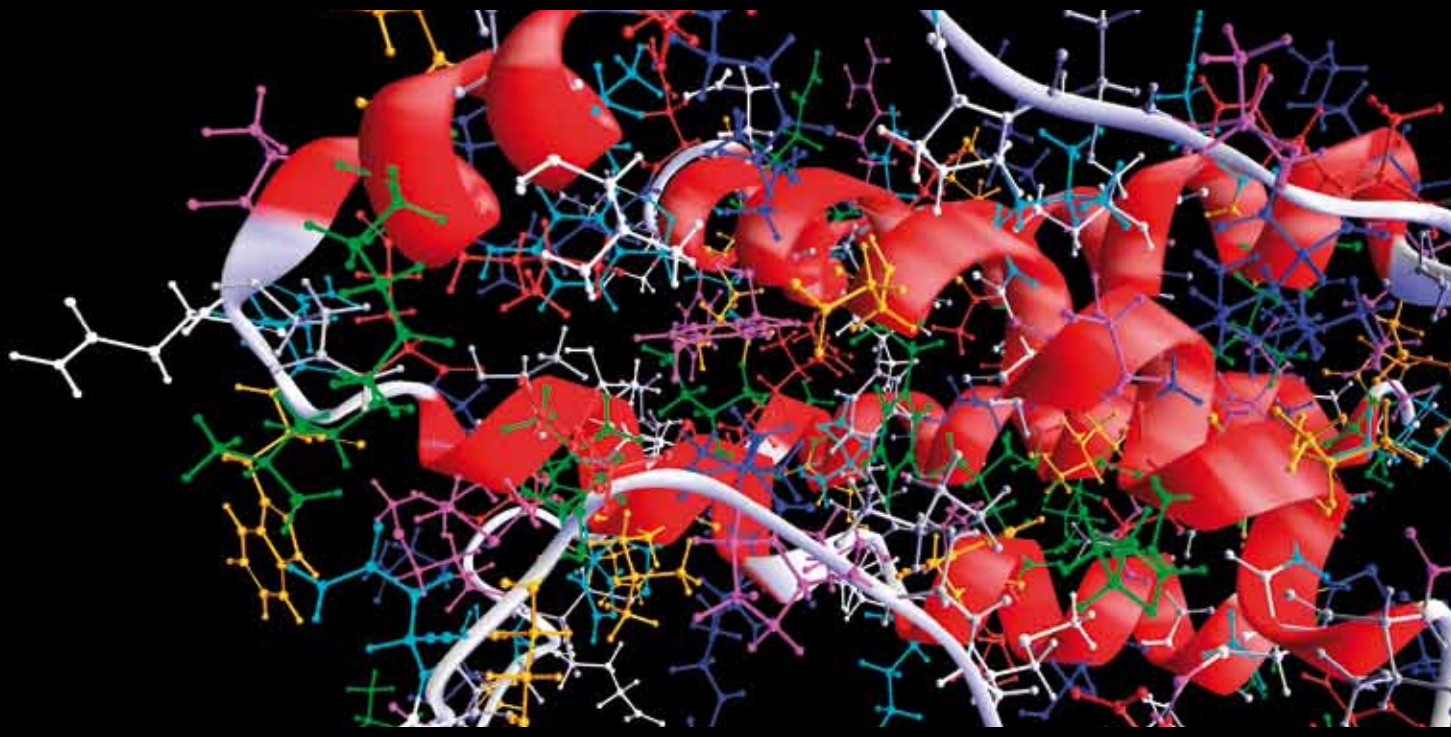
Computational and Mathematical Methods in Medicine

Translational Bioinformatics and Computational Systems Medicine

Guest Editors: Bairong Shen, Hong-Bin Shen, TianHai Tian, Qiang Lü, and Guang Hu



Translational Bioinformatics and Computational Systems Medicine

Computational and Mathematical Methods in Medicine

Translational Bioinformatics and Computational Systems Medicine

Guest Editors: Bairong Shen, Hong-Bin Shen, Tianhai Tian, Qiang Lü, and Guang Hu

Editorial Board

Contents

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 375641, 2 pages
http://dx.doi.org/10.1155/2013/375641

Editorial

Translational Bioinformatics and Computational Systems Medicine

Bairong Shen,[1] Hong-Bin Shen,[2] Tianhai Tian,[3] Qiang Lü,[4] and Guang Hu[1]

[1] *Center for Systems Biology, Soochow University, P.O. Box 206, Suzhou 215006, China*
[2] *Department of Automation, Shanghai Jiao Tong University, No. 800 Dongchuan Road, Shanghai 200240, China*
[3] *School of Mathematical Sciences, Monash University, Melbourne, VIC 3800, Australia*
[4] *School of Computer Science and Technologies, Soochow University, Suzhou 215006, China*

Correspondence should be addressed to Bairong Shen; bairong.shen@suda.edu.cn

Received 21 April 2013; Accepted 21 April 2013

This special issue is dedicated to Translational bioinformatics and Computational Systems Medicine. Translational biomedical informatics is a rapidly emerging discipline to integrate data from medical research, biotechnologies, and electronic medical records, and computational systems medicine is to apply computational and systems biology approaches to solve complex problems in medical research, aiming to improve the diagnosis, prognosis, and treatment of complex diseases. It is also well known that their development needs an integration of mathematical models, statistical methods, and computer algorithms.

Complex diseases such as cancers are coursed by a combination of genetic, environmental, and lifestyle factors, and thus the research of complex diseases at a system level like gene sets, pathway level, or static/dynamic network is a necessity. T. Holden et al. present an exploratory bioinformatics study of long noncoding RNA in Alzheimer's disease. The authors also discuss the model for drug development based on fractal dimension and entropy correlation in this study consistent with a zebrafish model and a mouse model. M. Ding et al. explore the genome-wide chromatin localization of Estrogen receptor-DNA binding regions by analyzing ChIP-Seq data from MCF-7 breast cancer cell line. It reveals novel Estrogen receptor-regulated genes pathways for further experimental validation. Y.-w. Lv et al. perform gene ontology category and pathways analysis for relationships among statistically significant genes in Kawasaki disease. The importance of compelling immune pathway of NF-AT signal and leukocyte interactions combined with another transcription factor NF-κB in the pathogenesis of KD is investigated by network analysis. J. Chen et al. perform two case studies on colorectal and prostate cancer microarray datasets to proof two hypotheses that (1) the expression signatures of different cancer microarray datasets are more similar at pathway level than at gene level; (2) the comparability of the cancer molecular mechanisms of different individuals is related to their genetic similarities.

The concept of entropy suggests that systems naturally progress from order to disorder. Entropy-based methods provide a novel insight into understanding many phenomena in biological systems. V. Oswal et al. present an automated entropy-based thresholding system for segmentation and quantification of cell nuclei from histologically stained images. The contributions to the application of this entropy-based system to detect cancerous cell nuclei and observe overlapping cellular events occurring during wound healing process in the human body are also presented. H.-T. Wu et al. investigate the feasibility and sensitivity of a novel multiscale entropy index in detecting subtle vascular abnormalities in healthy and diabetic subjects. The authors further focus on four groups of subjects to discuss the application of multiscale entropy index.

The role of protein structures in understanding diseases becomes more and more important, due to the following two reasons. One is that there are a lot of disease-associated proteins that were discovered, while the other fact is that many diseases are believed to result from misfolded proteins. Moreover, protein structures can be considered as complex

systems, and thus network theory can be used to characterize and to analyze protein structures. J. Sun et al. present a contribution which focuses on the effect of edge definition of complex networks on protein structure identification. After the performance on 2847 proteins, the authors argue that the optimal cutoff value for constructing the protein structure networks is 5.0 Å. X. Jiao et al. improve the amino acid network for proteins by introducing similar weight and dissimilar weight. This work demonstrates that highly central residues of the amino acid network are highly correlated with the hot spots in disease-associated proteins.

Finally, two bioinformatics tools were also involved in this issue. H. Wu et al. contribute a bioinformatics tool called patGPCR, to predict the 3D structures of transmembrane helices of G-protein-coupled receptors. patGPCR, a parallelized multitemplate approach, extends a bundle-packing related energy function to RosettaMem energy, which improves the TM RMSD (root mean square deviation of the transmembrane helices) of the predicted models. N. Deng et al. contribute another platform, called crcTRP, for colorectal cancer. This server provides the translational research of colorectal cancer by providing various types of biomedical information, including clinical data, epidemiology data, individual omics data, and public omics data.

Bairong Shen
Hong-Bin Shen
Tianhai Tian
Qiang Lü
Guang Hu

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 579136, 8 pages
http://dx.doi.org/10.1155/2013/579136

Research Article

Exploratory Bioinformatics Study of lncRNAs in Alzheimer's Disease mRNA Sequences with Application to Drug Development

T. Holden,[1] A. Nguyen,[1] E. Lin,[2] E. Cheung,[1] S. Dehipawala,[1] J. Ye,[1] G. Tremberger Jr.,[1] D. Lieberman,[1] and T. Cheung[1]

[1] *Queensborough Community College of CUNY, 222-05 56th Avenue Bayside, NY 11364, USA*
[2] *Albert Einstein College of Medicine, Department of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, USA*

Correspondence should be addressed to T. Holden; tholden@qcc.cuny.edu

Received 9 December 2012; Revised 11 March 2013; Accepted 15 March 2013

Academic Editor: Bairong Shen

Long noncoding RNA (lncRNA) within mRNA sequences of Alzheimer's disease genes, namely, APP, APOE, PSEN1, and PSEN2, has been analyzed using fractal dimension (FD) computation and correlation analysis. We examined lncRNA by comparing mRNA FD to corresponding coding DNA sequences (CDSs) FD. APP, APOE, and PSEN1 CDSs select slightly higher FDs compared to the mRNA, while PSEN2 CDSs FDs are lower. The correlation coefficient for these sequences is 0.969. A comparative study of differentially expressed MAPK signaling pathway lncRNAs in pancreatic cancer cells shows a correlation of 0.771. Selection of higher FD CDSs could indicate interaction of Alzheimer's gene products APP, APOE, and PSEN1. Including hypocretin sequences (where all CDSs have higher fractal dimensions than mRNA) in the APP, APOE, and PSEN1 sequence analyses improves correlation, but the inclusion of erythropoietin (where all CDSs have higher FD than mRNA) would suppress correlation, suggesting that HCRT, a hypothalamus neurotransmitter related to the wake/sleep cycle, might be better when compared to EPO, a glycoprotein hormone, for targeting Alzheimer's disease drug development. Fractal dimension and entropy correlation have provided supporting evidence, consistent with evolutionary studies, for using a zebrafish model together with a mouse model, in HCRT drug development.

1. Introduction

The instructions of a genetic sequence are carried by the fluctuations or variations in the nucleotide bases along the sequence. The bioinformatics of a sequence can be studied if the sequence is modeled as a series based on the nucleotide atomic number of the nucleotides A, T, C, and G. A recent study on such fluctuation in the FOXP2 gene pathway has been reported [1]. The fractal dimension and the Shannon entropy were found to have a negative correlation ($R^2 = 0.85$ $N = 12$) for the FOXP2 regulated "accelerated conserved-non-coding" sequences in human fetal brain. In general, fractal dimension and the Shannon entropy generate a 2D map representing a set of genetic sequences. For example, using the human Y chromosome, which contains 429 genetic sequences according to the http://ncbi.nlm.nih.gov/mapview/ database, the listed sequences have fractal dimension values from 1.92 to 2.06 and the Shannon dinucleotide entropy from 3.0 to 3.8 bits per base. Fractal dimension, being a nucleotide position sensitive measure, would be related to the richness in the embedded informatics associated with the sequence. In terms of transcription and translation, fractal dimension may be related to a docking energy parameter with similarity to the concept of roughness in a zipper and assembly traffic analogy for the docking interactions. In addition to CDS, mRNA sequences are often embedded with intronic regions. Noncoding RNA sequence with more than 200 base pair in length has been used to label a long non-coding RNA (lncRNA) regardless of intronic or intergenic in origin. This project uses fractal dimension of mRNA and coding DNA sequences (CDS) to probe the lncRNAs within the mRNAs (but not the CDS) in Alzheimer's disease genes, namely, APP or AD1, APOE or AD2, PSEN1 or AD3, and PSEN2

or AD4. The lncRNA sequences have been shown to be involved in significant regulatory functions. The lncRNA SPRY4-IT1 sequence was reported to have migrated to the cytoplasm, and upregulation was observed in melanoma cells [2]. Differential expressions in lncRNA upon radiation of HeLa and MCF-7 cells and in glioblastoma pathogenesis have been reported [3, 4]. The lncRNA relationship to cellular genetic product stability has been reported in the mouse model [5]. Microarray data analysis linked lncRNA to Huntington's disease [6], and the lncRNA role in neurological disorders and cancers has been reviewed [7, 8]. The lncRNAs becoming a new cancer diagnostic and therapeutic gold mine have been postulated [9]. Despite all these activities, only a few computation analysis results on the lncRNAs have been reported to our knowledge. A comparative study of lncRNAs with $3'$ untranslated regions ($3'$ UTRs) in protein coding RNA sequences has revealed parallel structure in the studied sequences, consistent with the presence of similar evolutionary constraints [10].

This project focuses on the study of disease related genetic sequences. The lncRNAs within the mRNA sequences in Alzheimer's disease genes, namely, APP or AD1, APOE or AD2, PSEN1 or AD3, and PSEN2 or AD4, have been analyzed in terms of fractal dimension computation and correlation analysis. The exploratory hypothesis that the lncRNA sequences embedded in a transcribed mRNA sequence would exhibit correlation in Alzheimer's disease genes has been studied in a comparative fractal dimension model of mRNA sequences versus coding DNA sequences (CDSs), which do not include the lncRNA sequences.

2. Materials and Methods

The data used in this study was downloaded from GenBank according to the following Gen-ID numbers. The studied human genes are APP-Gen-ID-351 containing 10 mRNA variants, APOE-Gen-ID-348, PSEN1-Gen-ID-5663 having 2 mRNA variants, PSEN2-Gen-ID-5664 having two mRNA variants, HCRT-Gen-ID-3060, HCRTR1-Gen-ID-3061 (HCRT Receptor-1), HCRTR2-Gen-ID-3062 (HCRT Receptor-2), EPO-Gen-ID-2056, and EPOR-Gen-ID-2057. The MAPK signaling pathway gene accession numbers have been listed in the report of differential expression of long non-coding intronic RNAs in pancreatic cancer cells [11]. The Allen Brain Atlas database has been accessed at http://brain-map.org.

A sequence with a relatively low nucleotide variety would have low Shannon's entropy (more constraints) in terms of the set of 16 possible dinucleotide pairs. A sequence's entropy can be computed as the sum of $(p_i) * \log(p_i)$ over all states i, and the probability p_i can be obtained from the empirical histogram of the 16 di-nucleotide pairs. The maximum entropy is 4 binary bits per pair for 16 possibilities (2^4). For mono-nucleotide consideration, the maximum entropy is two bits per mononucleotide with four possibilities (2^2). In general, the monoentropy is proportional to di-nucleotide entropy with $R^2 > 0.75$ for the reported sequences in the paper.

Roughly speaking, fractal dimension measures the complexity of a self-similar sequence. For a 1D sequence such as a DNA sequence, a fractal dimension near 2 indicates great complexity, while one closer to 1 would indicate little complexity, variety, or information. Among the various fractal dimension methods, the Higuchi fractal method is well suited for studying signal fluctuation [12]. A random spatial series with equal spatial steps can be modeled as a brightness signal in time such that the time series analysis tools can be used for spatial series analysis. The spatial intensity (Int) random series with equal intervals could be used to generate a difference series (Int(j) − Int(i)) for different lags in the spatial variable. The nonnormalized apparent length of the spatial series curve is simply $L(k) = \Sigma|(\text{Int}(j) - \text{Int}(i))|$ where the sum is for all pairs where j-$i = k$. The number of terms in a k-series varies, and normalization must be used to get the series length. If the Int(i) is a fractal function, then the $\log(L(k))$ versus $\log(1/k)$ should be a straight line with the slope equal to the fractal dimension. Higuchi incorporated a calibration division step such that the maximum theoretical value is calibrated to the topological value of 2. The details of the calculation method are given in the literature [12]. Numerical examples of the fractal dimension computation can be found in our earlier reports [13, 14]. For clarity, a Matlab implementation of the algorithm we used is listed below. "data" is an array loaded with the input sequences (one in each column). "width" is the number of sequences loaded in the data array. "max K" is cutoff for the maximum distance between j-i pairs. A max K of 7 was used for this paper, although other values for max K gave very similar results.

Consider

```
% calculate Length vectors for each column
L = zeros(max K, width);
for k = 1 : max K,
   data2 = circshift(data, k);
   data2 = abs(data2 − data);
   data2 (1 : k, :) = 0; % remove end effects
   L(k, :) = sum(data2)/k/k ∗ (height − 1)/(height − k);
end


% calculate slopes (FDs)
slope = zeros(1, width);
for i = 1 : width,
   temp = 1 : 1 : max K;
   log k = log (1./temp)′;
   X = [ones(size(log k)) log k];
   Y = log(L(:, i));
   a = X \ Y;
   FD(i) = a(2);
End
```

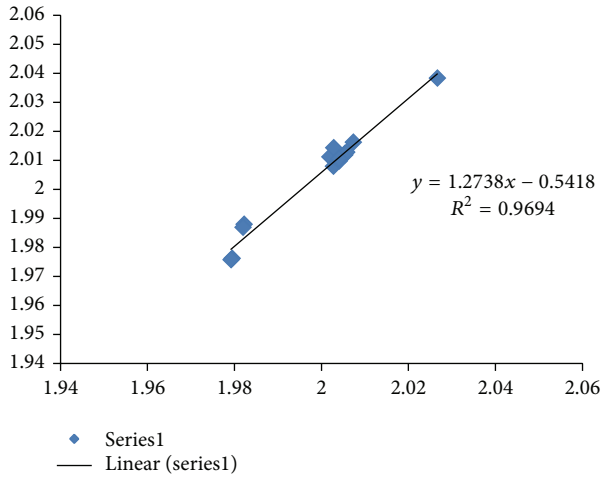


Figure 1: The fractal dimension correlation of the APP, APOE, PSEN1, and PSEN2 CDSs versus mRNAs in Alzheimer's disease is displayed with R^2 of 0.969 (N = 15 Series1). The y-axis represents the fractal dimension (FD) of the CDSs, and the x-axis represents the FD of the mRNAs. Deletion of PSEN2 Variant 1 (1.9793, 1.9748) and Variant 2 (1.9791, 1.9743) where CDSs have lower fractal dimension values as compared to the mRNA sequences, which would give R^2 of 0.979, N = 13.

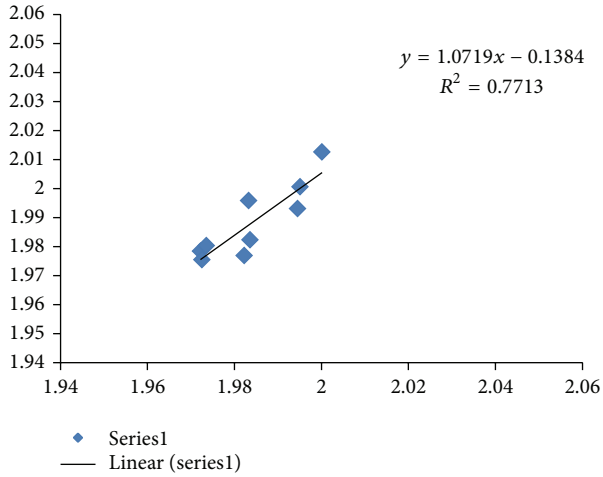


Figure 2: The fractal dimension correlation of the differentially expressed MAPK signaling pathway long noncoding intronic RNA in pancreatic cancer cells with a correlation of R^2 = 0.771 (N = 9 Series1). The y-axis represents the fractal dimension (FD) of the CDSs, and the x-axis represents the FD of the mRNAs. Deletion of MAP3K1 (1.9945, 1.9929), MAP3K14 (1.9822, 1.9768), and RAPGF2 (1.9835, 1.982) where CDSs have lower fractal dimension as compared to the mRNA sequences which would give R^2 of 0.950 (N = 6). The sequence GenBank accession numbers have been listed in [11].

3. Results of Fractal Analysis

The ratio of CDS length to mRNA length ranges from 0.23 to 0.78 in the studied Alzheimer's disease sequences. A negative correlation with R^2 of 0.61 was found for fractal dimension and the ratio of CDS length to mRNA length, using the 10 variant sequences in APP mRNA. The increase of fractal dimension with decreasing ratio could be related to some systematic properties in mRNA variant formation in the APP gene. It would appear that increasing the lncRNA length portion relative to the CDS length portion would correlate with increasing fractal dimension in the APP mRNA variants.

The fractal dimension correlation of the CDSs versus mRNAs in Alzheimer's disease is displayed in Figure 1 with R^2 of 0.969 (N = 15). All of the studied Alzheimer's disease sequences show higher fractal dimensions in the CDSs as compared to the mRNAs, except for the PSEN2 Variant 1 and Variant 2. Similarly, a comparative study of the differentially expressed MAPK signaling pathway of long non-coding intronic RNA in pancreatic cancer cells is displayed in Figure 2 with a correlation of R^2 = 0.771 [11]. The MAPK signaling pathway has been reported to involve 9 mRNAs in differential expression of long non-coding intronic RNA in [11]. They are ARRB1, ATF2, MAPK1, MAP2 K5, MAP3 K1, MAP3 K14, PPP3CB, RAPGF2, and TGFbR2. The CDSs of MAP3 K1, MAP3 K14, and RAPGF2 have lower fractal dimension values as compared to the mRNAs.

The systematic selection of higher fractal dimension CDSs could be indicative of certain characteristic interaction of the Alzheimer's gene products APP, APOE, and PSEN1 where a correlation with R^2 of 0.979 (N = 13) was obtained. Hypocretin (orexin) loss in Alzheimer's disease patients has been reported [15]. A brain scan study on a group of young adults has revealed 1/3 of them are PSEN1 E280A mutation carriers, an accepted hallmark for Alzheimer's disease [16]. The inclusion of hypocretin sequences (where all CDSs have higher fractal dimension values than mRNAs in HCRT, HCRT-R1, and HCRT-R2) in the APP, APOE, and PSEN1 sequence analysis would improve the correlation (R^2 = 0.985, N = 16) as shown in Figure 3. Erythropoietin EPO has been shown to have interaction with dopamine pathways [17–19] and offer protection for neuronal injury [20, 21]. The inclusion of erythropoietin (where all CDSs have higher dimension than mRNA in EPO and EPOR) in the correlation of APP, APOE, and PSEN1 would suppress the correlation (R^2 = 0.953, N = 15).

4. Discussion

The regression intercepts in Figures 1, 2, and 3 are negative, while the slopes are all greater than 1. This indicates selection pressure driving up CDS fractal dimension or a selection pressure against high FD in lncRNA. Whether the negative intercept value would suggest a minimum fractal dimension threshold in mRNA for containing a functional CDS in the studied set of Alzheimer's disease genes needs further investigation. The exploratory hypothesis that the lncRNA sequences embedded in the transcribed mRNAs would exhibit correlation in Alzheimer's disease genes receives supporting evidence in a comparative fractal dimension model of mRNA sequences versus coding DNA sequences (CDSs).

The correlation results suggest a hypothesis where HCRT, a neurotransmitter only produced in the hypothalamus and related to the wake/sleep cycle, could be a relatively more

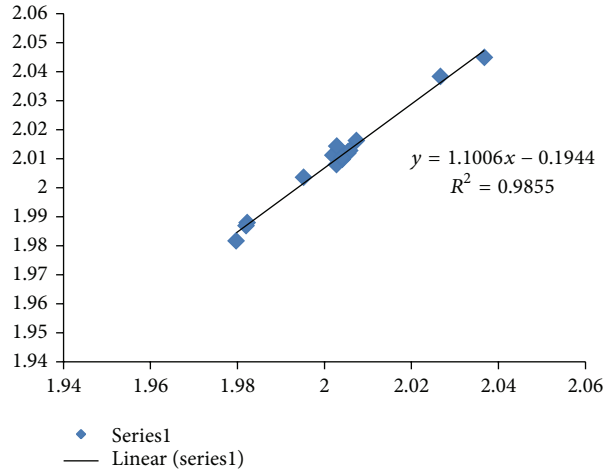


Figure 3: The fractal dimension correlation of the APP, APOE, PSEN1, HCRT, HCRT-R1, and HCRT-R2 CDSs versus mRNAs in Alzheimer's disease is displayed with R^2 of 0.985 ($N = 16$ Series1). The y-axis represents the fractal dimension (FD) of the CDSs, and the x-axis represents the FD of the mRNAs.

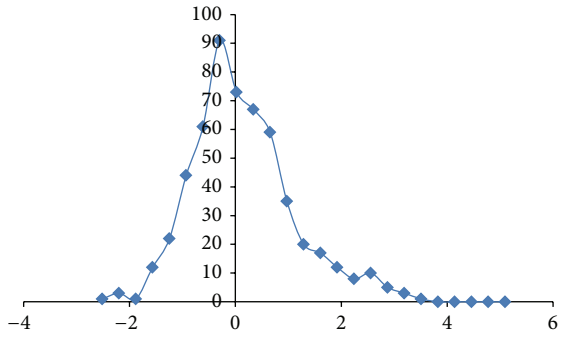


Figure 4: HCRT Receptor-2 expression level distribution in the brain regions (168 regions for each patient) using the 4-patient data from Allen Brain Atlas. The expression level z-score values are displayed in the x-axis. The displayed line is used as a visual guide. The Skewness of the distribution has been computed to be = 0.91.

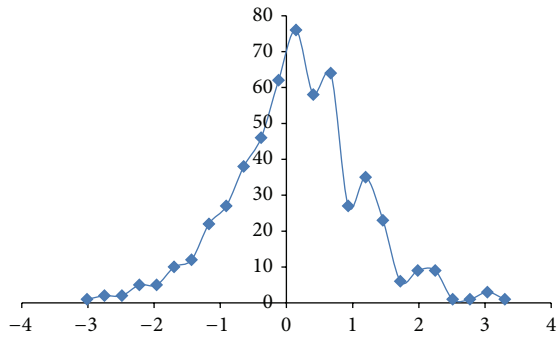


Figure 5: EPO Receptor expression level distribution in the brain regions (168 regions for each patient) using the 4-patient data from Allen Brain Atlas. The expression level z-score values are displayed in the x-axis. The displayed line is used as a visual guide. The Skewness of the distribution has been computed to be = 0.04.

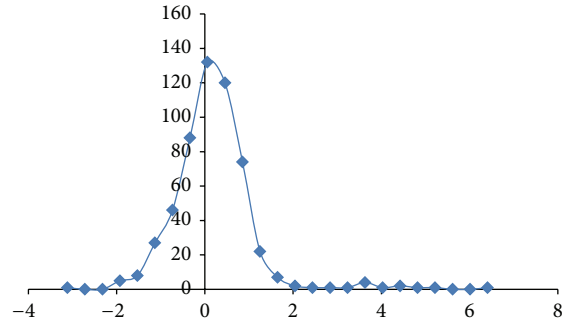


Figure 6: HCRT expression level distribution in the brain regions (168 regions for each patient) using the 4-patient data from Allen Brain Atlas. The expression level z-score values are displayed in the x-axis. The displayed line is used as a visual guide. The Skewness of the distribution has been computed to be = 2.2.

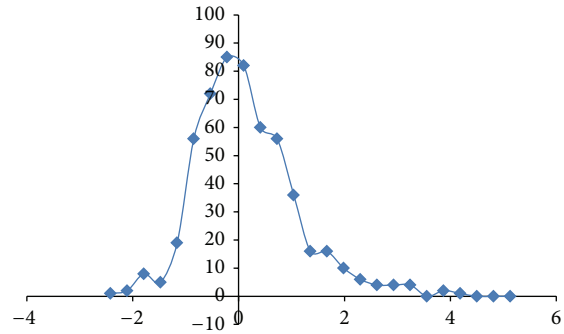


Figure 7: EPO expression level distribution in the brain regions (168 regions for each patient) using the 4-patient data from Allen Brain Atlas. The expression level z-score values are displayed in the x-axis. The displayed line is used as a visual guide. The Skewness of the distribution has been computed to be = 1.1.

important candidate as a blocker or promoter when compared to EPO, a glycoprotein hormone produced by kidney and liver, for targeting drug development with application to Alzheimer's disease clinical trials. The HCRT hypothesis would be consistent with MRI brain scans (168 regions) containing microarray array expression level data from the Allen Brain Atlas database. The brain scan data analysis has showed higher Skewness value in HCRT Receptor-2 expression level distribution (Figure 4) in the brain as compared to EPO Receptor distribution (Figure 5). The reasoning follows the fact that hypocretin Receptor expression level distribution would have a positive long tail representing high expression level and high demand for HCRT in the brain as compared to EPO receptor expression level distribution. The Skewness value would be 0.91 for HCRT Receptor-2 (Figure 4) versus 0.04 for EPO Receptor (Figure 5), a factor difference of about 20. The HCRT and EPO expression level distributions in the brain are displayed in Figures 6 and 7, respectively. The Skewness value would be 2.2 for HCRT (Figure 6) versus 1.1 for EPO (Figure 7), a factor difference of about 2. The large factor difference in receptor expression level in the brain would influence the selection of targeted receptors in drug development.

