. Pyo, L. A. Guethlein, Q. Vu et al., "Different patterns of evolution in the centromeric and telomeric regions of group A and B haplotypes of the human killer cell Ig-like receptor locus," *PLoS ONE*, vol. 5, no. 12, Article ID e15115, 2010.

[12] J. A. Hollenbach, I. Nocedal, M. B. Ladner, R. M. Single, and E. A. Trachtenberg, "Killer cell immunoglobulin-like receptor (KIR) gene content variation in the HGDP-CEPH populations," *Immunogenetics*, vol. 64, no. 10, pp. 719–737, 2012.

[13] J. Trowsdale, "Genetic and functional relationships between MHC and NK receptor genes," *Immunity*, vol. 15, no. 3, pp. 363–374, 2001.

[14] K. C. Hsu, S. Chida, D. E. Geraghty, and B. Dupont, "The killer cell immunoglobulin-like receptor (KIR) genomic region: gene-order, haplotypes and allelic polymorphism," *Immunological Reviews*, vol. 190, pp. 40–52, 2002.

[15] R. B. Herberman, M. E. Nunn, H. T. Holden, and D. H. Lavrin, "Natural cytotoxic reactivity of mouse lymphoid cells against syngeneic and allogeneic tumors. II. Characterization of effector cells," *International Journal of Cancer*, vol. 16, no. 2, pp. 230–239, 1975.

[16] R. Kiessling, E. Klein, and H. Wigzell, "'Natural' killer cells in the mouse. I. Cytotoxic cells with specificity for mouse Moloney leukemia cells. Specificity and distribution according to genotype," *European Journal of Immunology*, vol. 5, no. 2, pp. 112–117, 1975.

[17] R. Kiessling, E. Klein, H. Pross, and H. Wigzell, "'Natural' killer cells in the mouse. II. Cytotoxic cells with specificity for mouse Moloney leukemia cells. Characteristics of the killer cell," *European Journal of Immunology*, vol. 5, no. 2, pp. 117–121, 1975.

[18] R. T. Costello, C. Fauriat, S. Sivori, E. Marcenaro, and D. Olive, "NK cells: innate immunity against hematological malignancies?" *Trends in Immunology*, vol. 25, no. 6, pp. 328–333, 2004.

[19] L. Ruggeri, M. Capanni, M. Casucci et al., "Role of natural killer cell alloreactivity in HLA-mismatched hematopoietic stem cell transplantation," *Blood*, vol. 94, no. 1, pp. 333–339, 1999.

[20] S. Cooley, E. Trachtenberg, T. L. Bergemann et al., "Donors with group B KIR haplotypes improve relapse-free survival after unrelated hematopoietic cell transplantation for acute myelogenous leukemia," *Blood*, vol. 113, no. 3, pp. 726–732, 2009.

[21] S. M. Davies, L. Ruggieri, T. DeFor et al., "Evaluation of KIR ligand incompatibility in mismatched unrelated donor hematopoietic transplants. Killer immunoglobulin-like receptor," *Blood*, vol. 100, no. 10, pp. 3825–3827, 2002.

[22] K. Gagne, G. Brizard, B. Gueglio et al., "Relevance of KIR gene polymorphisms in bone marrow transplantation outcome," *Human Immunology*, vol. 63, no. 4, pp. 271–280, 2002.

[23] S. Giebel, F. Locatelli, T. Lamparelli et al., "Survival advantage with KIR ligand incompatibility in hematopoietic stem cell transplantation from unrelated donors," *Blood*, vol. 102, no. 3, pp. 814–819, 2003.

[24] K. C. Hsu, T. Gooley, M. Malkki et al., "KIR ligands and prediction of relapse after unrelated donor hematopoietic cell transplantation for hematologic malignancy," *Biology of Blood and Marrow Transplantation*, vol. 12, no. 8, pp. 828–836, 2006.

[25] K. Stringaris, S. Adams, M. Uribe et al., "Donor KIR Genes 2DL5A, 2DS1 and 3DS1 are associated with a reduced rate of leukemia relapse after HLA-identical sibling stem cell transplantation for acute myeloid leukemia but not other hematologic malignancies," *Biology of Blood and Marrow Transplantation*, vol. 16, no. 9, pp. 1257–1264, 2010.