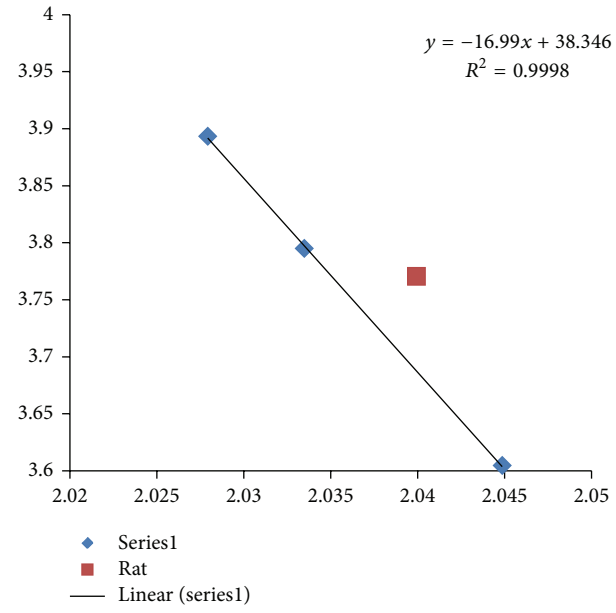


Figure 8: A plot of fractal dimension versus entropy for HCRT CDSs in human, mouse, and zebrafish is displayed with R^2 of 0.9998 and an adjusted R^2 of 0.9996 (Series1 diamonds). The fractal dimension is represented on the x-axis, and the dinucleotide entropy is represented on the y-axis. Rat HCRT CDS would be viewed as an outlier (square) from the regression analysis of human (highest fractal dimension), mouse and zebrafish (lowest fractal dimension). GenBank information: mouse Mus musculus HCRT has Gen-ID-15171, rat Rattus norvegicus HCRT has Gen-ID-25723, and zebrafish Danio rerio HCRT has Gen-ID-613239.

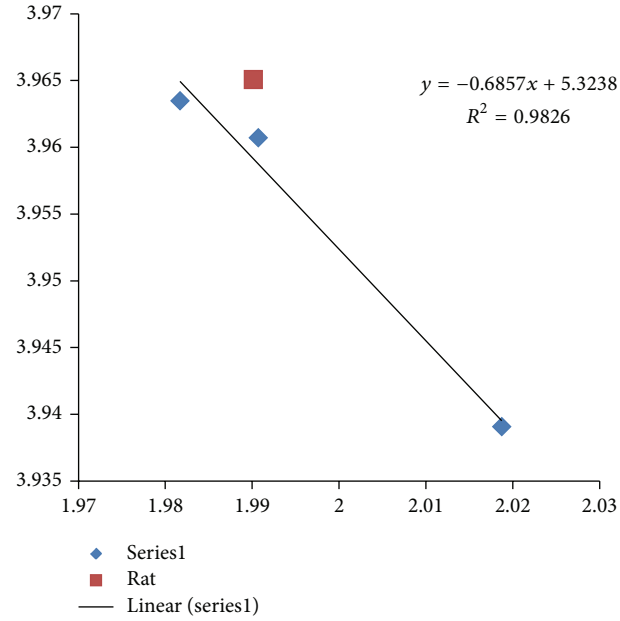


Figure 9: A plot of fractal dimension versus entropy for HCRT-R2 CDSs in human, mouse, and zebrafish is displayed with R^2 of 0.982 and an adjusted R^2 of 0.965 (Series1 diamonds). The fractal dimension is represented on the x-axis, and the dinucleotide entropy is represented on the y-axis. Rat HCRT-R2 CDS would be viewed as an outlier (square) from the regression analysis of human (lowest fractal dimension), mouse and zebrafish (highest fractal dimension). GenBank information: mouse Mus musculus HCRT has Gen-ID-387285, rat Rattus norvegicus HCRT has Gen-ID-25605, and zebrafish Danio rerio HCRT has Gen-ID-561260.

Mouse model has become a popular choice in drug development since evolution has been a corner stone for the understanding of biology. A plot of fractal dimension versus entropy for HCRT CDSs in human, mouse, and zebrafish is displayed in Figure 8 with an adjusted R^2 of 0.9996, and rat HCRT CDS would be viewed as an outlier from the regression analysis of human, mouse, and zebrafish. The regression result would be consistent with an evolutionary trend where the human HCRT has the highest fractal dimension, and zebrafish HCRT has the lowest fractal dimension. Similar fractal dimension entropy plot on HCRT-R2 is displayed in Figure 9 with an adjusted R^2 of 0.965, and rat HCRT-R2 CDS would be viewed as an outlier. The regression result would be consistent with an evolutionary trend, where the human HCRT-R2 has the lowest fractal dimension, and zebrafish HCRT-R2 has the highest fractal dimension. The high fractal dimension human HCRT combination of low fractal dimension human receptor HCRT-R2 would be consistent with a docking complimentary relationship as discussed above in terms of the significance of fractal dimension as an associated parameter for roughness matching in a zipper analogy for the understanding of transcription and translation. The folding of lncRNA could be important for further studies of docking and regulation in terms of sequence fractal dimension computation, and metal controlled folding RNA could serve as a starting platform with UV Circular Dichroism and Synchrotron based X-ray absorption spectroscopy structural data [22]. In any event, when a fractal dimension entropy map is used as a tool for evolutionary pressure study beyond simple classification, a strong correlation would lend quantitative support to the choice of mouse model and zebrafish model in HCRT drug development. Extension of fractal dimension computation and correlation analysis to the recently published dataset of expressed lncRNAs in zebrafish embryogenesis would help HCRT drug development [23].

Recently the R47H variant of TREM2 was reported to be associated with Late-onset Alzheimer's disease (LOAD) [24, 25]. The human TREM2 is known as an innate immune receptor and signals through TYROBP (agonist with 4 mRNA variants and 4 CDS variants) to clear the damaged tissue and reduce inflammation. The fractal dimension computation and correlation is displayed in Figure 10 with $R^2 = 0.999$, $N = 5$. The BLASTN comparison of TREM2 (lowest left corner data point in Figure 10) shows $E = 0.11$, given a CDS of 693-nucleotide sequence versus a 366-nucleotide sequence within the mRNA obtained by adding the beginning and ending non-coding regions together. Similar BLASTN comparison of TYROBP Variant 1 (uppermost right corner data point in Figure 10) had returned a null result, given a CDS of 342-nucleotide sequence versus a 266-nucleotide sequence within the mRNA obtained as described

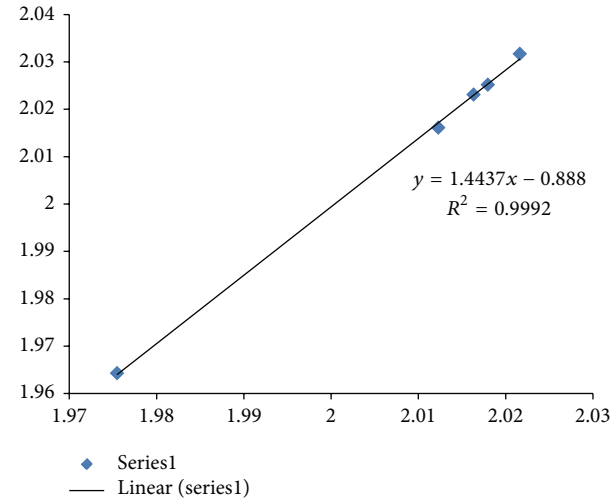


Figure 10: The fractal dimension correlation of the TYROBP-Variants 1–4 and TREM2 CDSs versus mRNAs involved in Late-onset Alzheimer's disease. $R^2 = 0.9992$ ($N = 5$). Inclusion of APOE (2.0267, 2.0382) would appear at the uppermost right corner in contrast to TREM2 at the lowest left corner and would give $R^2 = 0.9993$, $N = 6$ (adjusted $R^2 = 0.99915$, $N = 6$). For comparison, similar correlation analysis on the mouse and Bos taurus TYROBP, TREM2, and APOE would give R^2 values of 0.89 ($N = 3$), and 0.45 ($N = 3$) respectively. GenBank information: human TYROBP has Gen-ID-7305, human TREM2 has Gen-ID-54209, mouse TYROP has Gen-ID-22177, mouse TREM2 has Gen-ID-83433, mouse APOE has Gen-ID-11816, Bos taurus TYROP has Gen-ID-282390, Bos taurus TREM2 has Gen-ID-506467, and Bos taurus APOE has Gen-ID-281004.

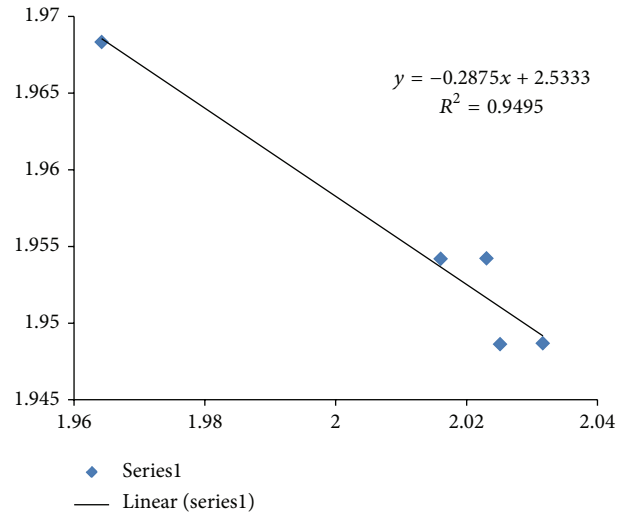


Figure 11: Fractal dimension versus entropy for TYROBP-Variant 1, TYROBP-Variant 2, TYROBP-Variant 3, TYROBP-Variant 4, and TREM2 CDSs involved in Late-onset Alzheimer's disease in human. The fractal dimension is represented on the x-axis, and the mononucleotide entropy is represented on the y-axis. TYROBP-Variant 1 has the highest fractal dimension and TYROBP-Variant 2 has the second lowest fractal dimension, and TREM2 has the lowest fractal dimension. Inclusion of APOE (2.0382, 1.8673) would suppress R^2 to 0.32.

above. The fact that the CDS and mRNA of the studied sequences have similar fractal dimension values but show little or no relationship under BLAST investigation would suggest a selection process, and the correlation showing R^2 of 0.9992 ($N = 5$) among the 4 variants and the receptor would suggest a systematic selection process, consistent with the CDSs entropy versus fractal dimension plot having R^2 of 0.949 ($N = 5$) in Figure 11.

As [24] pointed out, the apolipoprotein E (APOE) malfunction still remains as the most important sequence variant that would be risk of Late-onset Alzheimer's disease. The inclusion of APOE in Figure 10 would give R^2 of 0.9993 ($N = 6$), suggesting a very stringent regulation in selecting CDSs from mRNAs in the studied LOAD sequences. For comparison, similar correlation analysis on the mouse and Bos taurus TYROBP, TREM2, and APOE would give R^2 values of 0.89 ($N = 3$) and 0.45 ($N = 3$), respectively. The BLASTN comparison of human APOE had returned $E = 0.11$, given a CDS of 954-nucleotide sequence versus a 269-nucleotide sequence within the mRNA obtained by adding the beginning and ending non-coding regions together. The inclusion of HCRT and EPO informatics would suppress the correlation to R^2 values of 0.973 ($N = 9$) and 0.927 ($N = 8$), respectively, suggesting that HCRT drugs could be a better choice for treating Late-onset Alzheimer's disease as compared to EPO drugs. The inclusion of APOE would give R^2 of 0.32 in the entropy versus fractal dimension graph in Figure 11. The APOE sequence has the lowest entropy among the studied sequences, and all CDSs have lower entropy than mRNAs in the Late-onset Alzheimer's disease studied sequences. The APOE mononucleotide entropy of 1.8673 would suppress very slightly the mono-nucleotide correlation of mRNA versus CDS in Late-onset Alzheimer's disease studied sequences from R^2 of 0.9948 ($N = 5$) to 0.9944 ($N = 6$). The Late-onset Alzheimer's disease studied mRNAs and CDSs show very high correlation in fractal dimension (R^2 of 0.999) and entropy (R^2 of 0.994), consistent with a Late-onset Alzheimer's disease lncRNA hypothesis of high fractal dimension satisfied by low entropy in CDSs selection by deleting the lncRNAs with low entropy values about 1.91 except for TREM2 lncRNA having 1.995 bits per nucleotide type.

High correlation results are also observed in two other neurodegenerative disease involving TYROBP. The Nasu-Hakola disease, a disorder affecting both brain and bone, is known to be related to the malfunctioning of TREM2 or TYROBP [26]. The CSF1R (a microglial receptor) where malfunctioning is associated with a corticobasal syndrome called hereditary diffuse leukoencephalopathy with spheroids was reported to be cosignaling with TYROBP [27]. The addition of CSF1R in Figure 10 would reduce the correlation from 0.999 ($N = 5$) to 0.992 ($N = 6$). The exploratory study of noncoding RNA by comparing mRNA versus CDS informatics has revealed regularity. An examination of Figure 1 with R^2 0.969 ($N = 15$) and Figure 10 with R^2 0.999 ($N = 6$ including APOE) would suggest that the noncoding RNA assembly process in Late-onset Alzheimer's disease would involve relatively highly systematic process or

processes as compared to familial early-onset Alzheimer's disease with APP, PSEN1, PSEN2, and APOE. The project has used R^2 differences in the order of 0.02 to be the demarcation in the lncRNA investigation by comparing mRNAs versus CDSs in a cluster of disease related sequences. The mRNA versus CDS informatics comparative method could be a supplement to the well-accepted BLAST method. Further investigations using fractal analysis for neurodegenerative disease sequences and currently targeted receptors would be productive.

5. Conclusions

The long noncoding RNAs (lncRNAs) within the mRNA sequences in Alzheimer's disease genes, namely, APP, APOE, PSEN1, and PSEN2, have been analyzed in terms of fractal dimension computation and correlation analysis. The results show that APP, APOE, and PSEN1 CDSs select slightly higher fractal dimensions as compared to the mRNA sequences with a pattern evidenced by correlation coefficient of $R^2 = 0.979$ ($N = 13$ including variants). Inclusion of the 2 variants in PSEN2 where CDSs have lower fractal dimension values than mRNAs would yield R^2 of 0.969 ($N = 15$). The systematic selection of higher fractal dimension CDSs could be indicative of characteristic interaction of Alzheimer's gene products APP, APOE, and PSEN1. The inclusion of hypocretin sequences would improve the correlation ($R^2 = 0.985$, $N = 16$) but inclusion of erythropoietin would suppress the correlation ($R^2 = 0.953$, $N = 15$), suggesting that HCRT could be a relatively more important candidate as a blocker or promoter when compared to EPO for targeting in drug development with application to Alzheimer's disease clinical trials. The HCRT hypothesis would be consistent with MRI brain scan containing microarray expression level data from the Allen Brain Atlas database that shows higher Skewness value in HCRT receptor expression level distribution as compared to EPO receptor expression level distribution in the brain. Study of sequence fractal dimension and entropy correlation has provided quantitative supporting evidence, consistent with evolutionary studies, for using zebrafish model together with mouse model, in HCRT drug development. The TREM2 and TYROBP mRNAs reported recently in Late-onset Alzheimer's disease also yield a correlation of R^2 of 0.999 ($N = 6$) using similar informatics analysis, but HCRT informatics inclusion would suppress the correlation slightly as compared to the EPO informatics inclusion.

Acknowledgments

The project was partially supported by CUNY research Grant J. Ye thanks the NSF-REU program for student support. E. Cheung and S. Dehipawala thank QCC Physics Department for the hospitality. The authors thank Dr. N. Le of Thomas Jefferson Medical School Pathology Department, 1020 Locust Street, Philadelphia, PA 19107, USA for discussion. The authors, thank the research groups cited in the paper for posting their data and software in the public domain.

References

[1] G. Tremberger Jr., S. Dehipawala, E. Cheung et al., "Fractal analysis of FOXP2 regulated accelerated conserved non-coding sequences in human fetal brain," *World Academy of Science, Engineering and Technology*, no. 67, pp. 881–886, 2012.

[2] D. Khaitan, M. E. Dinger, J. Mazar et al., "The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion," *Cancer Research*, vol. 71, no. 11, pp. 3852–3862, 2011.

[3] E. Ozgür, U. Mert, M. Isin, M. Okutan, N. Dalay, and U. Gezer, "Differential expression of long non-coding RNAs during genotoxic stress-induced apoptosis in HeLa and MCF-7 cells," *Clinical and Experimental Medicine*, 2012.

[4] L. Han, K. Zhang, Z. Shi et al., "LncRNA profile of glioblastoma reveals the potential role of lncRNAs in contributing to glioblastoma pathogenesis," *International Journal of Oncology*, vol. 40, no. 6, pp. 2004–2012, 2012.

[5] M. B. Clark, R. L. Johnston, M. Inostroza-Ponta et al., "Genome-wide analysis of long noncoding RNA stability," *Genome Research*, vol. 22, no. 5, pp. 885–898, 2012.

[6] R. Johnson, "Long non-coding RNAs in Huntington's disease neurodegeneration," *Neurobiology of Disease*, vol. 46, no. 2, pp. 245–254, 2011.

[7] C. N. Niland, C. R. Merry, and A. M. Khalil, "Emerging roles for long non-coding RNAs in cancer and neurological disorders," *Frontiers in Genetics*, vol. 3, article 25, 2012.

[8] V. A. Moran, R. J. Perera, and A. M. Khalil, "Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs," *Nucleic Acids Research*, vol. 40, no. 14, pp. 6391–6400, 2012.

[9] P. Qi and X. Du, "The long non-coding RNAs, a new cancer diagnostic and therapeutic gold mine," *Modern Pathology*, vol. 26, pp. 155–165, 2012.

[10] F. Niazi and S. Valadkhan, "Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs," *RNA*, vol. 18, no. 4, pp. 825–843, 2012.

[11] A. C. Tahira, M. S. Kubrusly, M. F. Faria et al., "Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer," *Molecular Cancer*, vol. 10, no. 1, 19 pages, 2011.

[12] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D*, vol. 31, no. 2, pp. 277–283, 1988.

[13] T. Holden, R. Subramaniam, R. Sullivan et al., "ATCG nucleotide fluctuation of deinococcus radiodurans radiation genes," *Proceedings of SPIE*, vol. 6694, Article ID 669417, 10 pages, 2007.

[14] T. Holden, G. Tremberger Jr., P. Marchese et al., "DNA sequance based comparative studies of between non-extremophile and extremophile organisms with implications in exobiology," *Proceedings of SPIE*, Article ID 70970Q, 12 pages, 2008.

[15] R. Fronczek, S. van Geest, M. Frölich et al., "Hypocretin (orexin) loss in Alzheimer's disease," *Neurobiology of Aging*, vol. 33, no. 8, pp. 1642–1650, 2012.

[16] Eric M. Reiman, Y. T. Quiroz, A. S. Fleisher et al., "Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant Alzheimer's disease in the presenilin 1 E280A kindred: a case-control study," *Lancet Neurology*, vol. 11, no. 12, pp. 1048–1056, 2012.

[17] H. H. Marti, "Erythropoietin and the hypoxic brain," *The Journal of Experimental Biology*, vol. 207, part 18, pp. 3233–3242, 2004.

[18] N. M. Kanaan, T. J. Collier, D. M. Marchionini, S. O. McGuire, M. F. Fleming, and C. E. Sortwell, "Exogenous erythropoietin provides neuroprotection of grafted dopamine neurons in a rodent model of Parkinson's disease," *Brain Research*, vol. 1068, no. 1, pp. 221–229, 2006.

[19] M. Alnaeeli, L. Wang, B. Piknova, H. Rogers, X. Li, and C. T. Noguchi, "Erythropoietin in brain development and beyond," *Anatomy Research International*, vol. 2012, Article ID 953264, 15 pages, 2012.

[20] A. L. Sirén, M. Fratelli, M. Brines et al., "Erythropoietin prevents neuronal apoptosis after cerebral ischemia and metabolic stress," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 7, pp. 4044–4049, 2001.

[21] P. Villa, P. Bigini, T. Mennini et al., "Erythropoietin selectively attenuates cytokine production and inflammation in cerebral ischemia by targeting neuronal apoptosis," *Journal of Experimental Medicine*, vol. 198, no. 6, pp. 971–975, 2003.

[22] S. S. Athavale, A. S. Petrov, C. Hsiao et al., "RNA folding and catalysis mediated by iron (II)," *PLoS One*, vol. 7, no. 5, Article ID e38024, 7 pages, 2012.

[23] A. Pauli, E. Valen, M. F. Lin et al., "Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis," *Genome Research*, vol. 22, no. 3, pp. 577–591, 2012.

[24] T. Jonsson, H. Stefansson, S. Steinberg et al., "Variant of TREM2 associated with the risk of Alzheimer's disease," *The New England Journal of Medicine*, vol. 368, no. 2, pp. 107–116, 2013.

[25] R. Guerreiro, A. Wojtas, J. Bras et al., "TREM2 variants in Alzheimer's disease," *The New England Journal of Medicine*, vol. 368, no. 2, pp. 117–127, 2013.

[26] H. Neumann and M. J. Daly, "Variant TREM2 as risk factor for Alzheimer's disease," *The New England Journal of Medicine*, vol. 368, no. 2, pp. 182–184, 2013.

[27] D. W. McVicar and G. Trinchieri, "CSF-1R, DAP12 and beta-catenin: a ménage à trios," *Nature Immunology*, vol. 10, no. 7, pp. 681–683, 2009.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 486125, 12 pages
http://dx.doi.org/10.1155/2013/486125

Research Article

patGPCR: A Multitemplate Approach for Improving 3D Structure Prediction of Transmembrane Helices of G-Protein-Coupled Receptors

Hongjie Wu,[1,2,3] Qiang Lü,[1,2] Lijun Quan,[1] Peide Qian,[1,2] and Xiaoyan Xia[1,2]

[1] *School of Computer Science and Technology, Soochow University, Suzhou 215006, China*
[2] *Jiangsu Provincial Key Lab for Information Processing Technologies, Suzhou 215006, China*
[3] *School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China*

Correspondence should be addressed to Qiang Lü; qiang@suda.edu.cn

Received 5 November 2012; Revised 10 January 2013; Accepted 16 January 2013

Academic Editor: Hong-Bin Shen

The structures of the seven transmembrane helices of G-protein-coupled receptors are critically involved in many aspects of these receptors, such as receptor stability, ligand docking, and molecular function. Most of the previous multitemplate approaches have built a "super" template with very little merging of aligned fragments from different templates. Here, we present a parallelized multitemplate approach, patGPCR, to predict the 3D structures of transmembrane helices of G-protein-coupled receptors. patGPCR, which employs a bundle-packing related energy function that extends on the RosettaMem energy, parallelizes eight pipelines for transmembrane helix refinement and exchanges the optimized helix structures from multiple templates. We have investigated the performance of patGPCR on a test set containing eight determined G-protein-coupled receptors. The results indicate that patGPCR improves the TM RMSD of the predicted models by 33.64% on average against a single-template method. Compared with other homology approaches, the best models for five of the eight targets built by patGPCR had a lower TM RMSD than that obtained from SWISS-MODEL; patGPCR also showed lower average TM RMSD than single-template and multiple-template MODELLER.

1. Introduction

G-protein-coupled receptors (GPCRs) are among the most heavily investigated drug targets in the pharmaceutical industry [1] because activated GPCRs trigger a cascade of responses inside the cell. Although about 800 GPCRs in the human genome still await determining, the annual revenue in the market for human therapeutics based on the currently available receptors is in excess of $40 billion, and more than 50% of modern drugs are related to GPCRs [2]. On the other hand, it is still a very difficult problem to determine the conformation of GPCRs in vivo. The lipid environment in which the receptors are embedded blocks the two major techniques, nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography, that are used to determine protein structures. It is really exciting that the Nobel Prize in Chemistry was awarded in 2012 to two researchers studying the structure of GPCRs.

Fortunately, as demonstrated by recent publications [3, 4], in silico methods for deducing the three-dimensional structure of GPCRs have been increasingly successful. However, the development of computational approaches to predicting the structure of GPCRs remains a challenging task [5]. A lot of effort has been put into modeling the structures of the full-chain and the loop sections of GPCRs [6–8] and of membrane proteins [9, 10], whereas comparatively little research has been done on building more accurate models of the transmembrane helix sections of GPCRs, because the transmembrane helical bundles are commonly regarded as a conserved domain that can be easily duplicated

from templates. In fact, the accuracy of the models of the transmembrane helix sections is still far away from the native structures, and it cannot meet the requirements for subsequent full-chain prediction and the modeling of ligand docking. For the GPCR Dock 2010 assessment [3] targets CXCR4/CVX15, CXCR4/IT1t, and D3, the averages of the TM RMSD values (the root mean square deviation of the backbone of the transmembrane helices) of all models submitted by the participants were 3.56 Å, 9.75 Å, and 2.55 Å, respectively, which are not high-resolution values (<2.0 Å). How can one build high-resolution models of the conformations of full-length GPCRs without an accurate transmembrane helical bundle? This paper addresses this problem using a parallelized multitemplate homology approach.

The classification of methods for the prediction of the three-dimensional structure of GPCRs in silico into homology modeling, threading, and ab initio folding follows the classification of methods for protein structure prediction. The homology method builds models starting from one or more template structures with a high sequence similarity. When the sequence identity between target and template is more than 50%, near-native models tend to be generated. When it is less than 30%, the accuracy of the models decreases sharply. SWISS-MODEL [11], Sybyl [12], Prime [13], MODELLER [14], NEST [15], SEGMOD/ENCAD [16], 3D-JIGSAW [17], and Builder [18] are widely used, stable, reliable, and accurate systems for homology modeling. The threading method operates by "threading" (i.e., placing and aligning) each amino acid (or amino acid segment) in the target sequence into a position in the three-dimensional structure and evaluating how well the target fits the template. Zhang et al. [6] used TASSER to generate structure predictions for all 907 putative GPCRs in the human genome. Based on a benchmarked confidence score, approximately 820 of the predicted models should have the correct folds. Ab initio prediction of protein structure involves modeling the dihedral angles for each residue based on the minimum-free-energy principle, without the use of any experimentally solved structures. Baker's group [19, 20] was successful in the recent CASP (the Critical Assessment of protein Structure Prediction) challenge and designed a membrane-environment-related energy function to guide membrane protein folding.

Homology approaches can be divided into two categories, namely, single-template and multiple-template, depending on the number of templates employed in modeling. Single-template homology approaches cannot always achieve the best results, owing to difficulties in detecting the best template, particularly when remote homology is detected [21]. Multitemplate approaches can effectively combine more reasonably aligned regions than the single-template approach Cheng [22] reported that a multitemplate combination algorithm improved the GDT-TS scores of the predicted models by 6.8% on average for 45 CASP7 comparative-modeling targets. Liu et al. [23] took into account the information represented by multiple templates and alignments at the three-dimensional level by mixing and matching regions between different initial comparative models, and the multitemplate approach produced conformational models of higher quality than the individual starting predictions. MODELLER [14] modeled 3D conformations by optimally satisfying spatial restraints derived from the alignment and expressed as probability density functions for the features restrained. NEST [15] produced a model by taking a sequence alignment of a target to one or multiple template PDB files as input. 3D-JIGSAW [17] used a convergence of algorithms for comparative modeling which led to more reliable structures by superimposing multiple known structures from a protein family. In our opinion, previous multitemplate approaches have generally built a "super" template, which might ignore the flexibility of aligned structures from different templates. The long-distance homology information from different templates should be exchanged in each iteration rather than directive merging structures.

This paper proposes a parallelized multitemplate approach, inspired by our previous method pacBackbone [24, 25], for the prediction of the three-dimensional structure of the transmembrane helices of GPCRs. The system proposed here is referred to as "patGPCR" (parallelized multitemplate approach to GPCR transmembrane helix structure prediction). Parallelization not only accelerates the running speed but also provides a novel and effective mechanism to exchange homology regions between templates softly. We have exploited our method to predict tertiary structures for all eight determined GPCRs published on the GPCR Network website (http://gpcr.scripps.edu/index.html). Compared with other homology approaches, the best models for five of the eight targets built by patGPCR had a lower TM RMSD than was obtained from SWISS-MODEL, patGPCR showed lower average TM RMSD than single-template MODELLER, and patGPCR showed only one higher average TM RMSD target compared with multiple-template MODELLER.

2. Materials and Methods

2.1. Parallelized Framework of patGPCR. GPCRs share a similar structural topology, composed of seven transmembrane (TM) helices packed into a 7-TM helical bundle, with three intracellular (icls) and three extracellular loops (ecls) [26]. Thus, single helical refinement should be paralleled in independent threads. The parallelized framework for patGPCR is depicted in Figure 1. At the beginning of patGPCR, top 2–4 templates were examined for sequence identities using the Protein Data Bank (http://www.rcsb.org/pdb), which were used as starting template conformations. Eight parallelized pipelines which involve independent subprocedures (or threads) were used to randomly select starting conformations of the templates. Each pipeline containing TM refinement and loop refinement optimized an adjacent helix pair region. The first and the last pipelines optimized only one helix and one terminus. To use reasonable structural regions within different templates, TM helix crossing step and elite pool were introduced into the parallelized framework at the end of the pipelines. In the crossing step, each pipeline shared the best-so-far helix with other pipelines. If lower energy was obtained by helix crossing, the helix substitution was accepted; otherwise, it was rejected. Optimized conformations are conserved in the elite pool prepared for the next

(1) Input: P is starting conformation, $TM[]$ is the set of residue numbers for TM regions, $Loop[]$ is the set of residue numbers for loop regions.
Output: conformation after TM refined
E_m is the Rosetta membrane energy function with S.
E_{score3} is the Rosetta score3 energy function.
(2) $\vec{d}_{TM}[] = getHelixAxis()$;
(3) $\vec{d}_M = getMembrane()$;
(4) **for** $i = 1$ to $STAGE2\_CYCLES$ **do**
(5) $P' = transferHelix(P, TM[], RND(-5, 5), RND(-5, 5), RND(-5, 5))$;
(6) if $(E_m(P') < E_m(P))$ $P \leftarrow P'$;
(7) **end for**
(8) **for** $i = -180$ to 180 **do**
(9) $P' = spinHelix(P, TM[], \vec{d}_{TM}[], i)$;
(10) if $(E_m(P') < E_m(P))$ $P \leftarrow P'$;
(11) **end for**
(12) **for** $i = 1$ to $STAGE4\_CYCLES$ **do**
(13) $P' = tiltGaussianHelix(P, TM[], \vec{d}_M, 30, 6)$;
(14) if $(E_m(P') < E_m(P))$ $P \leftarrow P'$;
(15) **end for**
(16) $LoopModelRandomly(Loop[], E_{score3})$
(17) $P = crossHelices(P, TM[])$;
(18) Output: P;

Algorithm 1: Refinement by single pipeline (P, $TM[]$, $Loop[]$).

iteration. Thus, the crossing step and elite pool are two critical mechanisms of patGPCR to identify reasonable structures from multiple templates to pipelines and iterations.

2.2. Multiple Templates for Eight GPCR Targets. patGPCR was evaluated by using blind prediction testing the set of the eight determined GPCRs published on the GPCR network. Amino acid sequence was the only input used for blind prediction, which is commonly employed in GPCRDock2008/2010 and CASP exercises. This is the largest set used for directly evaluating prediction results. patGPCR employed 2–4 templates (column 4 in Table 1) for each target in the test set. The top three or four templates were selected by standard protein BLAST based on default parameters. Most sequence similarities between the templates and targets were lower than 50% and average sequence similarity for the templates was 36% (column 6 in Table 1). In some cases, the templates have high-sequence identity. There are two reasons why we did not remove these cases from benchmark. First, we are interested in validating the modeling ability of patGPCR based on various templates with different sequence identities. Second, patGPCR, SWISS-MODEL, and MODELLER used the same templates in the comparing experiments, so the results we presented are fair for these methods, no matter high- or low-sequence identity the templates is.

2.3. TM Refinement Protocol. The 7-TM helical bundle is the primary topology of GPCRs, which comprises approximately 75% of amino acids in the entire protein chain. The TM helix has been conserved throughout evolution [5]. However, in a recent GPCRDock2008/GPCRDock2010 study, average TM RMSDs of CXCR4/CVX15, CXCR4/IT1t, and D3 were 3.56 Å, 9.75 Å, and 2.55 Å. The accuracy of predicting TM regions can be improved, and the correct TM position and orientation ensure that loop regions are properly oriented. TM refinement protocols employed using patGPCR are based on 7-TM geometrical topology reflecting the bundle structure using a set of geometrical parameters. Topologically, each TM helix is regarded as a rigid body, and relative positions of internal atoms remain fixed when moving or rotating the rigid body.

TM refinement is divided into four stages. In the first stage, patGPCR was used to identify TM boundaries by averaging the results of six existing methods: TopPred [27], UniProt [28], TMpred [29], HMMTOP [30], TMHMM [31], and OCTOPUS [32]. In the second stage, translation of each rigid body along, or perpendicular to, the axis of the helix was used to optimize the relative positions of seven helices. In the third and fourth stages, spin angels and tilt angles were refined, respectively.

2.4. Loop Refinement Protocol. patGPCR, which involves an entirely predictive approach for GPCRs, combines two Rosetta loop modeling protocols. Due to high flexibility in loop regions, ab initio methods typically involve calculation of possible loop conformations with the help of various energy functions and minimizations [33–35]. patGPCR paralleled two Rosetta loop modeling protocols, including CCD [36] and KIC [37] modeling with Rosetta energy function *score*3. Each pipeline was used to randomly choose a loop movement to refine the loop section.

2.5. The Algorithm Executed by Single Pipeline. Algorithm 1 executed by single pipeline contains TM and loop refinement

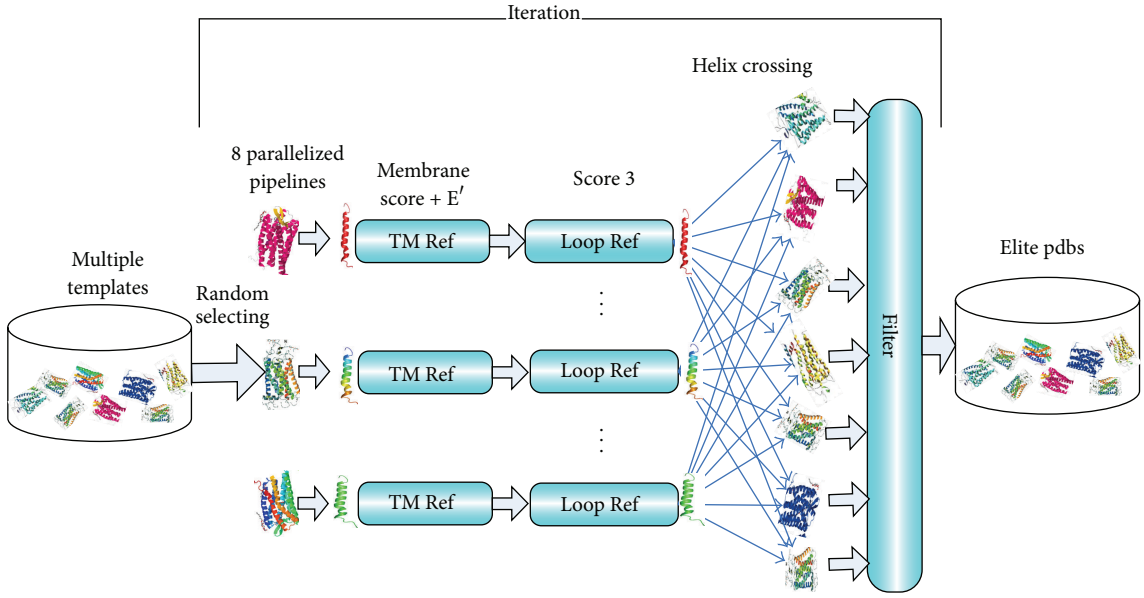


Figure 1: Parallelized framework of patGPCR.

at lines 2–15 and line 17, respectively. In the 2nd line, function *getHelixAxis*() gets the direction vector of the axis of the helix $\overrightarrow{d_{TM}}[]$. In the 3rd line, function *getMembrane*() gets the normal vector of the membrane plane $\overrightarrow{d_M}$. The 4th–7th lines execute second TM refinement stage and function $transferHelix(P, TM[], RND(-5, 5), RND(-5, 5), RND(-5, 5))$ randomly translates the helix from current position along x-, y-, or z-axis ranging from −5 Å to 5 Å. The new helix position would be accepted if the new energy E_m is lower than the energy before translating. The 8th–11th lines execute the third TM refinement stage and function $spinHelix(P, TM[], \overrightarrow{d_{TM}}[], i)$ spins the helix from −180 to 180 degrees and validates the new position using energy function E_m. The 12th–15th lines execute the fourth TM refinement stage and function $tiltGaussianHelix(P, TM[], \overrightarrow{d_M}, 30, 5)$ samples the tilt angles between the helix and the normal plane of transmembrane according to gaussian distribution (expected value is 30 and the variance is 6). In the 16th line, function *LoopModelRandomly*() refines the loop regions using Rosetta KIC or CCD remodel protocol randomly. In the 17th line, function *crossHelices*() exchanges the helixes among pipelines.

2.6. Energy Item for Evaluating the Helical Bundles. Developing an appropriate energy function remains a challenge in predicting GPCR 3D structures. An accurate energy function can be used to distinguish near-native models from candidates, while an imprecise energy function may not recognize near-native models even if they have been sampled by using an accurate search algorithm. patGPCR improves the Rosetta membrane energy function [19] by including a novel energy item E' for evaluating rigid helix packing. After

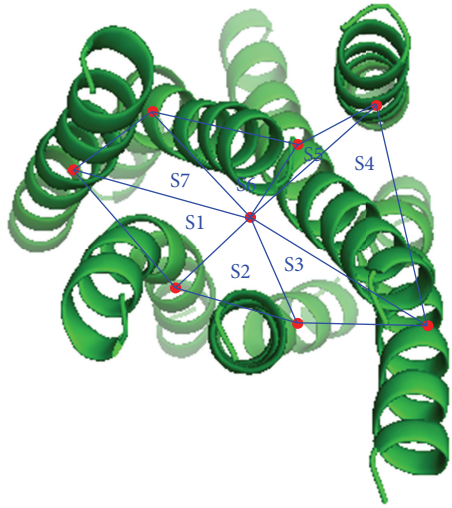


Figure 2: Parallelized framework of patGPCR.

projecting helices onto the membrane plane, Figure 2 was constructed to show $S1$–$S7$, which is the area of triangles constituted by the center point O and two intersection points of adjacent helix axes with the membrane plane. In (1), S is the sum of $S1$–$S7$ and $S_{\min}$ and $S_{\max}$ are the maximum and minimum values of S in known structures of GPCRs. A smaller E' indicates a tighter 7-TM helical bundle, while a greater E' indicates a looser bundle. Detection of collisions between residues is included in the Rosetta membrane energy function. Thus, the Rosetta membrane energy function which

Table 1: Eight determined GPCR targets and their templates.

Target names	Full names	No. of MT pat[a]	Templates	Chain	SS[b]	Avg ST pat[c]	Avg MT pat[d]	SWISS[e]	Avg ST MOD[f]	Avg MT MOD[g]
CXCR4	CXCR4	932	1GZM	A	24%	6.34	4.26 ± 1.40	√	7.69	11.4 ± 0.16
			2KS9	A	30%	4.33		√		
			2RH1	A	31%	5.77		×		
KOR1	kappa opioid receptor	1218	4AMJ	A	27%	3.23	2.98 ± 0.22	√	3.37	7.11 ± 0.20
			4EIY	A	56%	3.40		√		
HH1R	Histamine receptor H1	1533	2Y03	A	36%	1.56	12.17 ± 10.65	√	3.87	17.32 ± 0.24
			3SN6	R	47%	20.64		√		
			4AMJ	A	33%	11.17		√		
D3	D3 dopamine receptor	1776	2RH1	A	44%	2.64	1.69 ± 0.81	√	1.75	2.61 ± 0.16
			2Y00	A	34%	3.40		√		
			2VT4	A	36%	3.20		√		
A2A	adenosine A2a receptor	2076	1U19	A	31%	4.35	3.53 ± 21.52	√	3.85	7.88 ± 0.30
			2RH1	A	44%	26.58		√		
			2VT4	A	35%	4.16		√		
			2Z73	A	23%	5.80		√		
S1P1	Sphingosine 1-phosphate receptor 1	1031	2RH1	A	27%	4.39	6.67 ± 2.69	√	10.83	23.71 ± 0.77
			4EIY	A	54%	5.43		√		
			4AMJ	A	29%	5.52		√		
KOR3	nociceptin opioid receptor	1409	2RH1	A	29%	5.00	3.41 ± 1.67	√	8.71	16.82 ± 0.43
			3SN6	R	35%	7.73		×		
			4AMJ	A	37%	4.29		√		
Beta2	beta-2-adrener gic receptor	895	3EML	A	33%	5.60	3.82 ± 2.60	×	1.36	1.54 ± 0.28
			4AMJ	A	64%	2.19		√		
Avg		1358.75			36%	6.38	4.81 ± 5.20		5.11	11.05 ± 0.32

[a]Number of conformations generated by MT (multiple-template) patGPCR.
[b]Sequence similarity.
[c]Average TM RMSD of conformations generated by ST (single-template) patGPCR.
[d]Average TM RMSD of conformations generated by MT (multiple-template) patGPCR.
[e]SWISS output.
[f]Average TM RMSD of conformations generated by ST (single-template) MODELLER.
[g]Average TM RMSD of conformations generated by MT (multiple-template) MODELLER.

Table 2: Comparison with GPCRDocking2010 participants.

Number	Group name	CXCR4 TM RMSD	Group name	D3 TM RMSD
1	UMich-Pogozheva	2.05	UMich-Zhang	1.26
2	UMich-Zhang	2.08	Monash-Hall	1.35
3	VU-MedChem	2.14	WUStL	1.35
4	Monash-Sexton-1	2.18	Monash-Sexton-2	1.37
5	UMich-Zhang	2.18	WUStL	1.37
6	patGPCR	2.165	patGPCR	0.930

includes an additional item E' is employed by the patGPCR TM refinement algorithm:

$$E' = \begin{cases} \dfrac{(S - S_{\min})^2}{S_{\min}^2}, & S < S_{\min}, \\ 0, & S_{\min} < S < S_{\max}, \\ \dfrac{(S - S_{\max})^2}{S_{\max}^2}, & S < S_{\max}, \end{cases} \quad (1)$$

3. Results

3.1. The Contribution of TM Refinement Protocol Employed by Single Pipeline. The contributions of TM refinement protocols were investigated on the comparison with models submitted by GPCRDOCK2010 participants. GPCRDOCK2010 exercise is a community-wide assessment for GPCR homology-modeling and docking. The participants submitted the best five candidates of targets CXCR4 and D3. We executed patGPCR to optimize the helical bundles starting from the submitted conformations downloaded from GPCRDOCK2010 official website. The TM RMSD of the best model after refining and top 5 models published on GPCRDOCK2010 are listed in Table 2. For CXCR4 (D3) target, the TM RMSD of the best model after refining is 2.165 Å (0.93 Å), which ranks 4th (1st) among 161 (168) submitted models.

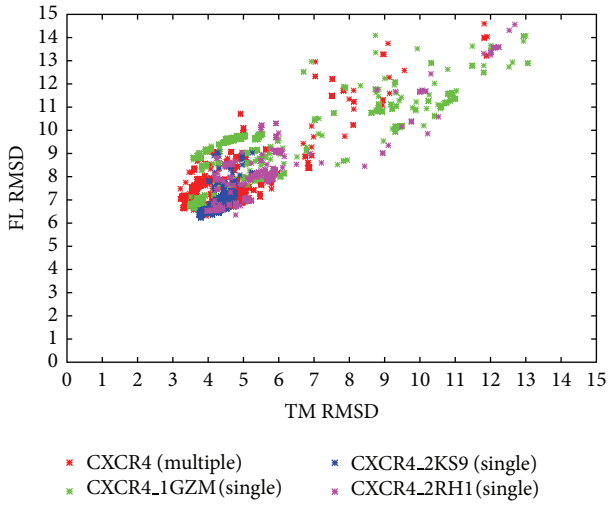
3.2. The Contribution of Parallelized Multitemplate. We examined the contributions of multiple-templates versus a single-template of patGPCR with the same settings. The number of templates used by the patGPCR ranged from two to four. Average of sequence identities between the templates and targets was 36%. Compared to multitemplate patGPCR, single-template patGPCR employed only one template and did not utilize the helix crossing step.

Full-length RMSD (FL RMSD) of GPCR and TM RMSD comparisons are plotted in Figure 3; dots with higher RMSD values (>15 Å) are not included. The average improvement in TM RMSD using the multiple-template patGPCR was 33.64% (raw TM RMSD increase 1.57 Å, columns 4 and 8 in Table 1).

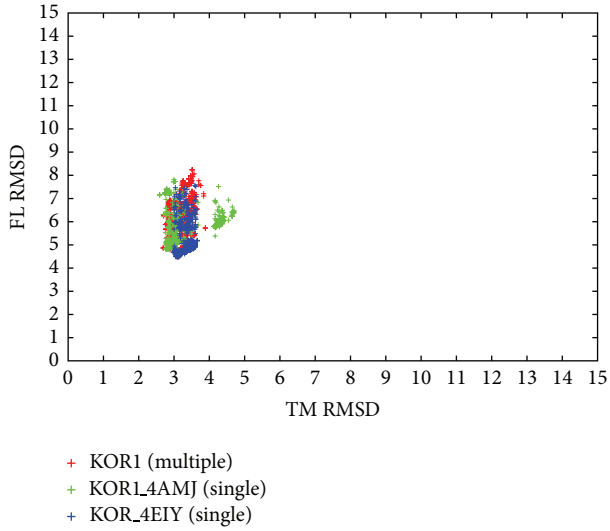
Multiple-template patGPCR yielded a higher number of low TM RMSD conformations than the single-template patGPCR in most cases (five out of eight targets), including CXCR4 (Figure 3(a)), D3 (Figure 3(d)), A2A (Figure 3(e)), KOR3 (Figure 3(g)), and Beta2 (Figure 3(h)). In other cases, the multitemplate approach showed similar performance with single-template patGPCR, including KOR1, HH1R, and S1P1 in Figures 3(b), 3(c), and 3(f). For KOR1, the narrow sampling space in the single template may have resulted from similar performance between multiple-template and single-template patGPCR. For HH1R and S1P1, both multiple-template and single-template patGPCR appeared to show bottlenecks. One reason for this may be that the transmembrane refining protocol employed by multiple-template and single-template approaches failed in some cases. Further analysis is described in Section 4.3.

3.3. Comparison of patGPCR to Homology Approach SWISS-MODEL. The SWISS-MODEL [16] is a widely used homology modeling approach which provides online prediction and manual template specification. We executed the SWISS-MODEL to predict eight targets using the same templates as used in patGPCR (Table 1). Three predictions that failed using the SWISS-MODEL are indicated as "×" in column 9 in Table 1.

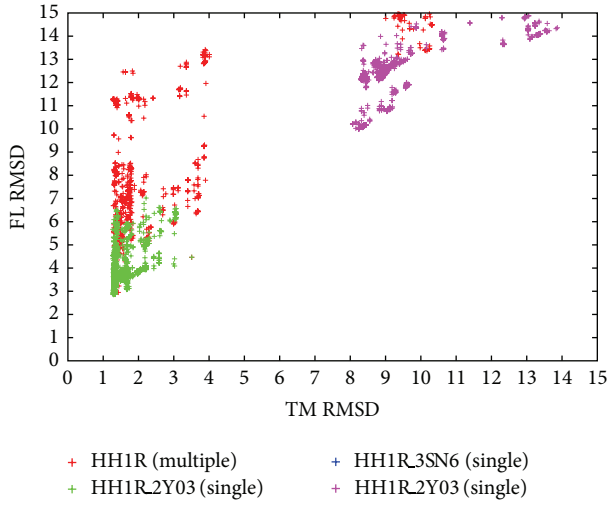
For each decoy (set of conformations of the predicted result) predicted by patGPCR, the top 400 conformations in terms of TM RMSD were reserved, while the others were eliminated from the decoy. Thus, we simplified selection of the nearest native prediction from the decoys. We depicted the comparison of decoys from the SWISS-MODEL results using a box-and-whisker plot (Figure 4). Prediction accuracy was expressed as the RMSD in angstroms, which was calculated after superimposing the corresponding coordinates of alpha-carbons determined by prediction and those of the native structure. To accurately determine transmembrane helices, the best model of five targets (CXCR4, KOR1, D3, A2A, and Beta2) yielded by patGPCR showed lower TM RMSD values than those determined using the SWISS-MODEL. patGPCR showed similar transmembrane accuracy on two targets (HH1R and S1P1); another target (KOR3), patGPCR, was inferior to the target identified using the SWISS-MODEL. In Figure 4, the accuracy of full-length chain and loop regions is shown. For four of the eight targets (KOR1, HH1R, KOR3, and Beta2), the best models predicted by patGPCR showed lower full-length RMSD values than those generated using the SWISS-MODEL. It is not surprising that only three targets (KOR1, KOR3, and Beta2) for patGPCR were more accurately predicted compared to the SWISS-MODEL regarding the accuracy of loop regions since patGPCR places emphasis on helix optimization, whereas relatively little emphasis is placed on loop optimization and



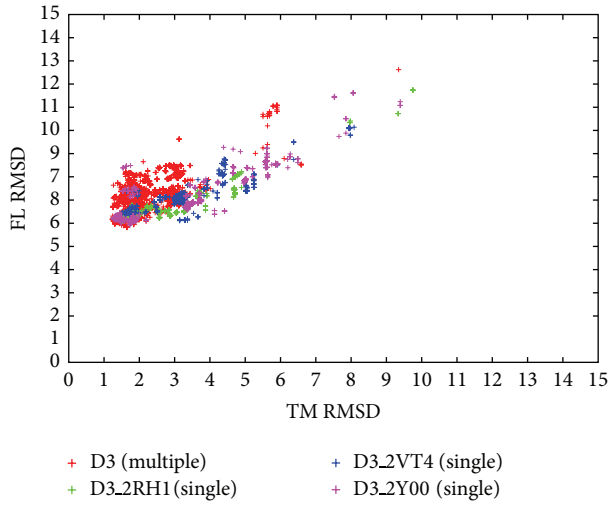
(a) Comparison between multiple-template approach and single-template approach for target CXCR4



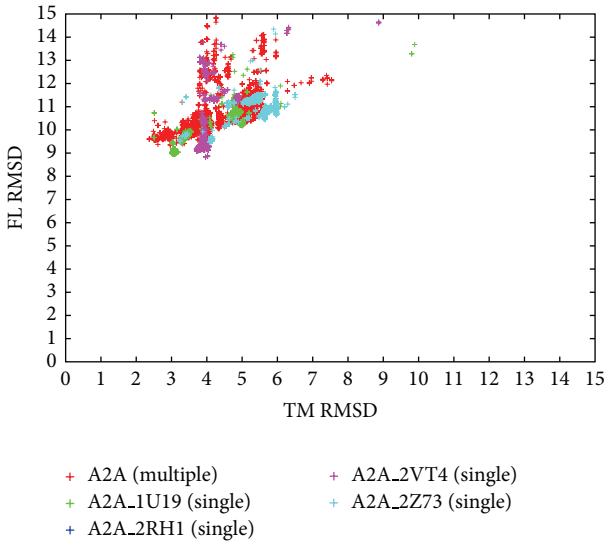
(b) Comparison between multiple-template approach and single-template approach for target KOR1



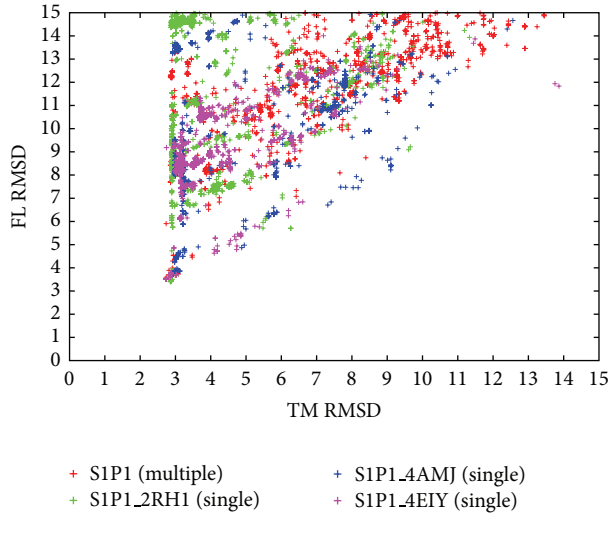
(c) Comparison between multiple-template and single-template approach for target HH1R



(d) Comparison between multiple-template and single-template approach for target D3

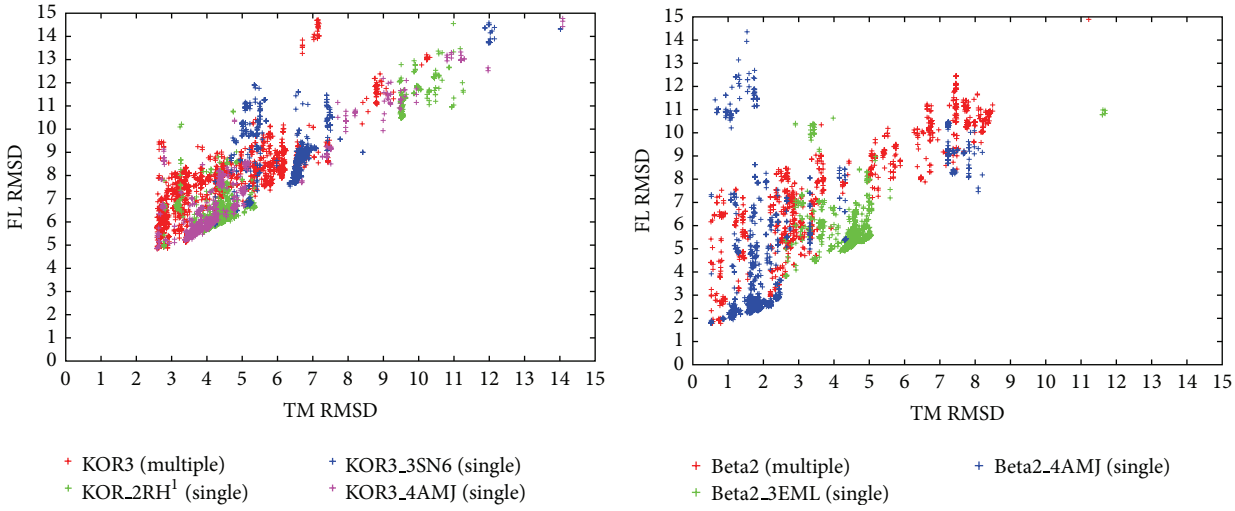


(e) Comparison between multiple-template and single-template approach for target A2A



(f) Comparison between multiple-template and single-template approach for target S1P1

Figure 3: Continued.



(g) Comparison between multiple-templates and single-template approach for target KOR3

(h) Comparison between multiple-template and single-template approach for target Beta2

Figure 3: Comparison of multiple-template and single-template patGPCR for eight GPCR targets. The x- and y-axes indicate TM RMSD and FL RMSD, respectively. Results generated using the multitemplate approach are shown in red and those generated using the single-template version of patGPCR are shown in green, blue, and purple. For five out of eight targets, CXCR4 (a), D3 (d), A2A (e), KOR3 (g), and Beta2 (h), the multitemplate approach yielded a higher number of conformations with lower TM RMSD values than the single-template approach. For KOR1 (b), the narrow sampling space in any single template may have resulted in similar performances for the multiple-template and single-template approaches.

the loop regions were excluded in the crossing step. Next, statistical properties of the representative solutions, the top 400 conformations regarding TM RMSD, for each decoy set were investigated. For each result decoy, Figure 5 shows the 1-percentile average (x-axis) and standard deviation (error bars of symbols) for every test instance, calculated from 50-fold bootstrap estimation using the BioShell package [38]. We choose the 1 percentile compared with the SWISS-MODEL results (y-axis) since only some of the decoys will be retained for further analysis, such as side chain refinement, ligand docking, and virtual screen. In Figure 5, patGPCR showed 14, 16, and 11 better results (below the line) for a total of 20 symbols on TM, FL, and loop RMSD values, respectively. In Figure 5, the lower deviations of the error bars indicate that patGPCR exhibits robust performance in most cases.