[26] H. J. Symons, M. S. Leffell, N. D. Rossiter, M. Zahurak, R. J. Jones, and E. J. Fuchs, "Improved survival with inhibitory killer immunoglobulin receptor (KIR) gene mismatches and KIR haplotype B donors after nonmyeloablative, HLA-haploidentical bone marrow transplantation," *Biology of Blood and Marrow Transplantation*, vol. 16, no. 4, pp. 533–542, 2010.

[27] L. Ruggeri, M. Capanni, E. Urbani et al., "Effectiveness of donor natural killer cell alloreactivity in mismatched hematopoietic transplants," *Science*, vol. 295, no. 5562, pp. 2097–2100, 2002.

[28] P. H. Basse, T. L. Whiteside, W. Chambers, and R. B. Herberman, "Therapeutic activity of NK cells against tumors," *International Reviews of Immunology*, vol. 20, no. 3-4, pp. 439–501, 2001.

[29] B. Gansuvd, M. Hagihara, Y. Yu et al., "Human umbilical cord blood NK T cells kill tumors by multiple cytotoxic mechanisms," *Human Immunology*, vol. 63, no. 3, pp. 164–175, 2002.

[30] M. J. Smyth, K. Y. T. Thia, S. E. A. Street et al., "Differential tumor surveillance by natural killer (NK) and NKT cells," *Journal of Experimental Medicine*, vol. 191, no. 4, pp. 661–668, 2000.

[31] B. Rosner, *Fundamentals of Biostatistics*, Duxbury Press, Pacific Grove, Calif, USA, 6th edition, 2006.

[32] J. Ignizio, *Introduction to Expert Systems*, McGraw-Hill, 1991.

[33] J. C. Cuevas Tello, D. Hernández-Ramírez, and C. A. García-Sepúlveda, "Support vector machine algorithms in the search of KIR gene associations with disease," *Computers in Biology and Medicine*, vol. 43, no. 12, pp. 2053–2062, 2013.

[34] C. Besson, S. Roetynck, F. Williams et al., "Association of killer cell immunoglobulin-like receptor genes with Hogkin's lymphoma in a familial study," *PLoS ONE*, vol. 2, no. 5, article e406, 2007.

[35] F. Shahsavar, N. Tajik, K.-Z. Entezami et al., "KIR2DS3 is associated with protection against acute myeloid leukemia," *Iranian Journal of Immunology*, vol. 7, no. 1, pp. 8–17, 2010.

[36] L. Karabon, A. Jedynak, S. Giebel et al., "KIR/HLA gene combinations influence susceptibility to B-cell chronic lymphocytic leukemia and the clinical course of disease," *Tissue Antigens*, vol. 78, no. 2, pp. 129–138, 2011.

[37] G. Q. Wu, Y. M. Zhao, X. Y. Lai et al., "The beneficial impact of missing KIR ligands and absence of donor KIR2DS3 gene on outcome following unrelated hematopoietic SCT for myeloid leukemia in the Chinese population," *Bone Marrow Transplantation*, vol. 45, no. 10, pp. 1514–1521, 2010.

[38] M. K. Gandhi, J. T. Tellam, and R. Khanna, "Epstein-Barr virus-associated Hodgkin's lymphoma," *British Journal of Haematology*, vol. 125, no. 3, pp. 267–281, 2004.

[39] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 207–216, Washington, DC, USA, May 1993.

[40] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," in *Advances in Knowledge Discovery and Data Mining*, pp. 1–34, American Association for Artificial Intelligence, Menlo Park, Calif, USA, 1996.

[41] P. Ning-Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Addison-Wesley, 2006.

[42] J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[43] C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan, and Q. Zou, "LibD3C: ensemble classifiers with a clustering and dynamic selection strategy," *Neurocomputing*, vol. 123, pp. 424–435, 2014.

[44] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, pp. 42–47, 2012.

[45] C. A. García-Sepúlveda, E. Carrillo-Acuña, S. E. Guerra-Palomares, and M. Barriga-Moreno, "Maxiprep genomic DNA extractions for molecular epidemiology studies and biorepositories," *Molecular Biology Reports*, vol. 37, no. 4, pp. 1883–1890, 2010.

[46] D. L. Alvarado-Hernández, D. Hernández-Ramírez, D. E. Noyola, and C. A. García-Sepúlveda, "KIR gene diversity in Mexican mestizos of San Luis Potosí," *Immunogenetics*, vol. 63, no. 9, pp. 561–575, 2011.

[47] R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, Calif, USA, 1993.

[48] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, August 1998.

[49] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[50] https://www.gnu.org/software/octave/.

[51] Statsoft, "STATISTICA," 2014, http://www.statsoft.com/.