3.4. Comparison of patGPCR to Homology Approach MODELLER. MODELLER [14] is a well-known tool used for single-template and multiple-template homologies or comparative modeling of protein three-dimensional structures. We conducted single-template and multiple-template experiments using MODELLER on the same Linux system and hardware as was used for patGPCR. For a reasonable comparison, single-template MODELLER generated 1300 predicted conformations, which was similar to the average predicted number of 1358.75 conformations generated by patGPCR, for each target and each template. Multiple-template MODELLER was also used to generate 1300 prediction conformations for each target. MODELLER also employed the same templates as patGPCR listed in Table 1. MODELLER aligning commands, including align2d() and salign(), which reportedly take into account structural information from the template when constructing an alignment, were used to align templates to the targets.

MODELLER experimental results are tabulated in columns 10 and 11 in Table 1. The 10th column shows the average and standard deviation of TM RMSD of conformations generated using the single-template MODELLER, and the 11th column shows the average and standard deviation of TM RMSD of conformations generated by the multiple-template MODELLER. For six out of eight targets, patGPCR showed lower average TM RMSD than single-template MODELLER, and patGPCR showed only one higher average TM RMSD target compared with multiple-template MODELLER. Multiple-template patGPCR yielded a larger number of lower TM RMSD conformations than the single-template patGPCR for most cases (five out of eight targets), including CXCR4, D3, A2A, KOR3, and Beta2. However, no target was improved by using multiple-template MODELLER. The average TM RMSD values for conformations generated by using multiple-template MODELLER were nearly two times higher than those generated using single-template MODELLER. The lower standard deviation in column 11 indicates that a narrow structural sampling strategy was adopted by multiple-template MODELLER, which may have led to the multiple-template MODELLER being trapped by an unaligned region.

Finally, a direct visual comparison of conformations yielded by native, patGPCR, SWISS-MODEL, and MODELLER are depicted in Figure 6.

4. Discussions

4.1. Complexity Analysis of patGPCR. Since patGPCR does not synchronize the exchange of helices between pipelines,

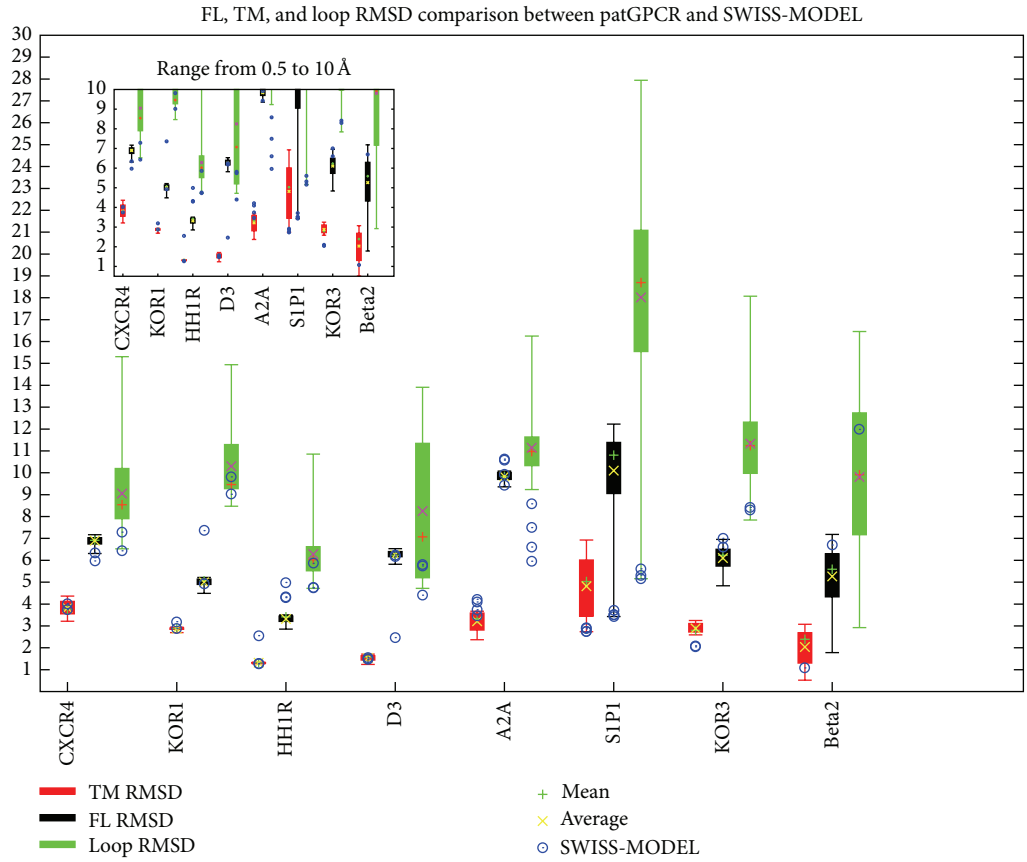


Figure 4: A box-and-whisker plot comparing FL, TM, and Loop RMSD between patGPCR and the SWISS-MODEL. The maximum, minimum, 1st quartile, 3rd quartile, mean (shown by +), and average (shown by ×) of RMSD (y-axis in angstrom) of FL, TM, and loop RMSD values are shown in red, black, and green for each target (x-axis). The results of the SWISS-MODEL are marked as blue circles in the corresponding box. The left-top panel is a detailed illustration in the range from 0.5 Å to 10 Å.

the complexity of patGPCR depends on that of a single pipeline repeating Algorithm 1 T times. For a GPCR target with n amino acid residues, the functions *transferHelix*(), *spinHelix*(), and *tiltGaussianHelix*() execute a movement of each residue in the helices, so these three functions have the same time complexity, $O(n)$. Therefore, the complexity of the TM refinement contained in the three stages from the fourth to the 15th line of Algorithm 1 is $3 \times O(n^2)$. The function *crossHelices*() exchanges the residues in each pair of helices and has complexity $O(n^2)$. The complexity of *LoopModelRandomly*() is determined by that of the Rosetta CCD and KIC refinement protocols, which are both composed of two-layer iterations repeating *outer_cycles* and *inner_cycles* (Rosetta parameters). Moreover, we generally set the parameters T, *STAGE2_CYCLES* (in the fourth line of Algorithm 1), *STAGE4_CYCLES* (in the 12th line of Algorithm 1), *outer_cycles*, and *inner_cycles* in a way that is linearly dependent on n. In summary, the time complexity of the parallelized patGPCR is eight times of that of a single pipeline, which is $n \times (3 \times O(n^2) + O(n^2) + O(n^2)) = O(n^3)$.

The space complexity of patGPCR is mainly determined by the elite conformation pool shown in Figure 1, the set of residue numbers for the transmembrane regions $TM[]$, and the set of residue numbers for the loop regions $Loop[]$. The pool is designed to store the coordinates of four backbone atoms of the elite conformations with the same size of template, so it requires $4 \times 3 \times n \times F$ space, where F is the number of available templates for the corresponding GPCR targets. $TM[]$ and $Loop[]$ obviously require $O(n)$ space. The space complexity of patGPCR is then $O(4 \times 3 \times n \times F) + 2 \times O(n) = O(n)$.

4.2. Why Does the patGPCR Approach Work in General? The following factors may contribute to the improvement in transmembrane helix structural prediction. First, a multi-template approach can effectively combine more reasonably aligned regions than a single-template approach. Single-template approaches use the top-ranked template and its alignment with the target protein to model the structure. These approaches cannot always achieve the best results, owing to difficulties in detecting the best template, particularly when remote homology is detected between targets and templates [39]. Because most of the sequence similarities among GPCRs are less than 50%, it is unreliable to attempt to

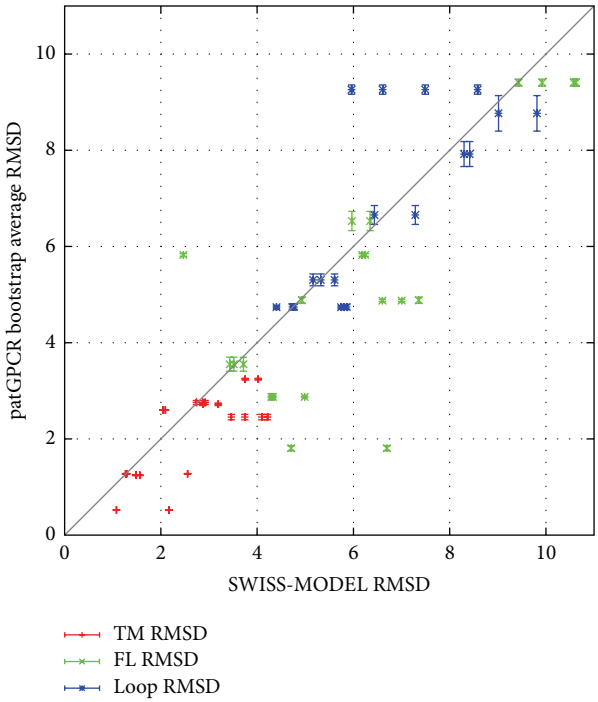


Figure 5: Statistical comparison between patGPCR after a 50-fold bootstrap estimation and the SWISS-MODEL on FL, TM, and Loop RMSD. x-axis indicates 1-percentile average RMSD of 50-fold bootstrap estimation of patGPCR prediction results and y-axis indicates the RMSD of the SWISS-MODEL. The error bars of the points indicate standard deviation 1 percentiles of patGPCR. Red, green, and blue show the comparison on TM, FL, and loop RMSD values, respectively. Points below the line represent targets where patGPCR yields higher accuracy, and above the line the SWISS-MODEL yields higher accuracy.

obtain high-resolution models using only a single template. We believe that each template, even if it has lower sequence similarity (<30%), contains a reasonable structure for the targets overall. Multiple-template approaches can not only increase the alignment coverage but also broaden the conformation search space by extracting aligned fragments from different templates.

Second, the parallelized approach enables an effective schema for exchanging reasonable regions among multiple templates. Previous multitemplate approaches [22, 40] have generally built a "super" template by methods that involve very little merging of aligned fragments from different templates, whereas the parallelized schema of patGPCR includes a soft resolution for introducing aligned regions and exchanges the reasonable helices between different pipelines at the end of an iteration. Therefore, the contribution of aligned regions from different templates will not only influence the beginning of the algorithm but also be preserved in subsequent iterations. Fortunately, because the helices that the GPCRs have in common are conserved, the transmembrane refinement protocol (see Section 2.3) will improve the accuracy of the aligned regions by moving the coordinates of the atoms toward lower free energy. This also helps the parallelized schema to broadcast the reasonable regions to the elite conformation pool.

Third, the additional energy item E related to the helical bundles (see Section 2.6) can reasonably guide the helices toward assembling into a tighter bundle, although not perfectly. The combination of the Rosetta membrane score with E based on the GPCR targets is effective in patGPCR, as shown in our experiments.

4.3. Limitations of patGPCR. On the other hand, patGPCR can also be improved in some respects. First, owing to the inadequately known structures of GPCRs, we could only validate the strengths and weaknesses of patGPCR on eight determined cases published on the GPCR Network website. These eight targets are the largest test data set for GPCRs at present. Therefore, the question of whether the current results of our experiments are occasional instances or have some universality cannot be answered immediately. Second, patGPCR seemingly encountered a bottleneck in some cases (in Figures 3(c) and 3(f)); it was difficult for patGPCR to obtain an accuracy results than single-template approach. The bottleneck may have two reasons: (a) the energy term in (1) improves the Rosetta energy function, but it is still not good enough to distinguish between the near-native conformations and (b) in patGPCR, the helix is treated as a rigid body, unlike real helices, which have disorder, irregular regions, and flexibility. The coordinates of the atoms in the interior of the helix are likely to cause a loss of precision. Third, although patGPCR generates a higher quality of decoys in the transmembrane regions, the difficulty of building accurate models of loop regions with high flexibility is a consequence of the difficult issues of the GPCR structure prediction problem.

4.4. Ability of New Energy Item. The ability of the new energy item E' was investigated for six GPCR targets, including CXCR4, KOR1, D3, A2A, KOR3, and Beta2, whose average TM RMSDs were lower than 5 Å. The relatively lower average TM RMSD values gained by using both patGPCR and MODELLER indicate that the accuracy of these targets depend on the energy functions. In Table 3, the top quarter of conformations in terms of TM RMSD values from the decoy of multiple-template patGPCR were chosen to compare how many conformations could be predicted using MEM (Rosetta membrane score) and E' (new energy item). For four targets (KOR1, D3, A2A, and KOR3), E' identified more nearly native conformations than MEM. Only one target (CXCR4) was identified to have fewer conformations than MEM.

5. Conclusion

In this paper, we have presented a parallelized multitemplate approach, patGPCR, to predicting the 3D structure of the transmembrane helices of G-protein-coupled receptors. patGPCR, which employs a bundle-packing-related energy function that improves on the RosettaMem energy, parallelizes eight pipelines for the refinement of transmembrane helix structures and exchanges the optimized helix structures

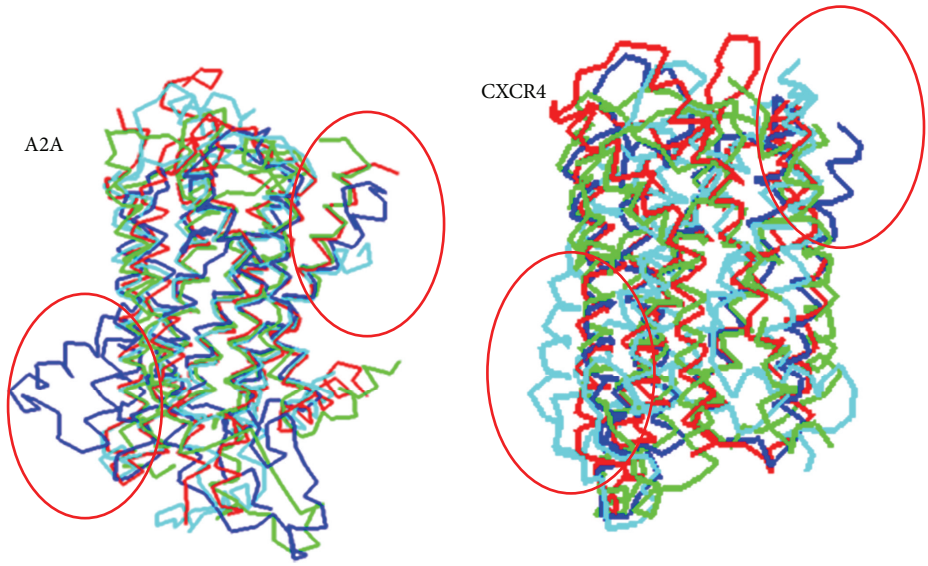


Figure 6: Superimposition of conformations yielded by native (red), patGPCR (green), SWISS-MODEL (blue), and MODELLER (cyan). The four red cycles indicate that patGPCR (green) backbones in the TM region are closer to native (red) than those predicted using the SWISS-MODEL (blue) and MODELLER (cyan) in targets A2A (a) and CXCR4 (b).

Table 3: Comparison of MEM and E'.

Target names	Number[a]	Neither[b]	Only E'[c]	Only MEM[d]	Both[e]
CXCR4	232	137	8	87	0
KOR1	304	107	86	52	59
D3	444	249	82	75	38
A2A	520	72	192	133	123
KOR3	352	172	84	74	22
Beta2	224	77	42	42	63

[a]The number of top quarter conformations in terms of TM RMSD from the decoy.
[b]The number of conformations that was neither in top quarter conformations in terms of MEM nor in top quarter conformations in terms of E'.
[c]The number of conformations that was only in top quarter conformations in terms of E'.
[d]The number of conformations that was only in top quarter conformations in terms of MEM.
[e]The number of conformations that was both in top quarter conformations in terms of MEM and in top quarter conformations in terms of E'.

among multiple templates. We have investigated the performance of patGPCR on a test set containing eight determined GPCRs published on the GPCR Network website. The results indicate that our multitemplate algorithm improves the TM RMSD values of the predicted models by 33.64% on average against a single-template method. Compared with other ab initio and homology approaches, patGPCR yielded more predicted conformations with high-resolution structures of the transmembrane helices. The best models for five of the eight targets built by patGPCR had a lower TM RMSD than had the models obtained from SWISS-MODEL, and half of them had a lower full-length RMSD. For six out of eight targets, patGPCR showed lower average TM RMSD than single-template MODELLER, and patGPCR showed only one higher average TM RMSD target compared with multiple-template MODELLER.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (http://www.nsfc.gov.cn) under Grants no. 60970055 and 61170125. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper. The authors thank Haiou Li and Rong Chen for helping with the analysis of the experiments and Lu Sun for helping with the preparation of the paper.

References

[1] D. Fillmore, "It's a GPCR world," *Modern Drug Discovery*, vol. 11, pp. 24–28, 2004.

[2] V. Cherezov, D. M. Rosenbaum, M. A. Hanson et al., "High-resolution crystal structure of an engineered human β2-adrenergic G protein-coupled receptor," *Science*, vol. 318, no. 5854, pp. 1258–1265, 2007.

[3] I. Kufareva, M. Rueda, V. Katritch et al., "Status of GPCR modeling and docking as reflected by community-wide GPCR Dock 2010 assessment," *Structure*, vol. 19, pp. 1108–1126, 2011.

[4] M. Michino, J. Chen, R. C. Stevens, and C. L. Brooks, "FoldGPCR: structure prediction protocol for the transmembrane domain of G proteincoupled receptors from class A," *Proteins*, vol. 78, no. 10, pp. 2189–2201, 2010.

[5] V. Katritch, M. Rueda, P. C. H. Lam, M. Yeager, and R. Abagyan, "GPCR 3D homology models for ligand screening: lessons learned from blind predictions of adenosine A2a receptor complex," *Proteins*, vol. 78, no. 1, pp. 197–211, 2010.

[6] Y. Zhang, M. E. Devries, and J. Skolnick, "Structure modeling of all identified G protein-coupled receptors in the human genome," *PLoS Computational Biology*, vol. 2, no. 2, pp. 88–99, 2006.

[7] C. de Graaf, N. Foata, O. Engkvist, and D. Rognan, "Molecular modeling of the second extracellular loop of G-protein coupled receptors and its implication on structure-based virtual screening," *Proteins*, vol. 71, no. 2, pp. 599–620, 2008.

[8] P. W. Hildebrand, A. Goede, R. A. Bauer et al., "SuperLooper—a prediction server for the modeling of loops in globular and membrane proteins," *Nucleic Acids Research*, vol. 37, no. 2, pp. W571–W574, 2009.

[9] L. Sun, X. Zeng, C. Yan et al., "Crystal structure of a bacterial homologue of glucose transporters GLUT1-4," *Nature*, vol. 490, pp. 361–366, 2012.

[10] T. A. Hopf, L. J. Colwell, R. Sheridan et al., "Three-dimensional structures of membrane proteins from genomic sequencing," *Cell*, vol. 149, pp. 1607–1621, 2012.

[11] L. Bordoli, F. Kiefer, K. Arnold, P. Benkert, J. Battey, and T. Schwede, "Protein structure homology modeling using SWISS-MODEL workspace," *Nature Protocols*, vol. 4, no. 1, pp. 1–13, 2009.

[12] *Sybyl-X, Version 1. 0*, Tripos, St. Louis, Mo, USA, 2009.

[13] *Prime, Version 1.6*, Schrodinger, LLC, New York, NY, USA, 2007.

[14] A. Sali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," *Journal of Molecular Biology*, vol. 234, no. 3, pp. 779–815, 1993.

[15] D. Petrey, Z. Xiang, C. L. Tang et al., "Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling," *Proteins*, vol. 53, supplement 6, pp. 430–435, 2003.

[16] M. Levitt, "Accurate modeling of protein conformation by automatic segment matching," *Journal of Molecular Biology*, vol. 226, no. 2, pp. 507–533, 1992.

[17] P. A. Bates, L. A. Kelley, R. M. MacCallum, and M. J. E. Sternberg, "Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM," *Proteins*, vol. 45, supplement 5, pp. 39–46, 2001.

[18] P. Koehl and M. Delarue, "Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy," *Journal of Molecular Biology*, vol. 239, no. 2, pp. 249–275, 1994.

[19] V. Yarov-Yarovoy, J. Schonbrun, and D. Baker, "Multipass membrane protein structure prediction using Rosetta," *Proteins*, vol. 62, no. 4, pp. 1010–1025, 2006.

[20] P. Barth, J. Schonbrun, and D. Baker, "Toward high-resolution prediction and design of transmembrane helical protein structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 40, pp. 15682–15687, 2007.

[21] C. N. Cavasotto and S. S. Phatak, "Homology modeling in drug discovery: current trends and applications," *Drug Discovery Today*, vol. 14, no. 13-14, pp. 676–683, 2009.

[22] J. Cheng, "A multi-template combination algorithm for protein comparative modeling," *BMC Structural Biology*, vol. 8, article 18, 2008.

[23] T. Liu, M. Guerquin, and R. Samudrala, "Improving the accuracy of template-based predictions by mixing and matching between initial models," *BMC Structural Biology*, vol. 8, article 24, 2008.

[24] Q. Lü, X. Xia, R. Chen et al., "When the lowest energy does not induce native structures: parallel minimization of multi-energy values by hybridizing searching intelligences," *PLoS ONE*, vol. 7, no. 9, Article ID e44967, 2012.

[25] Q. Lü, H. Wu, J. Wu, X. Huang, X. Luo, and P. Qian, "A parallel ant colonies approach to de novo prediction of protein backbone in CASP8/9," *Science China Information Sciences*, in press.

[26] R. J. Lefkowitz, "The superfamily of heptahelical receptors," *Nature Cell Biology*, vol. 2, no. 7, pp. E133–E136, 2000.

[27] M. G. Claros and G. von Heijne, "TopPred II: an improved software for membrane protein structure predictions," *Computer Applications in the Biosciences*, vol. 10, no. 6, pp. 685–686, 1994.

[28] uniprot, 2012, http://www.uniprot.org/.

[29] K. Hofmann and W. Stoffel, "Tmbase—a database of membrane spanning proteins segments," *Biological Chemistry Hoppe-Seyler*, vol. 374, no. 166, 1993.

[30] G. E. Tusnády and I. Simon, "The HMMTOP transmembrane topology prediction server," *Bioinformatics*, vol. 17, no. 9, pp. 849–850, 2001.

[31] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *Journal of Molecular Biology*, vol. 305, no. 3, pp. 567–580, 2001.

[32] H. Viklund and A. Elofsson, "OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar," *Bioinformatics*, vol. 24, no. 15, pp. 1662–1668, 2008.

[33] C. S. Soto, M. Fasnacht, J. Zhu, L. Forrest, and B. Honig, "Loop modeling: sampling, filtering, and scoring," *Proteins*, vol. 70, no. 3, pp. 834–843, 2008.

[34] C. S. Rapp, T. Strauss, A. Nederveen, and G. Fuentes, "Prediction of protein loop geometries in solution," *Proteins*, vol. 69, no. 1, pp. 69–74, 2007.

[35] K. Zhu, D. L. Pincus, S. Zhao, and R. A. Friesner, "Long loop prediction using the protein local optimization program," *Proteins*, vol. 65, no. 2, pp. 438–452, 2006.

[36] D. J. Mandell, E. A. Coutsias, and T. Kortemme, "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling," *Nature Methods*, vol. 6, no. 8, pp. 551–552, 2009.

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

[38] D. Gront and A. Kolinski, "BioShell—a package of tools for structural biology computations," *Bioinformatics*, vol. 22, no. 5, pp. 621–622, 2006.

[39] C. Venclovas and M. Margelevicius, "Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment," *Proteins*, vol. 61, supplement 7, pp. 99–105, 2005.

[40] B. Al-Lazikani, F. B. Sheinerman, and B. Honig, "Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 14796–14801, 2001.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 989307, 17 pages
http://dx.doi.org/10.1155/2013/989307

Research Article

Understanding the Pathogenesis of Kawasaki Disease by Network and Pathway Analysis

Yu-wen Lv,[1] Jing Wang,[2] Ling Sun,[1] Jian-min Zhang,[1] Lei Cao,[1] Yue-yue Ding,[1] Ye Chen,[1] Ji-juan Dou,[1] Jie Huang,[1] Yi-fei Tang,[2] Wen-tao Wu,[2] Wei-rong Cui,[2] and Hai-tao Lv[1]

[1] *Department of Pediatric Cardiology, Children's Hospital of Soochow University, Suzhou 215003, China*
[2] *Center for Systems Biology, Soochow University, Suzhou 215006, China*

Correspondence should be addressed to Hai-tao Lv; haitaosz@163.com

Received 22 November 2012; Accepted 4 January 2013

Academic Editor: Guang Hu

Kawasaki disease (KD) is a complex disease, leading to the damage of multisystems. The pathogen that triggers this sophisticated disease is still unknown since it was first reported in 1967. To increase our knowledge on the effects of genes in KD, we extracted statistically significant genes so far associated with this mysterious illness from candidate gene studies and genome-wide association studies. These genes contributed to susceptibility to KD, coronary artery lesions, resistance to initial IVIG treatment, incomplete KD, and so on. Gene ontology category and pathways were analyzed for relationships among these statistically significant genes. These genes were represented in a variety of functional categories, including immune response, inflammatory response, and cellular calcium ion homeostasis. They were mainly enriched in the pathway of immune response. We further highlighted the compelling immune pathway of NF-AT signal and leukocyte interactions combined with another transcription factor NF-κB in the pathogenesis of KD. STRING analysis, a network analysis focusing on protein interactions, validated close contact between these genes and implied the importance of this pathway. This data will contribute to understanding pathogenesis of KD.

1. Introduction

KD is a systemic vascular disease preferentially occurring in infants and children [1, 2]. It is characterized by the development of coronary artery aneurysms (CAA) which may result in fatal thrombosis and sudden cardiac failure. Clinical manifestations of KD include prolonged fever (1-2 weeks, mean 10-11 days), conjunctival infection, oral lesions, polymorphous skin rashes, extremity changes, and cervical lymphadenopathy, all of which comprise diagnostic criteria [3]. However, great majority of children failed to manifest typical characteristics. In addition to the diagnostic criteria, there are a broad range of nonspecific clinical features, including irritability, uveitis, aseptic meningitis, cough, vomiting, diarrhea, abdominal pain, gallbladder hydrops, urethritis, arthralgia, arthritis, hypoalbuminemia [4], liver function impairment, and heart failure [5, 6]. The peaked incidence at 9 to 11 months of age coincides with fading of maternal immunity, and symptoms partly similar to other infectious disorders suggest that some microorganisms may trigger this disease. Despite great efforts to identify the cause for nearly a half a century, the etiology of KD still remains unknown [7]. However, the role of genetic susceptibility to KD has long been evident through its striking predilection for children of Japanese ethnicity regardless of their country of residence; compared with Caucasian children, Japanese children have a relative risk of KD that is 10 to 15 times higher [8–10]. Siblings of KD children have a relative risk that is 6 to 10 times greater than that of children without a family history, and the parents of Japanese children with KD are twice as likely to have had KD themselves as children than other adults in the general Japanese population [11–14].

Candidate gene studies and genome-wide studies have been successively applied to explore the association between genetic effect and this mysterious disease [15, 16]. Many suspicious genes related to innate and acquired immune functions or to vascular remodeling have been studied [15, 17–19].

Genetic studies of KD were conducted not only to clarify the genetic background but also in the hope of providing clues about its etiology and pathogenesis. However, none of these studies have analyzed the internal association between these significant association genes and explored the possible pathogenic process in KD from overall level.

In this paper, we aim to extract statistically significant genes associated with KD (Up to September 2012, from all English databases) to explore their association and analyze their function in the pathogenesis of KD. This study is a systematic summary of previous research. Further studies on clinical validation will be summarized in our next study.

2. Methods and Materials

2.1. Extracting Genes with Statistical Significance. We performed a computerized search of Ovid, Google Scholar, and PubMed databases up to September 2012 and reviewed cited references to identify the relevant studies. Citations were screened at the title/abstract level and retrieved as full reports. Search keywords were "Kawasaki disease," "Kawasaki syndrome," "lymph node syndrome," "mucocutaneous lymph node syndrome" combined with "polymorphism," "gene," "genetic," "allele," and "genotype". The inclusion criteria of genes were those who have significant association with KD contributed to susceptibility, vascular lesions, resistance to initial IVIG treatment, late diagnosis of KD, and incomplete KD.

2.2. Data Analysis. DAVID (http://david.abcc.ncifcrf.gov/, version: 6.7) was used to process the bioinformatics analysis of these candidate gene markers, including gene classification (based on Biological Process Ontology and Molecular Function Ontology, resp.), enrichment analysis for significant gene ontology categories, KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway mapping, and significant pathway computing. GeneGo MetaCore (http://www.genego.com/, version: 6.5) was used to analyze the pathways of these significant genes. The association between these statistically significant genes were analyzed using STRING (http://string-db.org/), a database of known and predicted protein interactions.

3. Results

3.1. Extracting Genes with Statistical Significance. The characteristics of the genes are presented in Table 1 (candidate genetic studies) and Table 2 (genome-wide studies).

3.2. Gene Ontology Analysis. Genes with statistical significance were submitted to functional analysis using DAVID software. Defense response, response to wounding, and inflammatory response were identified as significantly enriched (Enrichment Score = 15.91). Furthermore, DAVID analysis identified clusters of genes with annotations related to cellular calcium ion homeostasis, cell chemotaxis (enrichment Score: 3.75), and positive regulation of immune system process (Enrichment Score: 3.58) which is involved in autoimmune thyroid disease (hsa05320), asthma (hsa05310), type I diabetes mellitus (hsa04940), and allograft rejection (hsa04672). The functional annotation table can be available in supplementary material available online at http://dx.doi.org/10.1155/2013/989307.

3.3. Enrichment Analysis. Enrichment analysis consists of matching genes in functional ontologies by GeneGo MetaCore (Figure 1). The probability of a random intersection between a set of gene list with ontology entities was estimated with the "*P*" value of the hypergeometric intersection. A lower "*P*" value means higher relevance of the entity to the dataset, which appears in higher rating for the entity. All maps were drawn by GeneGo. The height of the histogram corresponded to the relative expression value for a particular gene.

The most significant GeneGo Pathway Maps were (1) immune response: HSP60 and HSP70/TLR signaling pathway; (2) immune response: Inflammasome in inflammatory response; (3) cell adhesion: plasmin signaling; (4) immune response: NF-AT signaling and leukocyte interactions. In addition, there are other pathways including Role of HMGB1 in dendritic cell maturation and migration; histamine signaling in dendritic cells: plasmin signaling in cell adhesion; cross-talk between VEGF and angiopoietin 1 signaling pathways; regulation of epithelial-to-mesenchymal transition (EMT); TGF-beta-dependent induction of EMT via SMADs in Development; role of IAP-proteins in apoptosis pathway in apoptosis and survival, and so forth. Meanwhile, immune system process, defense response, and response to stress were the most significantly enriched GO processes of these genes. With the disease folders, representing over 112 human diseases annotated by GeneGo, these 76 genes were mainly related to autoimmune diseases and some kinds of vascular inflammatory diseases.

The abstracted genes involved in significant pathways are summarized in Table 3.

3.4. STRING Analysis. Now specifically, we are interested in finding functional associations among these genes. We broadcast our data to STRING (a database of known and predicted protein interactions), which responds by displaying a network of nodes (proteins) connected by colored edges representing functional relationships.

Figure 2 summarizes the network of predicted associations between proteins encoded by these genes. The results indicate that CASP3, IL18, BLK, FCGR2B, FCGR2A, CRP, CCR5, CCL5, CCR3, CCL3L1, TNFRSF1A, TNF, IL4, ERAP1, LTA, CD40, NOD1, CTLA4, NLRP1, TGFBR2, SMAD3, TGFB2, VEGFA, KDR, and CCR2 are associated according to experimental evidence, with involvement in many signaling pathways; TNF was the key of nodes, linking to CRP, IL-4, CD40, CD40LG, IL-18, IL-10, and so on. They linked to many immune and inflammatory responses. All of these proteins (encoded by genes) are interrelated, forming a large network. However, many proteins are not linked to others, indicating that their functions are unrelated or unknown.

Table 1: Candidate gene studies identified genes associated with KD.

Symbol	Region	Phenotype	Country	Reference
CD40	20q12-q13.2	KD	Taiwan	[20]
		CAL	Taiwan	[20]
CD209	19p13	KD	Taiwan	[21]
RETN	19p13.2	Incomplete KD	China	[22]
		KD	United States	[23]
		CAL	Japan	[24]
FCGR3B	1q23	IVIG nonresponse	United States	[23]
NOD1	7p15-p14	KD	Japan	[25]
NLRP1	17p13.2	KD	Japan	[25]
ITPKC	19q13.1	KD	Taiwan; Japan; China	[26–28]
		CAL	Japan	[29]
		IVIG nonresponse	Japan	[29]
TGFBR2	3p22	KD	European descent; Korea	[30, 31]
		CAL	European descent; Korea	[30, 31]
		IVIG nonresponse	European descent	[30]
		aortic root dilatation,	European descent	[30]
ABO	9q34.2	CAL	Japan	[32]
PELI1	2p13.3	CAL	Korea	[33]
SMAD3	15q22.33	KD	European descent; Taiwan	[30, 34]
		CAL	European descent	[30]
		IVIG nonresponse	European descent	[30]
		aortic root dilatation	European descent	[30]
TGFB2	1q41	KD	European descent; Taiwan	[30, 34]
		CAL	European descent	[30]
		IVIG nonresponse	European descent	[30]
		aortic root dilatation,	European descent	[30]
CASP3	4q34	CAL	Taiwan; Japan	[29, 35]
		IVIG nonresponse	Japan	[29]
ANGPT1	8q23.1	KD	Netherlands	[36]
VEGFA	6p12	KD	Netherlands; Taiwan; The Netherlands.	[36–38]
		CAL	Japan	[39]
MICB	6p21.3	KD	Taiwan	[40]
		CAL	Taiwan	[40]
MICA	6p21.33	KD	Taiwan	[40]
		CAL	Taiwan	[41]
BAG6	6p21.3	KD	Taiwan	[40, 42]
		CAA	Taiwan	[42]
MSH5	6p21.3	KD	Taiwan	[40]
VWA7	6p21.33	KD	Taiwan	[40]
FCGR2B	1q23	IVIG nonresponse	Pacific Northwest	[43]
IL10	1q31-q32	KD	Taiwan	[44, 45]
		CAL	China; Korea; Taiwan	[18, 46, 47]
CCL5	17q11.2-q12	CAL	India	[48]

Table 1: Continued.

Symbol	Region	Phenotype	Country	Reference
TNFRSF1A	12p13.2	KD	China	[49]
CTLA4	2q33	CAL (particularly in female patients)	Taiwan	[50]
MMP3	11q22.3	CAL	Korea; US-UK, tested in Japan	[51, 52]
MMP12	11q22.3	CAL	US-UK, tested in Japan	[52]
FGB	4q28	CAL	China	[53]
CCL3L1	17q21.1	KD	USA; Japan	[54, 55]
		IVIG nonresponse	Japan	[55]
CCR5	3p21.31	KD	USA; The Netherlands (Dutch Caucasian); Korea	[54, 56, 57]
		CAL	Japan	[55]
		IVIG nonresponse	Japan	[55]
PRRC2A	6p21.3	KD	Taiwan	[42]
		CAL	Taiwan	[42]
ABHD16A	6p21.3	KD	Taiwan	[42]
		CAL	Taiwan	[42]
ITPR3	6p21	CAL	Taiwan	[58]
COL11A2	6p21.3	KD	Taiwan	[59]
		CAL	Taiwan	[59]
MBL2	10q11.2	KD	China; Japan	[60, 61]
		CAL	The Netherlands; The Netherlands	[62, 63]
		Arterial stiffness	China	[64]
MMP11	22q11.23	KD	Korea	[65]
MIF	22q11.23	CAL	Italy	[66]
IL1B	2q14	IVIG nonresponse	Taiwan	[17]
BTNL2	6p21.3	KD	Taiwan	[67]
		CAL	Taiwan	[67]
TPH2	12q21.1	CAL	Korea	[68]
PDCD1	2q37.3	KD	Korean	[69]
IL18	11q22.2-q22.3	KD	Taiwan	[70, 71]
HLA-E	6p21.3	KD	Taiwan	[72]
		CAL	Taiwan	[72]
TIMP4	3p25	CAL	Korea	[73]
HLA-G	6p21.3	KD	Korea	[74]
CRP	1q21-q23	KD	China	[75]
		Carotid stiffness and carotid intima-media thickness	China	[75]
TNF	6p21.3	KD	China	[75]
		CAL	white	[76]
		Intima-media thickness	China	[75]
		IVIG nonresponse	China	[46]
MMP13	11q22.3	CAL	Japan	[77]
HLA-B	6p21.3	KD	Korea	[78]
HLA-C	6p21.3	KD	Korea	[78]
CCR3	3p21.3	KD	Netherlands (Dutch Caucasian)	[56]
CCR2	3p21.31	KD	Netherlands (Dutch Caucasian)	[56]
TIMP2	17q25	CAL	Japan	[79]

Table 1: Continued.

Symbol	Region	Phenotype	Country	Reference
ACE	17q23.3	KD	Taiwan; Korea	[80, 81]
		Coronary artery stenosis	Japan	[82, 83]
		Myocardial ischemia	Japan	[82]
PLA2G7	6p21.2-p12	IVIG nonresponse	Japan	[84]
IL1RN	2q14.2	KD	Taiwan	[85]
IL4	5q31.1	KD	USA	[86]
KDR	4q11-q12	CAL	Japan	[39]
CD40LG	Xq26	CAL: males affected	Japan	[87]
AGTR1	3q24	Coronary artery stenosis and myocardial ischemia	Japan	[82]
CD14	5q31.1	CAL	Japan	[88]
SLC11A1	2q35	KD	Japan	[89]
LTA	6p21.3	KD	white	[76]
MTHFR	1p36.3	CAL	Japan	[90]
HP	16q22.2	late diagnosis of KD	Taiwan	[90]

KD: kawasakidisease;CAL: coronary artery lesions; CAA: coronary artery aneurysms.

Table 2: Susceptibility genes for KD identified with association at genome-wide significance.

Gene	Locus	Methods	Reference
FCGR2A	1q23	GWAS	[91]
BLK	8p23-p22	GWAS	[92, 93]
CASP3	4q34	Genome wide Linkage analysis	[94]
ITPKC	19q13.1	Genome wide Linkage analysis; linkage disequilibrium mapping	[94, 95]
CD40L	Xq26	Genome wide Linkage analysis	[94]
CD40	20q12-q13.2	GWAS	[92, 93]
HLA-DQB2	6p21	GWAS	[92]
HLA-DOB	6p21.3	GWAS	[92]
NFKBIL1	6p21.3	GWAS	[92]
LTA	6p21.3	GWAS	[92]
NAALADL2	3q26.31	GWAS	[96]
ZFHX3	16q22.3	GWAS	[96]
DAB1	1p32-p31	GWAS	[97]
PELI1	2p13.3	GWAS	[97]
COPB2	3q23	GWAS	[98]
ERAP1	5q15	GWAS	[98]
IGHV	14q32.33	GWAS	[98]
ABCC4	13q32	Genome-wide linkage and association mapping	[99]

GWAS: genome-wide association study.

4. Discussion

4.1. Immune Response in the Pathogenesis of KD. KD has long been considered as an abnormal immune disease. The activation of immune system and the cascade release of inflammatory factors are the important features in KD. A large number of T cells (increased activated CD4 T cells, depressed CD8 T cells and CD4+ CD25+ regulatory T cells), large mononuclear cells, macrophages and plasma cells, with a smaller number of neutrophils, are observed in various organ tissues of fatal cases of acute KD [102–106]. Additionally, various inflammatory cytokines and chemokines [107, 108], matrix metalloproteinases, nitric oxide production [109], autoantibody production [110, 111], and adhesive molecule expression [112, 113] are also overactivated in the acute stage of KD which are considered to facilitate vascular endothelial inflammation and then participate in the pathogenesis of KD and CAL formation. Go processes and DAVID analysis revealed that these genes are significantly enriched in immune responses which have the parallel results with clinical and laboratory findings. In addition, these genes are widely involved in other immune systemic and inflammatory diseases, for example, autoimmune thyroid disease, asthma, type I diabetes mellitus, allograft rejection, inflammatory bowel

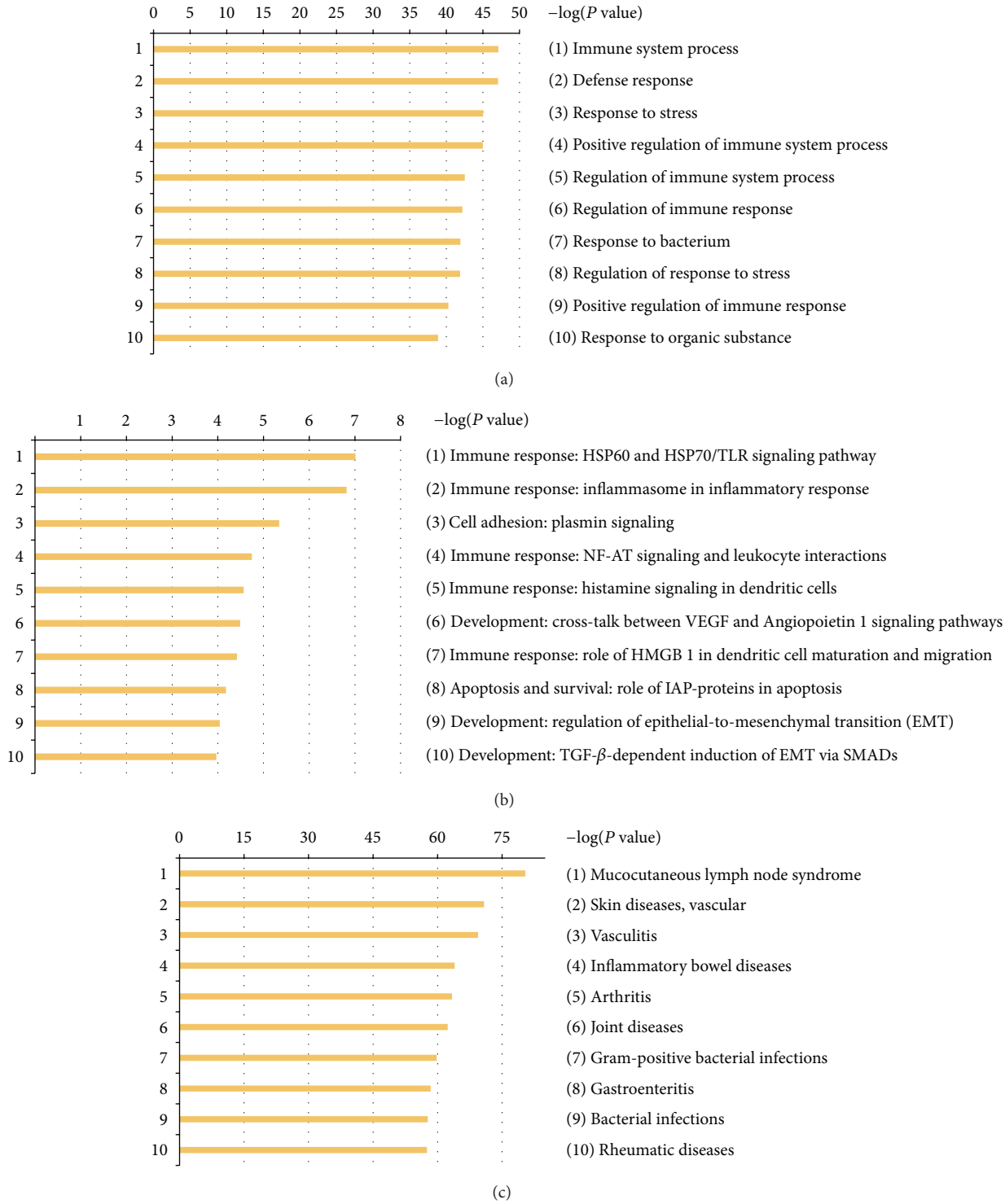


Figure 1: Enrichment analysis of the genes by GeneGo MetaCore: (a) GO Processes, (b) Go Pathway Maps, (c) Go Diseases (by Biomarkers). MetaCore version 6.11 build 41105.

disease, vasculitis, arthritis, and rheumatic disease. Furthermore, the signal pathway produced in GeneGo contains many immune response pathways that participate in inflammation, apoptosis, injury, and remodeling process, which have been listed in Table 3.

4.2. ECM-Remodeling and Plasmin Signaling Pathway in the Pathogenesis of KD. In addition to the signal pathway of the immune response, ECM-remodeling and plasmin signaling pathway associated with cell adhesion were enriched in GeneGo MetaCore software (FDR < 0.01, $P < 0.005$).

Table 3: Pathways analyzed by GeneGo Meta core.

Pathway categories	Pathways	Functions	Enrichment genes
Immune response	(1) HSP60 and HSP70/TLR signaling pathway (2) Inflammasome in inflammatory response (3) NF-AT signaling and leukocyte interactions (4) Role of HMGB1 in dendritic cell maturation and migration (5) Histamine signaling in dendritic cells (6) CD16 signaling in NK cells (7) MIF in innate immunity response (8) Th1 and Th2 cell differentiation (9) HMGB1 release from the cell (10) PGE2 signaling in immune response (11) Histamine H1 receptor signaling in immune response (12) Role of DAP12 receptors in NK cells	Pro-inflammatory response and anti-inflammatory response; cellular and humoral immune response; NO production; apoptosis and antiapoptosis; secretion of leukotrienes and prostaglandins; proliferation and differentiation of eosinophils; chemotaxis; proliferation, differentiation, activation of T cell; cell necrosis; smooth muscle construction; vascular permeability; blood coagulation; cytoskeleton remodeling	CD14, HSP70, IL-10, TNF-α, IL-1β, CD40, MHC class I, IL-4, NOD1, CARD7, IL-18, TNF-R1, CD40L, IP3receptor, CCL5, HLA-E, PLA2, MIF, CCR5, MMP13, HLA-C, HLA-B, HLA-G, HLA-E, Stromelysin-1
Cell adhesion	Plasmin signaling ECM remodeling	Fibrinolysis; cell viability	TGF-β 2,VEGF-A, TGF-β receptor type 2, VEGFR-2, Fibrinogen, MMP-13, TIMP2, Stromelysin-1, MMP-13, MMP-12
Development	(1) Cross-talk between VEGF and Angiopoietin 1 signaling pathways (2) Regulation of epithelial-to-mesenchymal transition (EMT) (3) TGF-β-dependent induction of EMT via SMADs (4) PEDF signaling (5) Glucocorticoid receptor signaling	Leukocyte-endothelial adhesion; epithelial-to-mesenchymal transition; proteasomal degradation; inhibition of angiogenesis; immune response	VEGF-A, VEGFR-2, Angiopoietin 1, IP3 receptor, TGF-β 2, IL-1β, TNF-α, TNF-R1, TGF-β receptor type 2, SMAD3, TGF-β, HSP70, MMP13
Apoptosis and survival	(1) Role of IAP-proteins in apoptosis (2) Anti-apoptotic TNFs/NF-kB/Bcl-2 pathway	Caspase dependent and independent apoptosis; apoptosis and antiapoptosis	TNF-α, TNF-R1, HSP70, caspase3, CD40L, CD40
Transcription	NF-kB signaling pathway	Activate the transcription of target genes	TNF-α, TGF-β, TNF-R1, CD14

FDR = 0.01.

Numerous studies suggest that they participated in the pathophysiological process of KD. Activation of the fibrinolytic system, vascular injury, and remodeling were the prominent outcome in these pathways. Activated plasmin in the plasmin signaling pathway which is a major fibrinolytic protease can directly degrade fibrinogen, laminin, and fibronectin [114]. On the cell surface, plasmin can activate a number of matrix metalloproteinases (MMPs) MMP1, MMP13 [115]. Other MMPs (MMP-9 and so on) were subsequently activated. Moreover, IL-1 β, IL-6, TNF-α, and IFN-γ can stimulate the endothelial cells to produce more MMP-9. These MMPs degrade extracellular matrix proteins and components of basal membranes leading to the disruption of the internal elastic lamina and the trilaminar structure of the vascular wall [116–118]. Many examinations have showed that many MMPs were highly expressed in the acute stage of KD. MMPs are prominent during the remodeling process, contributing to the formation of coronary artery lesions [119], and consequently the intima proliferates and thickens, while in rare cases the vessel wall becomes stenotic or occluded by either stenosis or thrombosis. Endogenous tissue inhibitors of metalloproteinases (TIMPs) such as TIMP1, TIMP2, and TIMP3 can reduce excessive proteolytic ECM degradation by MMPs. The balance between MMPs and TIMPs controls the extent of ECM remodeling [120, 121]. One study indicated that MMPs and TIMPs were in a state of imbalance in KD patients [122]. Therefore, ECM-remodeling and plasmin signaling pathway may have played a certain role in the vascular damage in KD.

4.3. NF-AT Signaling and Leukocyte Interactions. NF-AT signaling and leukocyte interactions (P value = 2.28×10^{-5})

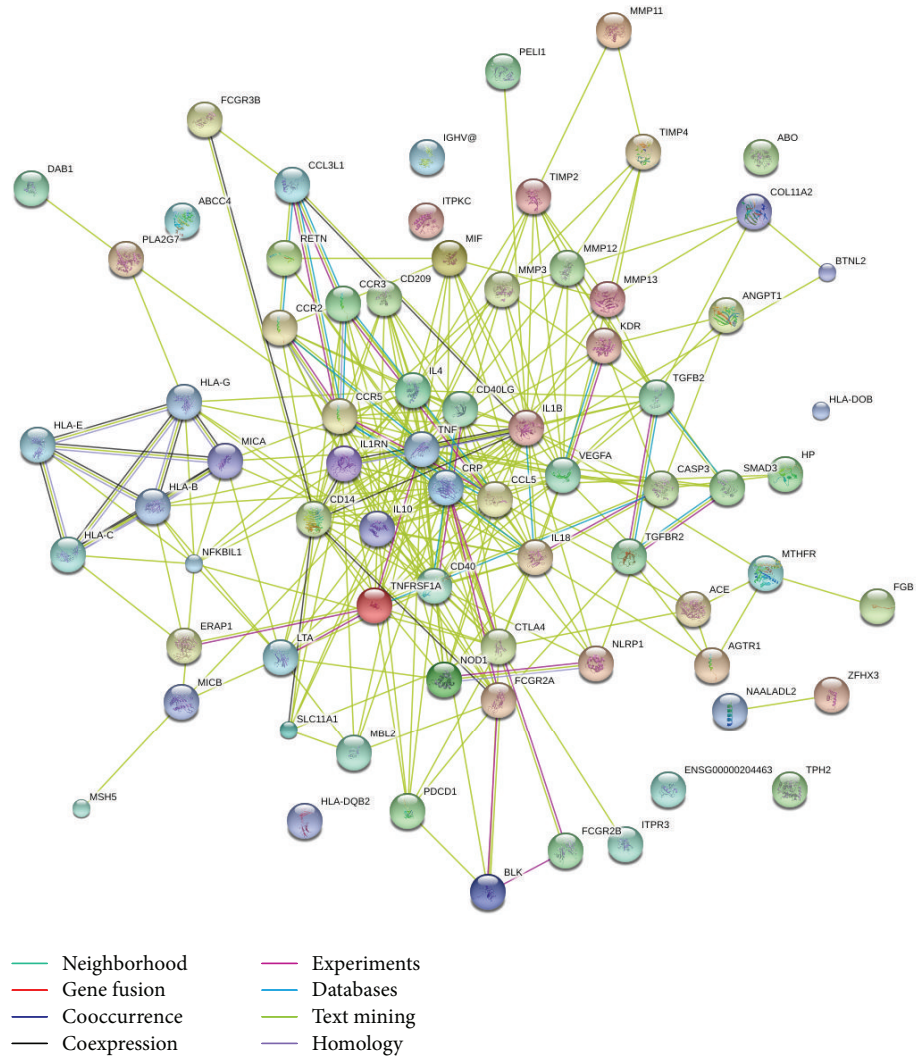


Figure 2: STRING analysis of the relationship between genes. The network nodes represent the proteins encoded by the genes. Seven different colored lines link a number of nodes and represent seven types of evidence used in predicting associations. Among these significant genes, VWA7, PRRC2A, and ABHD16A were not identified. A red line indicates the presence of fusion evidence; a green line represents neighborhood evidence; a blue line represents cooccurrence evidence; a purple line represents experimental evidence; a yellow line represents text mining evidence; a light blue line represents database evidence [100, 101] and a black line represents coexpression evidence.

in the immune response cause our great concern. In this pathway, the activation of NFAT proteins is induced by the engagement of receptors that are coupled to the calcium/calcineurin signals, such as the antigen receptors that are expressed by T cells (TCR) and B cells (BCR), the Fc-epsilon receptors (e.g, Fc epsilon R1) that are expressed by mast cells and basophil cells or receptors coupled to heterotrimeric G-proteins (e.g., CCR3 on eosinophils) [123, 124] (Figure 3).

The NFAT signal is activated in T cell and can promote the expression of the immune-related genes. Antigen presenting cells present antigenic peptides to the T helper cell via major histocompatibility complex, class (II) (MHC class II). MHC class II can upregulate the expression of CD4+T cells and downregulate the expression of CD8+T cells which has been confirmed in acute phase of KD. Then, MHC class II peptides activate the T-cell receptor (TCR alpha/beta-CD3 complex) that starts a signal leading to the increase in cytosolic Ca(II) through both the transient release of calcium from intracellular stores and the influx of calcium through Ca(II) channels. That leads to activation of the calcium-regulated phosphatase, Calcineurin A. The activated Calcineurin A cleaves an inhibitory phosphate residue from the transcription fator NF-AT (e.g., NF-AT1 and NF-AT2). Consequently, NF-AT is transported into the nucleus, where it cooperates with other transcription factors for promoter binding and thereby induces the expression of cytokines and many other T-cell-activation-induced proteins. NF-AT in T cells is critical for the expression of a number of immunologically important genes, including IL-2, IL-4, IL-5, and IL-13, as well as several related membrane-bound proteins such as CD40 Ligand (CD40L) and Fas Ligand (Fasl) [125–127].

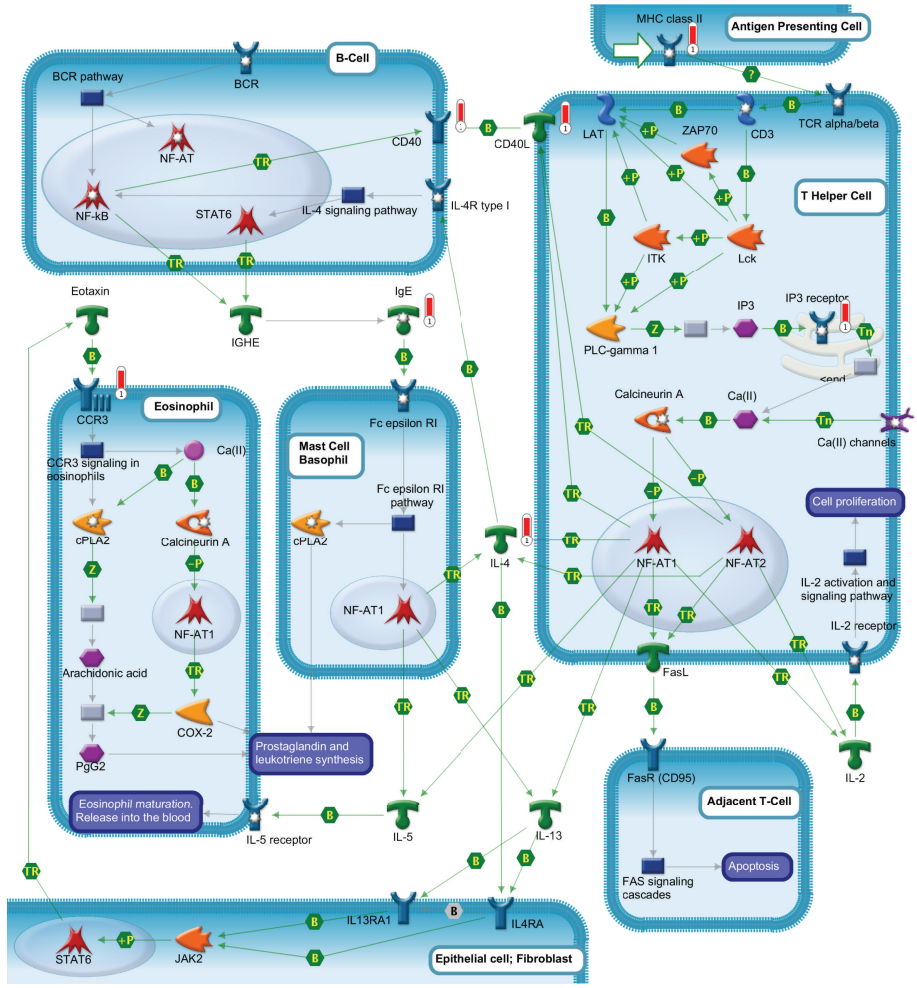


Figure 3: NF-AT signaling and leukocyte interactions have been enriched by GeneGo.

IL-4 plays an important role in cell-to-cell activation to activate NFAT signal to release leukotrienes and prostaglandins. Activated by NFAT signal in T cell, IL-4 activates nearby B cells that express corresponding receptor, IL-4R. In conjunction with BCR, IL-4 signaling pathway leads to the activation of several transcription factors, including nuclear factor kappa-B(NF-κB), signal transducer, and activator of transcription 6 (STAT6), that regulate immunoglobulin class switching and the production of immunoglobulin E (IgE) by some B cells [128–130]. IgE in turn activates NF-AT1 translocation and function in mast cells and basophils through the IgE receptor (Fc epsilon R1) leading to production of an array of cytokines, including IL-4, IL-5, and IL-13 [131, 132]. Fc epsilon R1 pathway also leads to activation of the cytosolic phospholipase A2 (Cpla2) that contributes to the secretion of leukotrienes and prostaglandins, the main mediators of inflammatory response [133]. IL-4 and IL-13, in turn, activate epithelial cells and/or fibroblasts to release eosinophil-activating cytokines, such as chemokine ligand 11 (Eotaxin). These cytokines recruit eosinophils to the inflammatory focus in the tissue and induce intracellular signaling, mainly via chemokine receptor 3 (CCR3) activation, which leads to the leukotrienes and prostaglandins synthesis and also can use NF-AT1 transcription complex to activate cytokines and chemokines. IL-4 plays an important role in the interaction between the leukocytes and induces the release of variety of inflammatory mediators.

Additionally, CD40L activates nearby B cells that express corresponding receptor CD40. IL-2 binds to IL-2 receptors at the T Cells surface to drive clonal expansion of the activated cell that induces autocrine proliferation [124]. Fasl activates the adjacent T Cells via binding to its receptors; FasR (CD95) [134] mediates apoptosis through the FAS signaling cascades (apoptosis). Fas-Fas ligand system has been considered to be involved in inducing apoptosis in KD resulting in marked decrease of peripheral blood lymphocytes [135].

4.3.1. What Is the NFATs? NFATs are nuclear factors of activated T cells. The NFAT family consists of five members: NFAT1, NFAT2, NFAT3, NFAT4, NFAT4, and NFAT5. Four (except NFAT5) of these proteins are regulated by calcium signaling and four (except NFAT3) are expressed in the immune system [124]. They are initially identified as

Ca^{2+}-sensitive transcription factors that regulate gene transcription in response to intracellular Ca^{2+} signals. NFAT family members are expressed by almost every cell type, including the immune system and nonimmune cells, contributed to the regulation of immune response, as well as development and differentiation. In the immune system, NFATs have pivotal roles in the development and function of immune organs and regulate numerous physiological processes. With the best described effects on T cell activation and phenotype, NFATs also regulate gene expression in other immune cells such as B cells [136], mast cells [137, 138], eosinophils [139], basophils [140] and NK cells [141], macrophage [142], and dendritic cells [143]. They can regulate the release of various cytokines in immune cells. In nonimmune cells, they regulate development and differentiation in a variety of organ systems [134]. It has been examined that they control gene expression during remodeling and are activated by growth factors [144, 145] or histamine [146] in the endothelium, contributing to cell growth, remodeling of smooth muscle cells [147–149], and vascular development and angiogenesis [150–152] (including the isoforms c1 and c3) and are activated in response to inflammatory processes [153] and high intravascular pressure [154] in the vascular system. The isoforms NFATc3 and NFATc4 are active during pathophysiological conditions that affect the cardiovascular system, including atrial fibrillation [155, 156] and hypertrophy [157]. Loss of specific NFAT isoforms has been found to result in cardiovascular, skeletal muscle, cartilage, neuronal, or immune system defects [158–162]. Therefore, we can conclude that the Ca^{2+}/NFAT pathway plays a wide range role in inflammatory processes, immune responses, and the remodeling of vascular tissues. All of these physiological processes occur in KD. It is suggested that the Ca^{2+}/NFAT pathway may involve in the pathological processes of KD.

4.3.2. The Upstream Adjustment Signals of NFAT Signal. NFATs are mainly Ca^{2+}-sensitive transcription factors that regulate gene transcription in response to intracellular Ca^{2+} signals. Four (except NFAT5) of these proteins are regulated by calcium signaling. Activity of NFATs is regulated by phosphorylation. Inactive NFATs are highly phosphorylated and localized in the cytoplasm. Intracellular Ca^{2+} signals activate the calmodulin-dependent serine/threonine phosphatase calcineurin (CaN), which dephosphorylates NFATs and induces translocation to the nucleus.

Inositol-trisphosphate 3-kinase C (ITPKC) is a negative regulator of the Ca^{2+}/NFAT pathway. NF-AT signaling was first mentioned to be associated with regulation of ITPKC in the KD. ITPKC is a kinase of inositol 1,4,5-triphosphate (IP3) which is a second messenger molecule that releases calcium from the endoplasmic and sarcoplasmic reticulum. First identified by genome-wide study and following confirmation by candidate genetic studies in both Japanese, Taiwanese and US children, ITPKC was considered to be associated with KD which confers both susceptibility to KD and the risk for CAL and IVIG resistance [26, 94, 95, 163], which has been thought to be involved in the Ca^{2+}-dependent NFAT signaling pathways in T cells. It has been considered that C allele of rs28493229 in ITPKC can reduce the splicing efficiency of the ITPKC mRNA, inducing the hyperactivation of Ca^{2+}-dependent NFAT signal in T cells, leading to a reduction in the phosphorylation of IP3 to IP4, resulting in the increase of IP3 levels. This would result in an increase of calcium levels and excessive activation of the NFAT signal, thus leading to immune dysregulation.

Caspase-3 (CASP3) is a key molecule of activation-induced cell death (AICD) [164]; it is profoundly related to the apoptosis of immune cells. It has also been reported to cleave the inositol 1,4,5-triphosphate receptor, type 1 (ITPR1) in apoptotic T cells (ITPR1 is a receptor for inositol 1,4,5-trisphosphate (IP3), a substrate for ITPKC in T cells [165]). Thereby, it is a positive regulatory factor of NFAT signal. Additionally, the mutation of CASP3 (rs113420705) can reduce the binding of NFAT to the DNA surrounding the SNP. Its gene variant (4q34-35, rs113420705) has been identified contributing to KD susceptibility in Euro-American triads and Taiwanese [35, 166]. Other studies [167, 168] also stated that CASP3 plays an important role in the execution phase of apoptosis of immune cells in KD.

Calcineurin inhibitors (e.g., CsA, FK506) have been extensively used as immunosuppressive agents to improve graft survival and to treat autoimmune diseases [127]. They act by blocking calcineurin enzymatic activity. CsA has been an effective [169–171] therapeutic drug in the treatment of IVIG resistance patients in KD.

4.3.3. The Downstream of Adjustment Signals of NFAT Signal: NF-κB (Nuclear Factor Kappa-B). NF-κB is another transcription factor of eukaryotes, which is evolutionarily related to the NF-AT family of transcription factors. It is activated in response to signals that lead to cell growth, differentiation, apoptosis, and other events. It takes part in expression of numerous cytokines and adhesion molecules which are critical elements involved in the regulation of immune responses.

NF-κB plays pivotal roles in the immune and inflammatory responses by regulating the interaction between CD40 and CD40L in T cells and B cells. NF-κB can be activated by IL-4 signaling pathway in B cells to induce the expression of CD40 which has been illustrated above. CD40 plays a crucial role as a costimulatory molecule in the cooperation between T and B cells. It is important in the pathogenesis of autoimmune diseases in humans and animal models such as autoimmune thyroiditis, inflammatory bowel disease, psoriasis, systemic lupus erythematosus, allergic encephalomyelitis, multiple sclerosis, rheumatoid arthritis, collagen-induced arthritis, and autoimmune type of diabetes mellitus [172–174]. CD40 signaling leads to isotype switching and autoantibody production in B cells and in T-cell priming, altering TCR expression through the expression and nuclear translocation of recombinases, which increases the risk of developing autoimmunity [173]. CD40 engagement in both T or B cells leads to the production of cytokines, such as IL-12, IL-2, TNF-α, IFN-α, and CD80, developing an environment which is conducive to autoimmune diseases [172–174].

Additionally, the interaction between CD40 and CD40L regulated by NF-κB can regulate the expression of numerous biomolecules in other cells. They can enhance the expression of cytokines (such as IL-2, IL-6, IL-10, TNF-α, lymphotoxin-α, and transforming growth factor-β by B cells; the synthesis of granulocyte macrophage colony-stimulating factor (GM-CSF) by dendritic cells and eosinophils and the synthesis of TNF-α, IL-1, IL-6, and IL-8 by peripheral blood mononuclear cells), chemokines (monocyte chemotactic protein-1 (MCP-1), IL-8, MCP-1, matrix metalloproteinases (MMP-1,-2,-3,-9,-11, and -13) by peripheral blood mononuclear cells, macrophages, endothelial and smooth muscle cells endothelial), adhesion molecules (E-selectin, vascular cell adhesion molecule-1 (VCAM-1) and intercellular adhesion molecule-1 (ICAM-1) in endothelial cells and fibroblasts), platelet-activating factor [175], prostaglandin E2 [176], vascular endothelial growth factor [177, 178], and NO [172]. which are involved in the pathophysiology of inflammatory and autoimmune diseases.

NF-κB may participate in the pathogenesis of vasculitis of KD in acute stage. Some studies have indicated that NF-κB is excessively activated in the acute phase of KD and the inhibition of NF-κB can reduce the generation of inflammatory cytokines which plays important roles in vascular damage of KD [179, 180]. NF-κB signaling pathway is a complex system; it perhaps involves in immune damage of KD in different levels. Activation of NF-κB can be used as the trigger of key links of the inflammatory response and induce the cascade release of inflammatory response factor, eventually leading to inflammatory pathological damage.

4.4. The NF-AT Signaling and Leukocyte Interactions and NF-κB Signaling Together May Be Involved in the Pathogenetic Process of KD. Given the important role of NFAT signaling and NF-κB signaling in the activation of immune system and the regulating of vascular remodeling, we speculate that the interaction between NFAT signaling and NF-κB signaling together may also be involved in the pathogenesis of KD.

Initially due to exposure to some inflammatory stimuli or certain pathogens, antigen presenting cells present antigenic peptides to the T-cell receptors via MHC class II leading to the stimulation of PLC-gamma 1 and hydrolyzation of PIP2. The second messengers IP3 in the T cells start a signal leading to the increase in cytosolic Ca(II) through both the transient release of calcium from intracellular stores and influx of calcium through Ca(II) channels. The high calcium levels lead to activation of the calcium-regulated phosphatase, Calcineurin A. The activated Calcineurin A cleaves an inhibitory phosphate residue from the transcription fator NF-AT (e.g., NF-AT1 and NF-AT2). Consequently, NF-AT is transported into the nucleus, where it cooperates with other transcription factors for promoter binding and activates T cells inducing the expression of a number of immunologically important genes including IL-2, IL-4, IL-5, IL-13, CD40 Ligand (CD40L), and Fas Ligand (Fasl). Through the leukocyte interactions, other immune cells were activated and release other inflammatory cytokines, such as leukotrienes and prostaglandins. In B cells and T-cell, CD40 signaling leads to isotype switching, autoantibody production, and altering TCR expression. CD40 signaling can also enhance the expression of cytokines, chemokines, matrix metalloproteinases, adhesion molecules, platelet-activating factors, prostaglandin E2, vascular endothelial growth factor, and NO. in other cells. The combined effect of these factors causes the vascular damage and formation of coronary artery lesions in KD. The process of NFAT signal in regulating development and differentiation was also excessively induced by the pathological damage of vasculature and then contributed to the remodeling of vascular system.

IL-4, CD40, and CD40L, which are enriched in the pathway of NF-AT signaling and leukocyte interactions and play a crucial role in the immune response and remodeling process, are located in the center position of the network (analysed by STRING) and are closely linked with the other factors. It further demonstrate the importance of this pathway.

5. Conclusions

KD is a complex disease. Many studies have shown that it is associated with a variety of gene polymorphism. Through GeneGo and DAVID analysis, we speculated that NF-AT signaling and leukocyte interactions combined with another transcription factor NF-κB may play an important role in pathological damage of KD. Their importance needs our follow-up clinical validation.

Acknowledgments

This work was financially supported by the Jiangsu Province Natural Science Foundation (no. BK2010032), Jiangsu province health department (H201127), and Suzhou Science and Technology Bureau (SYS201144). Additionally, the authors would like to thank the Systems Biology Center of Soochow University of China for their technical support.

References

[1] T. Kawasaki, "Acute febrile mucocutaneous syndrome with lymphoid involvement with specific desquamation of the fingers and toes in children," *Japanese Journal of Allergology*, vol. 16, no. 3, pp. 178–222, 1967.

[2] J. C. Burns, "Commentary: translation of Dr. Tomisaku Kawasaki's original report of fifty patients in 1967," *Pediatric Infectious Disease Journal*, vol. 21, no. 11, pp. 993–995, 2002.

[3] M. Ayusawa, T. Sonobe, S. Uemura et al., "Revision of diagnostic guidelines for Kawasaki disease (the 5th revised edition)," *Pediatrics International*, vol. 47, no. 2, pp. 232–234, 2005.

[4] H. C. Kuo, C. D. Liang, C. L. Wang, H. R. Yu, K. P. Hwang, and K. D. Yang, "Serum albumin level predicts initial intravenous immunoglobulin treatment failure in Kawasaki disease," *Acta Paediatrica*, vol. 99, no. 10, pp. 1578–1583, 2010.

[5] J. C. Burns and M. P. Glodé, "Kawasaki syndrome," *The Lancet*, vol. 364, no. 9433, pp. 533–544, 2004.

[6] Y. C. Liu, C. P. Hou, C. M. Kuo, C. D. Liang, and H. C. Kuo, "Atypical Kawasaki disease: literature review and clinical nursing," *Journal of Nursing*, vol. 57, no. 6, pp. 104–110, 2010.

[7] A. H. Rowley, S. C. Baker, J. M. Orenstein, and S. T. Shulman, "Searching for the cause of Kawasaki disease—cytoplasmic inclusion bodies provide new insight," *Nature Reviews Microbiology*, vol. 6, no. 5, pp. 394–401, 2008.

[8] Y. Nakamura, N. Yashiro, R. Uehara et al., "Epidemiologic features of Kawasaki disease in Japan: results of the 2009-2010 nationwide survey," *Journal of Epidemiology*, vol. 22, no. 3, pp. 216–221, 2012.

[9] R. Uehara and E. D. Belay, "Epidemiology of kawasaki disease in Asia, Europe, and the United States," *Journal of Epidemiology*, vol. 22, no. 2, pp. 79–85, 2012.

[10] R. C. Holman, A. T. Curns, E. D. Belay et al., "Kawasaki syndrome in Hawaii," *Pediatric Infectious Disease Journal*, vol. 24, no. 5, pp. 429–433, 2005.

[11] R. Uehara, M. Yashiro, Y. Nakamura, and H. Yanagawa, "Kawasaki disease in parents and children," *Acta Paediatrica*, vol. 92, no. 6, pp. 694–697, 2003.

[12] H. Yanagawa, Y. Nakamura, M. Yashiro et al., "Results of the nationwide epidemiologic survey of Kawasaki disease in 1995 and 1996 in Japan," *Pediatrics*, vol. 102, no. 6, article E65, 1998.

[13] Y. Nakamura, M. Yashiro, R. Uehara et al., "Epidemiologic features of Kawasaki disease in Japan: results of the 2007-2008 nationwide survey," *Journal of Epidemiology*, vol. 20, no. 4, pp. 302–307, 2010.

[14] R. Uehara, M. Yashiro, Y. Nakamura, and H. Yanagawa, "Parents with a history of Kawasaki disease whose child also had the same disease," *Pediatrics International*, vol. 53, no. 4, pp. 511–514, 2011.

[15] Y. Onouchi, "Molecular genetics of Kawasaki disease," *Pediatric Research*, vol. 65, no. 5, part 2, pp. 46R–54R, 2009.

[16] Y. Onouchi, "Genetics of Kawasaki disease: what we know and don't know," *Circulation Journal*, vol. 76, no. 7, pp. 1581–1586, 2012.

[17] K. P. Weng, K. S. Hsieh, T. Y. Ho et al., "IL-1B polymorphism in association with initial intravenous immunoglobulin treatment failure in Taiwanese children with Kawasaki disease," *Circulation Journal*, vol. 74, no. 3, pp. 544–551, 2010.

[18] K. P. Weng, K. S. Hsieh, Y. T. Hwang et al., "IL-10 polymorphisms are associated with coronary artery lesions in acute stage of Kawasaki disease," *Circulation Journal*, vol. 74, no. 5, pp. 983–989, 2010.

[19] K. P. Weng, T. Y. Ho, Y. H. Chiao et al., "Cytokine genetic polymorphisms and susceptibility to Kawasaki disease in Taiwanese children," *Circulation Journal*, vol. 74, no. 12, pp. 2726–2733, 2010.

[20] H. C. Kuo, M. C. Chao, Y. W. Hsu et al., "*CD40* gene polymorphisms associated with susceptibility and coronary artery lesions of Kawasaki disease in the Taiwanese population," *The Scientific World Journal*, vol. 2012, Article ID 520865, 5 pages, 2012.

[21] H. R. Yu, W. P. Chang, L. Wang et al., "*DC-SIGN* (CD209) promoter-336 A/G (rs4804803) polymorphism associated with susceptibility of Kawasaki disease," *The Scientific World Journal*, vol. 2012, Article ID 634835, 5 pages, 2012.

[22] R. Liu, B. He, F. Gao, Q. Liu, and Q. Yi, "Association of the resistin gene promoter region polymorphism with Kawasaki disease in Chinese children," *Mediators of Inflammation*, vol. 2012, Article ID 356362, 8 pages, 2012.

[23] S. Shrestha, H. Wiener, A. Shendre et al., "Role of activating *FcγR* gene polymorphisms in Kawasaki disease susceptibility and intravenous immunoglobulin response," *Circulation: Cardiovascular Genetics*, vol. 5, no. 3, pp. 309–316, 2012.

[24] S. Taniuchi, M. Masuda, M. Teraguchi et al., "Polymorphism of Fcγ RIIa may affect the efficacy of γ-globulin therapy in Kawasaki disease," *Journal of Clinical Immunology*, vol. 25, no. 4, pp. 309–313, 2005.

[25] S. Onoyama, K. Ihara, Y. Yamaguchi et al., "Genetic susceptibility to Kawasaki disease: analysis of pattern recognition receptor genes," *Human Immunology*, vol. 73, no. 6, pp. 654–660, 2012.

[26] M. T. Lin, J. K. Wang, J. I. Yeh et al., "Clinical implication of the C allele of the *ITPKC* gene SNP rs28493229 in kawasaki disease: association with disease susceptibility and BCG scar reactivation," *Pediatric Infectious Disease Journal*, vol. 30, no. 2, pp. 148–152, 2011.

[27] Y. Onouchi, "Identification of susceptibility genes for Kawasaki disease," *Japanese Journal of Clinical Immunology*, vol. 33, no. 2, pp. 73–80, 2010.

[28] Q. Peng, C. Chen, Y. Zhang et al., "Single-nucleotide polymorphism rs2290692 in the 3′UTR of *ITPKC* associated with susceptibility to Kawasaki disease in a Han Chinese population," *Pediatric Cardiology*, vol. 33, no. 7, pp. 1046–1053, 2012.

[29] Y. Onouchi, Y. Suzuki, H. Suzuki et al., "*ITPKC* and *CASP3* polymorphisms and risks for IVIG unresponsiveness and coronary artery lesion formation in Kawasaki disease," *Pharmacogenomics Journal*, vol. 13, no. 1, pp. 52–59, 2013.

[30] C. Shimizu, S. Jain, M. L. Hibberd et al., "Transforming growth factor-β signaling pathway in patients with Kawasaki disease," *Circulation: Cardiovascular Genetics*, vol. 4, no. 1, pp. 16–25, 2011.

[31] Y. M. Choi, K. S. Shim, K. L. Yoon et al., "Transforming growth factor beta receptor II polymorphisms are associated with Kawasaki disease," *Korean Journal of Pediatrics*, vol. 55, no. 1, pp. 18–23, 2012.

[32] K. Yamamura, K. Ihara, K. Ikeda, H. Nagata, Y. Mizuno, and T. Hara, "Histo-blood group gene polymorphisms as potential genetic modifiers of the development of coronary artery lesions in patients with Kawasaki disease," *International Journal of Immunogenetics*, vol. 39, no. 2, pp. 119–125, 2012.

[33] J. J. Kim, Y. M. Hong, S. W. Yun et al., "Assessment of risk factors for Korean children with Kawasaki disease," *Pediatric Cardiology*, vol. 33, no. 4, pp. 513–520, 2012.

[34] H.-C. Kuo, Y. Onouchi, Y.-W. Hsu et al., "Polymorphisms of transforming growth factor-β signaling pathway and Kawasaki disease in the Taiwanese population," *Journal of Human Genetics*, vol. 56, no. 12, pp. 840–845, 2011.

[35] H. C. Kuo, H. R. Yu, S. H. H. Juo et al., "*CASP3* gene single-nucleotide polymorphism (rs72689236) and Kawasaki disease in Taiwanese children," *Journal of Human Genetics*, vol. 56, no. 2, pp. 161–165, 2011.

[36] W. B. Breunis, S. Davila, C. Shimizu et al., "Disruption of vascular homeostasis in patients with Kawasaki disease: involvement of vascular endothelial growth factor and angiopoietins," *Arthritis and Rheumatism*, vol. 64, no. 1, pp. 306–315, 2012.

[37] W. B. Breunis, M. H. Biezeveld, J. Geissler et al., "Vascular endothelial growth factor gene haplotypes in Kawasaki disease," *Arthritis and Rheumatism*, vol. 54, no. 5, pp. 1588–1594, 2006.

[38] K. C. Hsueh, Y. J. Lin, J. S. Chang et al., "Association of vascular endothelial growth factor C-634 G polymorphism in Taiwanese children with Kawasaki disease," *Pediatric Cardiology*, vol. 29, no. 2, pp. 292–296, 2008.

[39] H. Kariyazono, T. Ohno, V. Khajoee et al., "Association of vascular endothelial growth factor (VEGF) and VEGF receptor gene polymorphisms with coronary artery lesions of Kawasaki disease," *Pediatric Research*, vol. 56, no. 6, pp. 953–959, 2004.

[40] Y. Y. Hsieh, C. C. Chang, C. M. Hsu, S. Y. Chen, W. H. Lin, and F. J. Tsai, "Major histocompatibility complex class I chain-related gene polymorphisms: associated with susceptibility to kawasaki disease and coronary artery aneurysms," *Genetic Testing and Molecular Biomarkers*, vol. 15, no. 11, pp. 755–763, 2011.

[41] F. Y. Huang, Y. J. Lee, M. R. Chen et al., "Polymorphism of transmembrane region of MICA gene and Kawasaki disease," *Experimental and Clinical Immunogenetics*, vol. 17, no. 3, pp. 130–137, 2000.

[42] Y. Y. Hsieh, Y. J. Lin, C. C. Chang et al., "Human lymphocyte antigen B-associated transcript 2, 3, and 5 polymorphisms and haplotypes are associated with susceptibility of Kawasaki disease and coronary artery aneurysm," *Journal of Clinical Laboratory Analysis*, vol. 24, no. 4, pp. 262–268, 2010.

[43] S. Shrestha, H. W. Wiener, A. K. Olson et al., "Functional *FCGR2B* gene variants influence intravenous immunoglobulin response in patients with Kawasaki disease," *Journal of Allergy and Clinical Immunology*, vol. 128, no. 3, pp. 677.e1–680.e1, 2011.

[44] K. C. Hsueh, Y. J. Lin, J. S. Chang et al., "Association of interleukin-10 A-592C polymorphism in Taiwanese children with Kawasaki disease," *Journal of Korean Medical Science*, vol. 24, no. 3, pp. 438–442, 2009.

[45] K. S. Hsieh, T. J. Lai, Y. T. Hwang et al., "IL-10 promoter genetic polymorphisms and risk of Kawasaki disease in Taiwan," *Disease Markers*, vol. 30, no. 1, pp. 51–59, 2011.

[46] J. Yang, C. R. Li, Y. B. Li et al., "The correlation between Kawasaki disease and polymorphisms of tumor necrosis factor α and interleukin-10 gene promoter," *Zhonghua Er Ke Za Zhi*, vol. 41, no. 8, pp. 598–602, 2003.

[47] H. S. Jin, H. B. Kim, B. S. Kim et al., "The IL-10 (-627 A/C) promoter polymorphism may be associated with coronary aneurysms and low serum albumin in Korean children with Kawasaki disease," *Pediatric Research*, vol. 61, no. 5, pp. 584–587, 2007.

[48] K. Chaudhuri, T. Singh Ahluwalia, S. Singh, G. Binepal, and M. Khullar, "Polymorphism in the promoter of the CCL5 gene (CCL5 G-403A) in a cohort of North Indian children with Kawasaki disease. A preliminary study," *Clinical and Experimental Rheumatology*, vol. 29, no. 1, supplement 64, pp. S126–S130, 2011.

[49] G. B. Wang, C. R. Li, J. Yang, P. Q. Wen, and S. L. Jia, "A regulatory polymorphism in promoter region of TNFR1 gene is associated with Kawasaki disease in Chinese individuals," *Human Immunology*, vol. 72, no. 5, pp. 451–457, 2011.

[50] H. C. Kuo, C. D. Liang, H. R. Yu et al., "CTLA-4, position 49 A/G polymorphism associated with coronary artery lesions in Kawasaki disease," *Journal of Clinical Immunology*, vol. 31, no. 2, pp. 240–244, 2011.

[51] J. A. Park, K. S. Shin, and W. K. Youn, "Polymorphism of matrix metalloproteinase-3 promoter gene as a risk factor for coronary artery lesions in Kawasaki disease," *Journal of Korean Medical Science*, vol. 20, no. 4, pp. 607–611, 2005.

[52] C. Shimizu, T. Matsubara, Y. Onouchi et al., "Matrix metalloproteinase haplotypes associated with coronary artery aneurysm formation in patients with Kawasaki disease," *Journal of Human Genetics*, vol. 55, no. 12, pp. 779–784, 2010.

[53] J. Gao, H. Y. Wang, N. J. Wu, and S. H. Zhang, "Relationship between fibrinogen Bβ-148C/T polymorphism and coronary artery lesions in children with Kawasaki disease," *Zhongguo Dang Dai Er Ke Za Zhi*, vol. 12, no. 7, pp. 518–520, 2010.

[54] J. C. Burns, C. Shimizu, E. Gonzalez et al., "Genetic variations in the receptor-ligand pair *CCR5* and *CCL3L1* are important determinants of susceptibility to Kawasaki disease," *Journal of Infectious Diseases*, vol. 192, no. 2, pp. 344–349, 2005.

[55] M. Mamtani, T. Matsubara, C. Shimizu et al., "Association of *CCR2-CCR5* haplotypes and *CCL3L7* copy number with kawasaki disease, coronary artery lesions, and IVIG responses in Japanese children," *PLoS ONE*, vol. 5, no. 7, Article ID e11458, 2010.

[56] W. B. Breunis, M. H. Biezeveld, J. Geissler et al., "Polymorphisms in chemokine receptor genes and susceptibility to Kawasaki disease," *Clinical and Experimental Immunology*, vol. 150, no. 1, pp. 83–90, 2007.

[57] W. K. Jhang, M. J. Kang, H. S. Jin et al., "The *CCR5* (-2135C/T) polymorphism may be associated with the development of kawasaki disease in Korean children," *Journal of Clinical Immunology*, vol. 29, no. 1, pp. 22–28, 2009.

[58] Y. C. Huang, Y. J. Lin, J. S. Chang et al., "Single nucleotide polymorphism rs2229634 in the *ITPR3* gene is associated with the risk of developing coronary artery aneurysm in children with Kawasaki disease," *International Journal of Immunogenetics*, vol. 37, no. 6, pp. 439–443, 2010.

[59] J. J. Sheu, Y. J. Lin, J. S. Chang et al., "Association of *COL11A2* polymorphism with susceptibility to Kawasaki disease and development of coronary artery lesions," *International Journal of Immunogenetics*, vol. 37, no. 6, pp. 487–492, 2010.

[60] J. Yang, C. R. Li, Y. B. Li, H. J. Huang, R. X. Li, and G. B. Wang, "Correlation between mannose-binding lectin gene codon 54 polymorphism and susceptibility of Kawasaki disease," *Zhonghua Er Ke Za Zhi*, vol. 42, no. 3, pp. 176–179, 2004.

[61] S. Sato, H. Kawashima, Y. Kashiwagi, T. Fujioka, K. Takekuma, and A. Hoshika, "Association of mannose-binding lectin gene polymorphisms with Kawasaki disease in the Japanese," *International Journal of Rheumatic Diseases*, vol. 12, no. 4, pp. 307–310, 2009.

[62] M. H. Biezeveld, I. M. Kuipers, J. Geissler et al., "Association of mannose-binding lectin genotype with cardiovascular abnormalities in Kawasaki disease," *The Lancet*, vol. 361, no. 9365, pp. 1268–1270, 2003.

[63] M. H. Biezeveld, J. Geissler, G. J. Weverling et al., "Polymorphisms in the mannose-binding lectin gene as determinants of age-defined risk of coronary artery lesions in Kawasaki disease," *Arthritis and Rheumatism*, vol. 54, no. 1, pp. 369–376, 2006.

[64] Y. F. Cheung, M. H. K. Ho, W. K. Ip, S. F. S. Fok, T. C. Yung, and Y. L. Lau, "Modulating effects of mannose binding lectin genotype on arterial stiffness in children after Kawasaki disease," *Pediatric Research*, vol. 56, no. 4, pp. 591–596, 2004.

[65] J. Y. Ban, S. K. Kim, S. W. Kang, K. L. Yoon, and J. H. Chung, "Association between polymorphisms of matrix metalloproteinase 11 (MMP-11) and Kawasaki disease in the Korean population," *Life Sciences*, vol. 86, no. 19-20, pp. 756–759, 2010.

[66] G. Simonini, E. Corinaldesi, C. Massai et al., "Macrophage migration inhibitory factor-173 polymorphism and risk of coronary alterations in children with Kawasaki disease," *Clinical and Experimental Rheumatology*, vol. 27, no. 6, pp. 1026–1030, 2009.

[67] K. C. Hsueh, Y. J. Lin, J. S. Chang, L. Wan, and F. J. Tsai, "*BTNL2* gene polymorphisms may be associated with susceptibility to Kawasaki disease and formation of coronary artery lesions in Taiwanese children," *European Journal of Pediatrics*, vol. 169, no. 6, pp. 713–719, 2010.

[68] S. W. Park, J. Y. Ban, K. L. Yoon et al., "Involvement of tryptophan hydroxylase 2 (TPH2) gene polymorphisms in

susceptibility to coronary artery lesions in Korean children with Kawasaki disease," *European Journal of Pediatrics*, vol. 169, no. 4, pp. 457–461, 2010.

[69] J. K. Chun, D. W. Kang, B. W. Yoo, J. S. Shin, and D. S. Kim, "Programmed death-1 (PD-1) gene polymorphisms lodged in the genetic predispositions of Kawasaki disease," *European Journal of Pediatrics*, vol. 169, no. 2, pp. 181–185, 2010.

[70] K. C. Hsueh, Y. J. Lin, J. S. Chang et al., "Influence of interleukin 18 promoter polymorphisms in susceptibility to Kawasaki disease in Taiwan," *Journal of Rheumatology*, vol. 35, no. 7, pp. 1408–1413, 2008.

[71] S. Y. Chen, L. Wan, Y. C. Huang et al., "Interleukin-18 gene 105A/C genetic polymorphism is associated with the susceptibility of Kawasaki disease," *Journal of Clinical Laboratory Analysis*, vol. 23, no. 2, pp. 71–76, 2009.

[72] Y. J. Lin, L. Wan, J. Y. Wu et al., "*HLA-E* gene polymorphism associated with susceptibility to Kawasaki disease and formation of coronary artery aneurysms," *Arthritis and Rheumatism*, vol. 60, no. 2, pp. 604–610, 2009.

[73] J. Y. Ban, K. L. Yoon, S. K. Kim, S. Kang, and J. H. Chung, "Promoter polymorphism (rs3755724, -55C/T) of tissue inhibitor of metalloproteinase 4 (TIMP4) as a risk factor for Kawasaki disease with coronary artery lesions in a Korean population," *Pediatric Cardiology*, vol. 30, no. 3, pp. 331–335, 2009.

[74] J. J. Kim, S. J. Hong, Y. M. Hong et al., "Genetic variants in the HLA-G region are associated with Kawasaki disease," *Human Immunology*, vol. 69, no. 12, pp. 867–871, 2008.

[75] Y. F. Cheung, G. Y. Huang, S. B. Chen et al., "Inflammatory gene polymorphisms and susceptibility to kawasaki disease and its arterial sequelae," *Pediatrics*, vol. 122, no. 3, pp. e608–e614, 2008.

[76] M. W. Quasney, D. E. Bronstein, R. M. Cantor et al., "Increased frequency of alleles associated with elevated tumor necrosis factor-α levels in children with Kawasaki disease," *Pediatric Research*, vol. 49, no. 5, pp. 686–690, 2001.

[77] K. Ikeda, K. Ihara, K. Yamaguchi et al., "Genetic analysis of MMP gene polymorphisms in patients with Kawasaki disease," *Pediatric Research*, vol. 63, no. 2, pp. 182–185, 2008.

[78] J. H. Oh, J. W. Han, S. J. Lee et al., "Polymorphisms of human leukocyte antigen genes in Korean children with Kawasaki disease," *Pediatric Cardiology*, vol. 29, no. 2, pp. 402–408, 2008.

[79] K. Furuno, H. Takada, K. Yamamoto et al., "Tissue inhibitor of metalloproteinase 2 and coronary artery lesions in Kawasaki disease," *Journal of Pediatrics*, vol. 151, no. 2, pp. 155.e1–160.e1, 2007.

[80] S. F. Wu, J. S. Chang, C. T. Peng, Y. R. Shi, and F. J. Tsai, "Polymorphism of angiotensin-1 converting enzyme gene and Kawasaki disease," *Pediatric Cardiology*, vol. 25, no. 5, pp. 529–533, 2004.

[81] H. S. Yoon, S. K. Hae, S. Sohn, and M. H. Young, "Insertion/deletion polymorphism of angiotensin converting enzyme gene in Kawasaki disease," *Journal of Korean Medical Science*, vol. 21, no. 2, pp. 208–211, 2006.

[82] R. Fukazawa, T. Sonobe, K. Hamamoto et al., "Possible synergic effect of angiotensin-I converting enzyme gene insertion/deletion polymorphism and angiotensin-II type-1 receptor 1166A/C gene polymorphism on ischemic heart disease in patients with Kawasaki disease," *Pediatric Research*, vol. 56, no. 4, pp. 597–601, 2004.

[83] K. Takeuchi, K. Yamamoto, S. Kataoka et al., "High incidence of angiotensin I converting enzyme genotype II its Kawasaki disease patients with coronary aneurysm," *European Journal of Pediatrics*, vol. 156, no. 4, pp. 266–268, 1997.

[84] T. Minami, H. Suzuki, T. Takeuchi, S. Uemura, J. Sugatani, and N. Yoshikawa, "A polymorphism in plasma platelet-activating factor acetylhydrolase is involved in resistance to immunoglobulin treatment in Kawasaki disease," *Journal of Pediatrics*, vol. 147, no. 1, pp. 78–83, 2005.

[85] S. F. Wu, J. S. Chang, L. Wan, C. H. Tsai, and F. J. Tsai, "Association of IL-1*Ra* gene polymorphism, but no association of IL-1β and IL-4 gene polymorphisms, with Kawasaki disease," *Journal of Clinical Laboratory Analysis*, vol. 19, no. 3, pp. 99–102, 2005.

[86] J. C. Burns, C. Shimizu, H. Shike et al., "Family-based association analysis implicates IL-4 in susceptibility to Kawasaki disease," *Genes and Immunity*, vol. 6, no. 5, pp. 438–444, 2005.

[87] Y. Onouchi, S. Onoue, M. Tamari et al., "CD40 ligand gene and Kawasaki disease," *European Journal of Human Genetics*, vol. 12, no. 12, pp. 1062–1068, 2004.

[88] S. Nishimura, M. Zaitsu, M. Hara et al., "A polymorphism in the promoter of the CD14 gene (CD14/-159) is associated with the development of coronary artery lesions in patients with kawasaki disease," *Journal of Pediatrics*, vol. 143, no. 3, pp. 357–362, 2003.

[89] K. Ouchi, Y. Suzuki, T. Shirakawa, and F. Kishi, "Polymorphism of *SLC11A1* (formerly *NRAMP1*) gene confers susceptibility to Kawasaki disease," *Journal of Infectious Diseases*, vol. 187, no. 2, pp. 326–329, 2003.

[90] H. Tsukahara, M. Hiraoka, M. Saito et al., "Methylenetetrahydrofolate reductase polymorphism in Kawasaki disease," *Pediatrics International*, vol. 42, no. 3, pp. 236–240, 2000.

[91] C. C. Khor, S. Davila, W. B. Breunis et al., "Genome-wide association study identifies *FCGR2A* as a susceptibility locus for Kawasaki disease," *Nature Genetics*, vol. 43, no. 12, pp. 1241–1246, 2011.

[92] Y. Onouchi, K. Ozaki, J. C. Burns et al., "A genome-wide association study identifies three new risk loci for Kawasaki disease," *Nature Genetics*, vol. 44, no. 5, pp. 517–521, 2012.

[93] Y. C. Lee, H. C. Kuo, J. S. Chang et al., "Two new susceptibility loci for Kawasaki disease identified through genome-wide association analysis," *Nature Genetics*, vol. 44, no. 5, pp. 522–525, 2012.

[94] Y. Onouchi, M. Tamari, A. Takahashi et al., "A genomewide linkage analysis of Kawasaki disease: evidence for linkage to chromosome 12," *Journal of Human Genetics*, vol. 52, no. 2, pp. 179–190, 2007.

[95] Y. Onouchi, T. Gunji, J. C. Burns et al., "*ITPKC* functional polymorphism associated with Kawasaki disease susceptibility and formation of coronary artery aneurysms," *Nature Genetics*, vol. 40, no. 1, pp. 35–42, 2008.

[96] D. Burgner, S. Davila, W. B. Breunis et al., "A genome-wide association study identifies novel and functionally related susceptibility loci for Kawasaki disease," *PLoS Genetics*, vol. 5, no. 1, Article ID e1000319, 2009.

[97] J. J. Kim, Y. M. Hong, S. Sohn et al., "A genome-wide association analysis reveals 1p31 and 2p13.3 as susceptibility loci for Kawasaki disease," *Human Genetics*, vol. 129, no. 5, pp. 487–495, 2011.

[98] F. J. Tsai, Y. C. Lee, J. S. Chang et al., "Identification of novel susceptibility loci for Kawasaki disease in a han Chinese population by a genome-wide association study," *PLoS ONE*, vol. 6, no. 2, Article ID e16853, 2011.

[99] C. C. Khor, S. Davila, C. Shimizu et al., "Genome-wide linkage and association mapping identify susceptibility alleles in

ABCC4 for Kawasaki disease," *Journal of Medical Genetics*, vol. 48, no. 7, pp. 467–472, 2011.

[100] I. H. Choi, Y. J. Chwae, W. S. Shim et al., "Clonal expansion of $CD8^+$ T Cells in Kawasaki disease," *Journal of Immunology*, vol. 159, no. 1, pp. 481–486, 1997.

[101] K. Furuno, T. Yuge, K. Kusuhara et al., "$CD25^+CD4^+$ regulatory T cells in patients with Kawasaki disease," *Journal of Pediatrics*, vol. 145, no. 3, pp. 385–390, 2004.

[102] H. Fujiwara and Y. Hamashima, "Pathology of the heart in Kawasaki disease," *Pediatrics*, vol. 61, no. 1, pp. 100–107, 1978.

[103] S. Amano, F. Hazama, and H. Kubagawa, "General pathology of Kawasaki disease. On the morphological alterations corresponding to the clinical manifestations," *Acta Pathologica Japonica*, vol. 30, no. 5, pp. 681–694, 1980.

[104] T. J. Brown, S. E. Crawford, M. L. Cornwall, F. Garcia, S. T. Shulman, and A. H. Rowley, "CD8 T lymphocytes and macrophages infiltrate coronary artery aneurysms in acute Kawasaki disease," *Journal of Infectious Diseases*, vol. 184, no. 7, pp. 940–943, 2001.

[105] A. H. Rowley, S. T. Shulman, C. A. Mask et al., "IgA plasma cell infiltration of proximal respiratory tract, pancreas, kidney, and coronary artery in acute Kawasaki disease," *Journal of Infectious Diseases*, vol. 182, no. 4, pp. 1183–1191, 2000.

[106] C. Galeotti, J. Bayry, I. Kone-Paut, and S. V. Kaveri, "Kawasaki disease: aetiopathogenesis and therapeutic utility of intravenous immunoglobulin," *Autoimmunity Reviews*, vol. 9, no. 6, pp. 441–448, 2010.

[107] J. Kimura, H. Takada, A. Nomura et al., "Th1 and Th2 cytokine production is suppressed at the level of transcriptional regulation in Kawasaki disease," *Clinical and Experimental Immunology*, vol. 137, no. 2, pp. 444–449, 2004.

[108] H. C. Kuo, C. L. Wang, C. D. Liang et al., "Association of lower eosinophil-related T helper 2 (Th2) cytokines with coronary artery lesions in Kawasaki disease," *Pediatric Allergy and Immunology*, vol. 20, no. 3, pp. 266–272, 2009.

[109] C. L. Wang, Y. T. Wu, C. J. Lee, H. C. Liu, L. T. Huang, and K. D. Yang, "Decreased nitric oxide production after intravenous immunoglobulin treatment in patients with Kawasaki disease," *Journal of Pediatrics*, vol. 141, no. 4, pp. 560–565, 2002.

[110] M. Fujieda, R. Karasawa, H. Takasugi et al., "A novel anti-peroxiredoxin autoantibody in patients with Kawasaki disease," *Microbiology and Immunology*, vol. 56, no. 1, pp. 56–61, 2012.

[111] J. K. Chun, T. J. Lee, K. M. Choi, K. H. Lee, and D. S. Kim, "Elevated anti-α-enolase antibody levels in Kawasaki disease," *Scandinavian Journal of Rheumatology*, vol. 37, no. 1, pp. 48–52, 2008.

[112] T. Kobayashi, H. Kimura, Y. Okada et al., "Increased CD11b expression on polymorphonuclear leucocytes and cytokine profiles in patients with Kawasaki disease," *Clinical and Experimental Immunology*, vol. 148, no. 1, pp. 112–118, 2007.

[113] Y. Mitani, H. Sawada, H. Hayakawa et al., "Elevated levels of high-sensitivity C-reactive protein and serum amyloid-A late after Kawasaki disease: association between inflammation and late coronary sequelae in Kawasaki disease," *Circulation*, vol. 111, no. 1, pp. 38–43, 2005.

[114] A. Bonnefoy and C. Legrand, "Proteolysis of subendothelial adhesive glycoproteins (fibronectin, thrombospondin, and von Willebrand factor) by plasmin, leukocyte cathepsin G, and elastase," *Thrombosis Research*, vol. 98, no. 4, pp. 323–332, 2000.

[115] H. Morgan and P. A. Hill, "Human breast cancer cell-mediated bone collagen degradation requires plasminogen activation and matrix metalloproteinase activity," *Cancer Cell International*, vol. 5, article 1, 2005.

[116] K. Imai, H. Shikata, and Y. Okada, "Degradation of vitronectin by matrix metalloproteinases-1, -2, -3, -7 and -9," *FEBS Letters*, vol. 369, no. 2-3, pp. 249–251, 1995.

[117] S. D. Shapiro, "Matrix metalloproteinase degradation of extracellular matrix: biological consequences," *Current Opinion in Cell Biology*, vol. 10, no. 5, pp. 602–608, 1998.

[118] S. Netzel-Arnett, D. J. Mitola, S. S. Yamada et al., "Collagen dissolution by keratinocytes requires cell surface plasminogen activation and matrix metalloproteinase activity," *The Journal of Biological Chemistry*, vol. 277, no. 47, pp. 45154–45161, 2002.

[119] K. Sakata, K. Hamaoka, S. Ozawa et al., "Matrix metalloproteinase-9 in vascular lesions and endothelial regulation in Kawasaki disease," *Circulation Journal*, vol. 74, no. 8, pp. 1670–1675, 2010.

[120] J. Verstappen and J. W. von den Hoff, "Tissue inhibitors of metalloproteinases (TIMPs): their biological functions and involvement in oral disease," *Journal of Dental Research*, vol. 85, no. 12, pp. 1075–1084, 2006.

[121] A. R. Hannas, J. C. Pereira, J. M. Granjeiro, and L. Tjäderhane, "The role of matrix metalloproteinases in the oral environment," *Acta Odontologica Scandinavica*, vol. 65, no. 1, pp. 1–13, 2007.

[122] H. Senzaki, S. Masutani, J. Kobayashi et al., "Circulating matrix metalloproteinases and their inhibitors in patients with Kawasaki disease," *Circulation*, vol. 104, no. 8, pp. 860–863, 2001.

[123] E. Serfling, F. Berberich-Siebelt, A. Avots et al., "NFAT and NF-κB factors-the distant relatives," *International Journal of Biochemistry and Cell Biology*, vol. 36, no. 7, pp. 1166–1170, 2004.

[124] F. Macian, "NFAT proteins: key regulators of T-cell development and function," *Nature Reviews Immunology*, vol. 5, no. 6, pp. 472–484, 2005.

[125] F. Rusnak and P. Mertz, "Calcineurin: form and function," *Physiological Reviews*, vol. 80, no. 4, pp. 1483–1521, 2000.

[126] F. Macián, C. López-Rodríguez, and A. Rao, "Partners in transcription: NFAT and AP-1," *Oncogene*, vol. 20, no. 19, pp. 2476–2489, 2001.

[127] M. Lee and J. Park, "Regulation of NFAT activation: a potential therapeutic target for immunosuppression," *Molecules and Cells*, vol. 22, no. 1, pp. 1–7, 2006.

[128] M. Benekli, M. R. Baer, H. Baumann, and M. Wetzler, "Signal transducer and activator of transcription proteins in leukemias," *Blood*, vol. 101, no. 8, pp. 2940–2954, 2003.

[129] K. Silver and R. J. Cornall, "Isotype control of B cell signaling," *Science's STKE*, vol. 2003, no. 184, article pe21, 2003.

[130] T. Mizuno and T. L. Rothstein, "B cell receptor (BCR) cross-talk: CD40 engagement enhances BCR-induced ERK activation," *Journal of Immunology*, vol. 174, no. 6, pp. 3369–3376, 2005.

[131] T. Kawakami and S. J. Galli, "Regulation of mast-cell and basophil function and survival by IgE," *Nature Reviews Immunology*, vol. 2, no. 10, pp. 773–786, 2002.

[132] A. Lorentz, I. Klopp, T. Gebhardt, M. P. Manns, and S. C. Bischoff, "Role of activator protein 1, nuclear factor-κB, and nuclear factor of activated T cells in IgE receptor-mediated cytokine expression in mature human mast cells," *Journal of Allergy and Clinical Immunology*, vol. 111, no. 5, pp. 1062–1068, 2003.

[133] C. C. Leslie, "Regulation of the specific release of arachidonic acid by cytosolic phospholipase A2," *Prostaglandins Leukotrienes and Essential Fatty Acids*, vol. 70, no. 4, pp. 373–376, 2004.

[134] G. R. Crabtree and E. N. Olson, "NFAT signaling: choreographing the social lives of cells," *Cell*, vol. 109, no. 2, pp. S67–S79, 2002.

[135] H. Y. Kim, H. G. Lee, and D. S. Kim, "Apoptosis of peripheral blood mononuclear cells in Kawasaki disease," *Journal of Rheumatology*, vol. 27, no. 3, pp. 801–806, 2000.

[136] M. M. Winslow, E. M. Gallo, J. R. Neilson, and G. R. Crabtree, "The calcineurin phosphatase complex modulates immunogenic B cell responses," *Immunity*, vol. 24, no. 2, pp. 141–152, 2006.

[137] S. Monticelli, D. C. Solymar, and A. Rao, "Role of NFAT proteins in IL13 gene transcription in mast cells," *The Journal of Biological Chemistry*, vol. 279, no. 35, pp. 36210–36218, 2004.

[138] E. Ullerås, M. Karlberg, C. M. Westerberg et al., "NFAT but not NF-κB is critical for transcriptional induction of the prosurvival gene A1 after IgE receptor activation in mast cells," *Blood*, vol. 111, no. 6, pp. 3081–3089, 2008.

[139] M. C. Seminario, J. Guo, B. S. Bochner, L. A. Beck, and S. N. Georas, "Human eosinophils constitutively express nuclear factor of activated T cells p and c," *Journal of Allergy and Clinical Immunology*, vol. 107, no. 1, pp. 143–152, 2001.

[140] J. T. Schroeder, K. Miura, H. H. Kim, A. Sin, A. Cianferoni, and V. Casolaro, "Selective expression of nuclear factor of activated T cells 2/c1 in human basophils: evidence for involvement in IgE-mediated IL-4 generation," *Journal of Allergy and Clinical Immunology*, vol. 109, no. 3, pp. 507–513, 2002.

[141] J. Aramburu, L. Azzoni, A. Rao, and B. Perussia, "Activation and expression of the nuclear factors of activated T cells, NFATp and NFATc, in human natural killer cells: regulation upon CD16 ligand binding," *Journal of Experimental Medicine*, vol. 182, no. 3, pp. 801–810, 1995.

[142] A. Yarilina, K. Xu, J. Chen, and L. B. Ivashkiv, "TNF activates calcium-nuclear factor of activated T cells (NFAT)c1 signaling pathways in human macrophages," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 4, pp. 1573–1578, 2011.

[143] I. Zanoni, R. Ostuni, G. Capuano et al., "CD14 regulates the dendritic cell life cycle after LPS exposure through NFAT activation," *Nature*, vol. 460, no. 7252, pp. 264–268, 2009.

[144] E. Hofer and B. Schweighofer, "Signal transduction induced in endothelial cells by growth factor receptors involved in angiogenesis," *Thrombosis and Haemostasis*, vol. 97, no. 3, pp. 355–363, 2007.

[145] L. Hadri, C. Pavoine, L. Lipskaia, S. Yacoubi, and A. M. Lompré, "Transcription of the sarcoplasmic/endoplasmic reticulum Ca^{2+}-ATPase type 3 gene, *ATP2A3*, is regulated by the calcineurin/NFAT pathway in endothelial cells," *Biochemical Journal*, vol. 394, no. 1, pp. 27–33, 2006.

[146] V. Boss, X. Wang, L. F. Koppelman, K. Xu, and T. J. Murphy, "Histamine induces nuclear factor of activated T cell-mediated transcription and cyclosporin A-sensitive interleukin-8 mRNA expression in human umbilical vein endothelial cells," *Molecular Pharmacology*, vol. 54, no. 2, pp. 264–272, 1998.

[147] B. R. Wamhoff, D. K. Bowles, and G. K. Owens, "Excitation-transcription coupling in arterial smooth muscle," *Circulation Research*, vol. 98, no. 7, pp. 868–878, 2006.

[148] V. Boss, K. L. Abbott, X. F. Wang, G. K. Pavlath, and T. J. Murphy, "The cyclosporin A-sensitive nuclear factor of activated T cells (NFAT) proteins are expressed in vascular smooth muscle cells. Differential localization of NFAT isoforms and induction of NFAT-mediated transcription by phospholipase C-coupled cell surface receptors," *The Journal of Biological Chemistry*, vol. 273, no. 31, pp. 19664–19671, 1998.

[149] S. de Frutos, R. Spangler, D. Alò, and L. V. González Bosc, "NFATc3 mediates chronic hypoxia-induced pulmonary arterial remodeling with α-actin up-regulation," *The Journal of Biological Chemistry*, vol. 282, no. 20, pp. 15081–15089, 2007.

[150] I. A. Graef, F. Chen, and G. R. Crabtree, "NFAT signaling in vertebrate development," *Current Opinion in Genetics and Development*, vol. 11, no. 5, pp. 505–512, 2001.

[151] R. A. Schulz and K. E. Yutzey, "Calcineurin signaling and NFAT activation in cardiovascular and skeletal muscle development," *Developmental Biology*, vol. 266, no. 1, pp. 1–16, 2004.

[152] V. Horsley and G. K. Pavlath, "NFAT: ubiquitous regulator of cell differentiation and adaptation," *Journal of Cell Biology*, vol. 156, no. 5, pp. 771–774, 2002.

[153] V. N. Bochkov, D. Mechtcheriakova, M. Lucerna et al., "Oxidized phospholipids stimulate tissue factor expression in human endothelial cells via activation of ERK/EGR-1 and Ca^{++}/NFAT," *Blood*, vol. 99, no. 1, pp. 199–206, 2002.

[154] L. V. Gonzalez Bosc, M. K. Wilkerson, K. N. Bradley, D. M. Eckman, D. C. Hill-Eubanks, and M. T. Nelson, "Intraluminal pressure is a stimulus for NFATc3 nuclear accumulation: role of calcium, endothelium-derived nitric oxide, and cGMP-dependent protein kinase," *The Journal of Biological Chemistry*, vol. 279, no. 11, pp. 10702–10709, 2004.

[155] C. C. Lin, J. L. Lin, C. S. Lin et al., "Activation of the calcineurin-nuclear factor of activated T-cell signal transduction pathway in atrial fibrillation," *Chest*, vol. 126, no. 6, pp. 1926–1932, 2004.

[156] P. Tavi, S. Pikkarainen, J. Ronkainen et al., "Pacing-induced calcineurin activation controls cardiac Ca^{2+} signalling and gene expression," *Journal of Physiology*, vol. 554, part 2, pp. 309–320, 2004.

[157] R. S. Williams, "Calcineurin signaling in human cardiac hypertrophy," *Circulation*, vol. 105, no. 19, pp. 2242–2243, 2002.

[158] P. B. Bushdid, H. Osinska, R. R. Wacław, J. D. Mołkentin, and K. E. Yutzey, "NFATc3 and NFATc4 are required for cardiac development and mitochondrial function," *Circulation Research*, vol. 92, no. 12, pp. 1305–1313, 2003.

[159] J. L. de la Pompa, L. A. Timmerman, H. Takimoto et al., "Role of the NF-ATc transcription factor in morphogenesis of cardiac valves and septum," *Nature*, vol. 392, no. 6672, pp. 182–186, 1998.

[160] I. A. Graef, F. Chen, L. Chen, A. Kuo, and G. R. Crabtree, "Signals transduced by Ca^{2+}/calcineurin and NFATc3/c4 pattern the developing vasculature," *Cell*, vol. 105, no. 7, pp. 863–875, 2001.

[161] K. M. Kegley, J. Gephart, G. L. Warren, and G. K. Pavlath, "Altered primary myogenesis in NFATC3$^{-/-}$ mice leads to decreased muscle size in the adult," *Developmental Biology*, vol. 232, no. 1, pp. 115–126, 2001.

[162] A. M. Ranger, M. J. Grusby, M. R. Hodge et al., "The transcription factor NF-ATc is essential for cardiac valve formation," *Nature*, vol. 392, no. 6672, pp. 186–190, 1998.

[163] H. C. Kuo, K. D. Yang, S. H. H. Juo et al., "*ITPKC* single nucleotide polymorphism associated with the kawasaki disease in a taiwanese population," *PLoS ONE*, vol. 6, no. 4, Article ID e17370, 2011.

[164] M. Woo, R. Hakem, M. S. Soengas et al., "Essential contribution of caspase 3/CPP32 to apoptosis and its associated nuclear changes," *Genes and Development*, vol. 12, no. 6, pp. 806–819, 1998.

[165] J. Hirota, T. Furuichi, and K. Mikoshiba, "Inositol 1,4,5-trisphosphate receptor type I is a substrate for caspase-3 and is

cleaved during apoptosis in a caspase-3-dependent manner," *The Journal of Biological Chemistry*, vol. 274, no. 48, pp. 34433–34437, 1999.

[166] Y. Onouchi, K. Ozaki, J. C. Buns et al., "Common variants in *CASP3* confer susceptibility to Kawasaki disease," *Human Molecular Genetics*, vol. 19, no. 14, pp. 2898–2906, 2010.

[167] Q. J. Yi, C. R. Li, and X. Q. Yang, "Effect of intravenous immunoglobulin on inhibiting peripheral blood lymphocyte apoptosis in acute Kawasaki disease," *Acta Paediatrica*, vol. 90, no. 6, pp. 623–627, 2001.

[168] H. Tsujimoto, S. Takeshita, K. Nakatani, Y. Kawamura, T. Tokutomi, and I. Sekine, "Delayed apoptosis of circulating neutrophils in Kawasaki disease," *Clinical and Experimental Immunology*, vol. 126, no. 2, pp. 355–364, 2001.

[169] H. Hamada, H. Suzuki, J. Abe et al., "Inflammatory cytokine profiles during cyclosporin treatment for immunoglobulin-resistant Kawasaki disease," *Cytokine*, vol. 60, no. 3, pp. 681–685, 2012.

[170] A. H. Tremoulet, P. Pancoast, A. Franco et al., "Calcineurin inhibitor treatment of intravenous immunoglobulin-resistant Kawasaki disease," *Journal of Pediatrics*, vol. 161, no. 3, pp. 506.e1–512.e1, 2012.

[171] V. Raman, J. Kim, A. Sharkey, and T. Chatila, "Response of refractory Kawasaki disease to pulse steroid and cyclosporin a therapy," *Pediatric Infectious Disease Journal*, vol. 20, no. 6, pp. 635–637, 2001.

[172] U. Schönbeck and P. Libby, "The CD40/CD154 receptor/ligand dyad," *Cellular and Molecular Life Sciences*, vol. 58, no. 1, pp. 4–43, 2001.

[173] M. E. Munroe and G. A. Bishop, "A costimulatory function for T cell CD40," *Journal of Immunology*, vol. 178, no. 2, pp. 671–682, 2007.

[174] J. D. Laman, M. de Boer, and B. A. Hart, "CD40 in clinical inflammation: from multiple sclerosis to atherosclerosis," *Developmental Immunology*, vol. 6, no. 3-4, pp. 215–222, 1998.

[175] S. Russo, B. Bussolati, I. Deambrosis, F. Mariano, and G. Camussi, "Platelet-activating factor mediates CD40-dependent angiogenesis and endothelial-smooth muscle cell interaction," *Journal of Immunology*, vol. 171, no. 10, pp. 5489–5497, 2003.

[176] Y. Inoue, T. Otsuka, H. Niiro et al., "Novel regulatory mechanisms of CD40-induced prostanoid synthesis by IL-4 and IL-10 in human monocytes," *Journal of Immunology*, vol. 172, no. 4, pp. 2147–2154, 2004.

[177] M. Melter, M. E. J. Reinders, M. Sho et al., "Ligation of CD40 induces the expression of vascular endothelial growth factor by endothelial cells and monocytes and promotes angiogenesis *in vivo*," *Blood*, vol. 96, no. 12, pp. 3801–3808, 2000.

[178] P. H. Lapchak, M. Melter, S. Pal et al., "CD40-induced transcriptional activation of vascular endothelial growth factor involves a 68-bp region of the promoter containing a CpG island," *American Journal of Physiology*, vol. 287, no. 3, pp. F512–F520, 2004.

[179] T. Ichiyama, T. Yoshitomi, M. Nishikawa et al., "NF-κB activation in peripheral blood monocytes/macrophages and T cells during acute Kawasaki disease," *Clinical Immunology*, vol. 99, no. 3, pp. 373–377, 2001.

[180] W. Yin, X. Wang, Y. Ding et al., "Expression of nuclear factor-κBp65 in mononuclear cells in Kawasaki disease and its relation to coronary artery lesions," *The Indian Journal of Pediatrics*, vol. 78, no. 11, pp. 1378–1382, 2011.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 592790, 10 pages
http://dx.doi.org/10.1155/2013/592790

Research Article

An Entropy-Based Automated Cell Nuclei Segmentation and Quantification: Application in Analysis of Wound Healing Process

Varun Oswal,[1] Ashwin Belle,[1] Robert Diegelmann,[2] and Kayvan Najarian[1]

[1] *Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA*
[2] *Department of Biochemistry & Molecular Biology, Virginia Commonwealth University, Richmond, VA 23298, USA*

Correspondence should be addressed to Ashwin Belle; bellea@vcu.edu

Received 23 October 2012; Revised 22 January 2013; Accepted 26 January 2013

Academic Editor: Tianhai Tian

The segmentation and quantification of cell nuclei are two very significant tasks in the analysis of histological images. Accurate results of cell nuclei segmentation are often adapted to a variety of applications such as the detection of cancerous cell nuclei and the observation of overlapping cellular events occurring during wound healing process in the human body. In this paper, an automated entropy-based thresholding system for segmentation and quantification of cell nuclei from histologically stained images has been presented. The proposed translational computation system aims to integrate clinical insight and computational analysis by identifying and segmenting objects of interest within histological images. Objects of interest and background regions are automatically distinguished by dynamically determining 3 optimal threshold values for the 3 color components of an input image. The threshold values are determined by means of entropy computations that are based on probability distributions of the color intensities of pixels and the spatial similarity of pixel intensities within neighborhoods. The effectiveness of the proposed system was tested over 21 histologically stained images containing approximately 1800 cell nuclei, and the overall performance of the algorithm was found to be promising, with high accuracy and precision values.

1. Introduction

Analysis of microscopy images is one of the most fundamental goals in the realm of immunohistochemistry. The primary tasks involved in the analysis of histologically stained tissue sections are cell nuclei counting, detecting abnormal cell nuclei, and the presence of antigens within the target cells. Results derived from these analyses are most frequently used in the clinical setting to help diagnose a wide spectrum of pathologies. In the past, pathologists accomplished most of these tasks by the means of manual measurements; for example, the quantification of total cells and abnormal cells was performed through manual hand counting. These manual methods are not only time consuming, but the results they yield are often susceptible to inconsistency due to human error. However, as the result of recent advancements in microscopic imaging technology and computational image processing techniques [1], there has been significant growth of research towards translational computational systems that can detect, analyze, classify, and quantify cell nuclei from microscopic images. Adapting to robust automated image processing techniques for primary tasks such as cell nuclei segmentation and quantification will not only prove to be time efficient for pathologists, but these techniques will also be capable of producing consistent results.

In recent years, numerous image processing techniques have been proposed for cell nuclei segmentation [2]. While some techniques only perform the task of cell nuclei segmentation and quantification, techniques that are capable of further detecting and classifying abnormal tumors (cell nuclei) that cause various types of cancer have also been proposed. A cell nuclei segmentation algorithm incorporating unsupervised color clustering, morphological operations, and local thresholding has been proposed to distinguish the cancerous and noncancerous areas in histologically stained images and then segment the clustered cell nuclei [3]. K-means clustering

is implemented as unsupervised color clustering technique for cell nuclei segmentation in [4]. Another technique that uses contour detection and contour optimization combined with local gradient information and color deconvolution has been used to detect the optimal threshold values for nuclei segmentation [5]. Entropic-based thresholding methods for cell nuclei segmentation are proposed by Wang and Gudla et al. [6, 7].

A popular technique in the realm of image processing known as region growing is combined with a graph-cuts-based algorithm that incorporates Laplacian of Gaussian (LoG) filtering to detect cell nuclei [8]. Stained endocrine cell nuclei are segmented by a sequential thresholding algorithm that uses a Support Vector Machine (SVM) type of artificial neural network [9]. An adaptive-attention-window-(AAW-) based cell nuclei segmentation technique that exploits quadtree decomposition is proposed by Ko et al. [10]; the size of the AAW dynamically adapts to the region of interest in the input image, and the final cell nuclei segmentation is performed within each AAW. Histogram analysis and optimal local thresholding are followed by morphological procedures to segment a variety of cell nuclei found within the bladder and skin tissue [11]. Watershed segmentation and adaptive thresholding methods are also widely used to achieve automated segmentation of cell nuclei [12–15]. Singh et al. propose the use of a feedforward backpropagation neural network for the classification of the segmented cell nuclei into two categories: benign and malignant breast tumor (nuclei); the proposed neural network is also capable of classifying the detected malignant breast tumor in terms of type 1, type 2, and type 3 [15].

The selection of the above publications exemplifies the wide range of image processing techniques and practical diagnostics applications that encompass the realm of cell nuclei segmentation in immunohistochemistry. The well-established cell nuclei segmentation and differential immunostaining techniques that have proven to be so valuable in the cancer field are now being applied to the field of wound healing research [16]. It is of interest to note that many of the characteristics and cell functions that are manifest in cancer are also found during wound healing [3]. New research strategies to explore human wound healing are now available and allow for the in-depth investigation of the specific cell types that participate in the highly orchestrated events that occur during tissue injury and repair [4, 17]. Development of such translational computation medical systems has been and will be providing invaluable insight into understanding the complex nature of the wound healing process [18, 19].

The vast amount of research in this realm also emphasizes on the need for automated computational systems for cell segmentation techniques that produce accurate and reproducible results. However, the task of cell segmentation is still one of the most challenging tasks in biomedical image processing mainly because the histological specimens that are used for the image acquisition process are 2-dimensional sections of 3-dimensional tissue samples [8]. Images acquired from 2-dimensional histological specimen often contain cells with uneven distribution of color intensities, weak edges, and even incomplete nuclei. These are some key characteristics of microscopic histology images due to which the development of robust automated cell segmentation techniques still remains a challenge. Marker-based watershed segmentation techniques rely on automated detection of marker positions to perform accurate segmentation, however, the task of detecting the number of markers and their positions is not trivial, and over segmentation is often evident in the results. Simple edge detection-based techniques perform well in regions with strong edges but tend to cause over segmentation in regions with weak or poorly defined edges. Active contour or snake methods perform better on cell nuclei with weak edges, but these techniques often require supervision or optimized configuration files with priori information for parameter settings [20]. Hence, there is a need for an automated segmentation system that extracts cell information without requiring any user input.

Thresholding techniques are fairly simple but still effective; they are widely used for the segmentation of histological images because the regions of interest within these images are distinguishable from the other components by visual features such as color and texture [21, 22]. The entropy-based thresholding algorithm presented in this paper uses the color intensity information of pixels and the spatial correlation between pixel intensity values to segment cell nuclei. The proposed technique segments a histological image by classifying it into object (cell nuclei) and background regions; all pixels with intensity values greater (or lesser) than a global threshold value are grouped as the objects, while the remaining pixels are classified as the background. Most of the aforementioned techniques usually perform the cell nuclei segmentation on histological images stained by the Hematoxylin & Eosin (H&E) stain, whereas the proposed entropy-based thresholding algorithm segments cell nuclei from images stained by H&E and three more immunostains for specific cell phenotypes. Endothelial lineage cells were identified by the presence of platelet endothelial cell adhesion molecule cluster of differentiation 31 on the cell surface (CD-31); macrophage functional cells were identified by the presence of a specific cytoplasmic granule found in macrophages called CD-68; and contractile functioning cells were characterized by the presence of intracellular alpha-smooth muscle actin (SMActin).

The following section explains the image acquisition and histological procedures involved in the preparation of the testing dataset. Section 3 thoroughly explains the methods and mathematical formulae involved in the computation of the proposed technique. Section 4 presents the results that were obtained by testing the proposed segmentation technique on a dataset of 21 immunohistochemically stained images, and finally conclusions and future work are briefly discussed in Section 5.

2. Data Preparation

The dataset used in the testing of the proposed algorithm consists of 21 immunohistochemically stained images. The images in the dataset were acquired from human tissue

sections derived from PTFE (expanded polytetrafluoroethylene) tubes that were removed at 5, 7, and 14 days after implantation [5]. The image acquisition process was performed in the described timely manner to characterize the 4 distinct overlapping phases: hemostasis, inflammation, proliferation, and remodeling that occur during the healing process of simple acute wounds. The 4 phases are associated with biological markers and some distinct but overlapping cellular events that can be observed through change in features, such as number of cell nuclei and size of the average nuclei in a tissue section.

The tissue collection and histological staining procedures involved in the preparation of the data set are as follows. Using alcohol and povidone-iodine topical antiseptic, the site of implantation was sterilized and anesthetized using 3 cc lidocaine (1%) without epinephrine. Five, 6.0 cm, of high-porosity PTFE (polytetrafluoroethylene Custom Profile Extrusions, Tempe, AZ) tubes were implanted subcutaneously into the inner aspect of the upper arms of a healthy volunteer subject. Standardized placement was made by a 5.5 cm cannulation of the subcutaneous tissue in a proximal direction. Using a sterile 14-gauge trochar containing PTFE tubing, the skin was punctured, and the trochar was inserted subcutaneously arising through the skin 5.5–6.0 cm away. The trochar was then removed, and the proximal and distal ends of the PTFE tubing were sutured to the skin using a single 5.0 nylon suture. The implantation site was covered with antibiotic ointment and a transparent surgical dressing. On day 14, the PTFE tube was removed and stored in 10% formalin. The wound tissue contained within the fixed PTFE tube was then processed and embedded in paraffin, and 5 micron sections were prepared using standardized histologic techniques. Positive and negative control sections were included to ensure reproducible staining. Hematoxylin & Eosin (H&E) stain was used to highlight the cellular components, and standard immunostaining techniques were used to identify endothelial cells (CD-31), macrophages (CD-68), and contractile cells (α-SMActin).

The derived tissue sections were examined using a Zeiss LSM 510 NLO Meta confocal/multiphoton laser scanning microscope. For confocal imaging, the 488, 561, and 633 nm laser lines were used for sample imaging. Images were collected using sequential illumination (i.e., one laser per channel) to avoid signal cross-talk amongst channels. The images were collected using a 63x/1.4 n.a. oil immersion lens (for single photon confocal imaging) or a 63x/1.2 n.a. IR water immersion objective (for multiphoton imaging). The human study was carried out under the approval of the Institutional Review Board of Virginia Commonwealth University, School of Medicine (IRB number 11087).

3. Methodology

This section provides an in-depth explanation of the proposed entropy-based image segmentation technique. The flowchart in Figure 1 illustrates the steps involved in the cell nuclei segmentation process. The proposed algorithm is composed of four steps; image preprocessing, entropy-based thresholding, post-processing, and cell nuclei quantification.

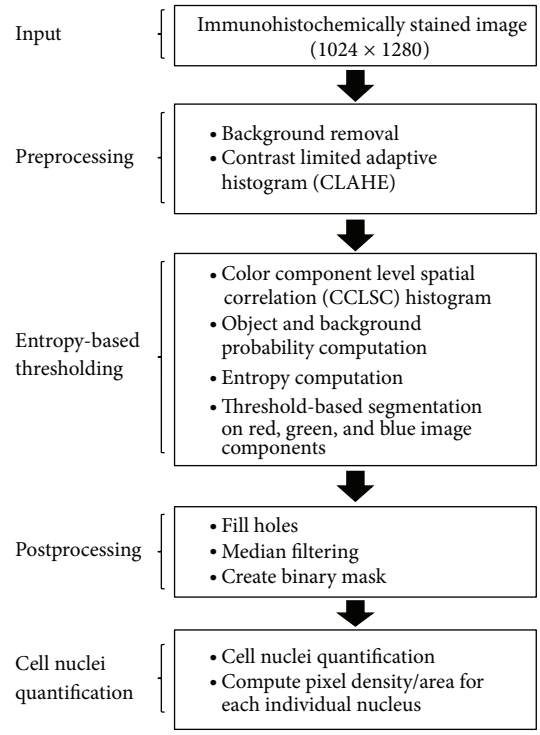


Figure 1: Overview of the proposed algorithm.

3.1. Preprocessing. The preprocessing of an immunohistochemically stained input image I begins with a background removal process. The background removal process eliminates all the white space (background) that is captured in image I due to the empty spaces present on microscopic slides.

Although there are several options with respect to color spaces wherein the processing of the image can be performed, for this project the cell extraction is primarily performed in the RGB color space. Other color spaces such as YCbCr, LAB, and HSV were tested, and in comparison the RGB color space consistently provided the best results. This is because the objective is to extract cell structures based on the color information present within the image, hence distinguishing the different biological objects within the image. The background removal process starts by separating the RGB color image I into its red, green, and blue color components to produce 3 color component images I_r, I_g, and I_b, respectively. The local range of the component images I_r, I_g, and I_b is then computed by finding range values for each individual pixel contained in I_r, I_g, and I_b. The result of this operation yields three output images J_r, J_g, and J_b, in which the value of each output pixel is its local range value, that is, the difference between the maximum and the minimum pixels values within a 3-by-3 neighbourhood surrounding the output pixel. Next, the images J_r, J_g, and J_b are multiplied together, and the resulting image is then converted to a binary mask B. The pixels with value 1 in mask B represent the region of interest,

and the pixels with value 0 represent the eliminated white space in the input image. Image B typically contains noisy components of isolated pixels that are eliminated by median filtering. Finally, the image B is individually multiplied to I_r, I_g, and I_b to produce the three color component images K_r, K_g, and K_b. The RGB image composed of images K_r, K_g, and K_b represents only the stained tissue section containing the cell nuclei. The results of the background removal process are illustrated in Figure 2.

A popular histogram equalization technique called Contrast Limited Adaptive Histogram Equalization (CLAHE) is then used in its orignal form to enhance the local contrast in the color component images K_r, K_g, and K_b. Lighting and illumination conditions are very crucial in the acquisition of microscopic images and these conditions are not necessarily the same in every microscopic setup. Different levels of lighting conditions can usually cause differences in the gray-level distribution of pixels in images [23, 24], and therefore, CLAHE is used in the proposed algorithm to uniformly equalize the varying gray-level distributions in any stained immunohistochemical input image.

In CLAHE, contrast enhancement is performed locally in small regions called "tiles", each tile's histogram is equalized to provide a better overall visual distinction between target objects (cell nuclei) and background (intercellular matter). Additionally, the use of CLAHE ensures that the stained cell nuclei in the tissue section are enhanced uniformly, thereby providing accurate recognition of cell nuclei irrespective of the influences of different staining procedures. The histograms derived from the operation of CLAHE are chosen to maintain a uniform shape. The computation of entropy in the proposed algorithm is based on Shannon's entropy, for which the net information values are calculated within 5-by-5 neighbourhoods throughout the image; the number of tiles for CLAHE's operation is chosen to be close to the total number of 5-by-5 neighbourhoods present in the input image. Once the contrast enhancement is performed on K_r, K_g, and K_b, a series of probability and entropy computations are performed on each component image to determine its optimal threshold value for cell nuclei segmentation.

3.2. Entropy-Based Thresholding. After CLAHE is performed on the 3 color component images, the thresholding technique described below is applied to each component image. The proposed entropy-based thresholding method has 3 steps: computation of the color component level spatial correlation (CCLSC) histogram, computation of object and background probabilities, and the computation of object and background entropies. Threshold values for each color component image are obtained once all the calculations are performed. The mathematical computations involved in each step are described in the following sections.

3.2.1. Computation of CCLSC Histogram. The entropy-based thresholding technique relies on the CCLSC histogram which is a modified version of the Grey-level spatial correlation histogram presented in [25]. Two probability distributions are required for the computation of the CCLSC histogram, a histogram distribution of each color component image and a distribution of similarity indices within pixel neighbourhoods that is defined below.

Let F denote a color component input image of size $P \times Q$ which has a color intensity value $f(x, y)$ for a pixel located at coordinate (x, y) in image F. The set of all color intensity values is denoted by the set $G = \{0, 1, \ldots, 255\}$. The similarity index $g(x, y)$ for a pixel located at (x, y) is computed by determining the number of surrounding pixels that have intensity values within an ϵ difference of that pixel's intensity value $f(x, y)$. Similarity indices are computed within $N \times N$ pixel neighbourhoods, where N is a positive odd number and ϵ is a number between 0 and $N \times N$. The choice of values for N and ϵ depends on characteristics of the input images. Smaller values for N and ϵ work well on images with lower magnifications such as 10x or 20x, and bigger N and ϵ values work better on images with higher magnification. Values for N and ϵ are both empirically chosen to be 5 for the implementation of this algorithm. The similarity index uses spatial correlation information of pixel intensity values to preserve important image information such as the edges of cell nuclei. The objects and background regions often tend to have high similarity indices, whereas the features such as edges produce discontinuities in the image's neighbourhoods and are therefore associated with lower similarity indices.

The similarity index is then computed within every pixel neighbourhood, and the degree of similarity is based on the difference of intensity values between the pixel located at the center of the 5×5 neighbourhood and all other pixels in the neighbourhood. The lowest similarity count that any neighbourhood can have is 1, that is, a neighbourhood with no other pixel values within an ε difference of the center pixel's intensity value will have a similarity index of 1. In any $N \times N$ neighbourhood where N is an odd number, the color intensity value of the pixel located at the center of that neighbourhood can be denoted by

$$f(\text{center}) = f\left(\frac{N+1}{2}, \frac{N+1}{2}\right). \tag{1}$$

The similarity index $g(x, y)$ for a $N \times N$ neighbourhood is mathematically expressed as

$$g(x, y) = \sum_{i=1}^{N}\sum_{j=1}^{N} \begin{cases} 1, & \text{if } \left|f(i, j) - f(\text{center})\right| \leq \epsilon, \\ 0, & \text{if } \left|f(i, j) - f(\text{center})\right| > \epsilon. \end{cases} \tag{2}$$

The CCLSC histogram is then computed by combining the probability distribution of the histogram of the color component image and the probability distribution of the similarity indices of neighbourhoods. The CCLSC histogram $h(k, m)$ is mathematically defined as

$$h(k, m) = \text{Prob}\left(f(x, y) = k, g(x, y) = m\right), \tag{3}$$

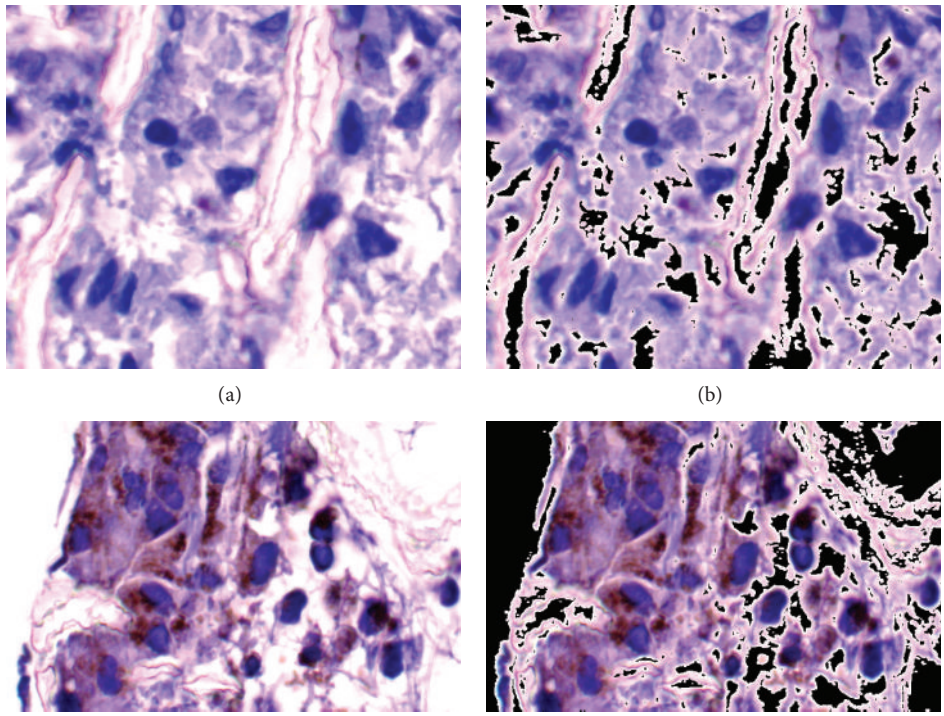


FIGURE 2: Results from background removal process. (a), (c) Input images. (b), (d) Resulting images composed of K_r, K_g, and K_b color component images.

where k is a value in the set G and m is a similarity index that can have values in the range $\{1, \ldots, N \times N\}$. The normalized CCLSC histogram $\widehat{h}(k, m)$ is then computed by

$$\widehat{h}(k,m) = \frac{\text{No. of pixels with } f(x,y) = k}{P \times Q} * \frac{\text{No. of neighbourhoods with } g(x,y) = m}{\text{Total no. of neighbourhoods}}. \quad (4)$$

Once the CCLSC $\widehat{h}(k, m)$ has been computed, it is used to determine the object and background entropies of the color component image. Figure 3 shows the surface plot of an instance of the CCLSC histogram.

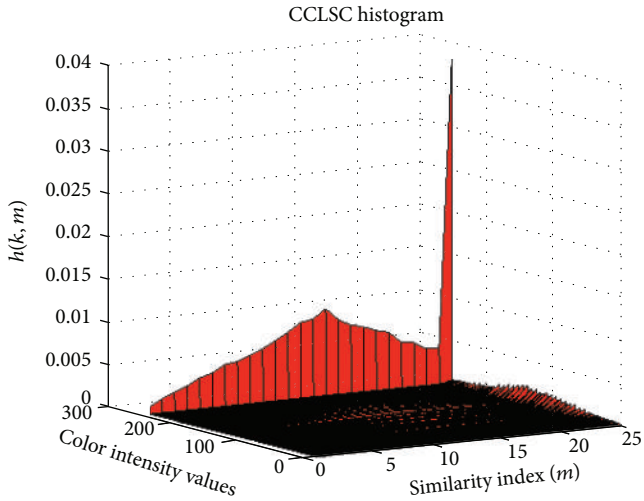


FIGURE 3: Surface plot of the CCLSC histogram.

3.2.2. Object and Background Probability Distributions. The calculation of object and background entropies require probability distributions associated with an image's object and background regions that are derived in the following way. A threshold value t is needed to segment an image's object from its background, t's value is chosen such that it partitions the set of color intensities G into 2 subsets, G_O and G_B. Let $G_O = \{0, 1, 2, \ldots, t\}$ be the set of pixel values that represent the objects, and let $G_B = \{t+1, t+2, \ldots, 255\}$ be the set of pixel values that represent the background region. The probability distribution for the image's object is expressed as

$$\left[\frac{\widehat{h}(0,1)}{P_O(t)}, \ldots, \frac{\widehat{h}(0, N \times N)}{P_O(t)}, \frac{\widehat{h}(1,1)}{P_O(t)}, \ldots, \frac{\widehat{h}(1, N \times N)}{P_O(t)}, \ldots, \frac{\widehat{h}(t, N \times N)}{P_O(t)}\right], \quad (5)$$

and the distribution associated with the background is given by

$$\left[\frac{\widehat{h}(t+1,1)}{P_B(t)},\ldots,\frac{\widehat{h}(t+1,N\times N)}{P_B(t)},\right.$$
$$\left.\frac{\widehat{h}(t+2,1)}{P_B(t)},\ldots,\frac{\widehat{h}(255,N\times N)}{P_B(t)}\right], \tag{6}$$

where

$$P_O(t)=\sum_{k=0}^{t}\sum_{m=1}^{N\times N}\widehat{h}(k,m),$$
$$P_B(t)=\sum_{k=t+1}^{255}\sum_{m=1}^{N\times N}\widehat{h}(k,m), \tag{7}$$
$$P_O(t)+P_B(t)=1.$$

3.2.3. Computation of Object and Background Entropies. The probability distributions described in the previous section are used to compute the object and background entropies. According to the principle of Shannon's entropy, the measure of uncertainty from a source equals the net value of information obtained from the source. Features such as noise and edges are associated with higher entropy values because they produce discontinuities between the object and the background which produce more uncertainity in images, that is, net information. As noted in Section 3.2.1, the background and object regions often have higher similarity index values (m), whereas edges often have values that lie in the mid range of the set $\{1,\ldots,N\times N\}$. A weight function is used in the computation of entropy to assign higher weights to the range of similarity indices that often represent edges of cell nuclei. The weight equation is

$$\text{weight}\,(m,N)=5e^{-(m-(N\times N/2))^2/32}, \tag{8}$$

where N is a positive odd number and m is a number in the set $\{1,\ldots,N\times N\}$. Figure 4 illustrates the weight function's emphasis in the calculation of the object and background entropy values.

The object entropy $H_O(t,N)$ is computed as

$$H_O(t,N)=-\sum_{k=0}^{t}\sum_{m=1}^{N\times N}\frac{\widehat{h}(k,m)}{P_O}\ln\left[\frac{\widehat{h}(k,m)}{P_O}\right]\text{weight}(m,N), \tag{9}$$

and the background entropy $H_B(t,N)$ is computed as

$$H_B(t,N)=-\sum_{k=t=1}^{255}\sum_{m=1}^{N\times N}\frac{\widehat{h}(k,m)}{P_B}\times\ln\left[\frac{\widehat{h}(k,m)}{P_B}\right]\text{weight}(m,N). \tag{10}$$

After the computation of entropies the function $\Phi(t,N)$ is maximized to yield the optimal threshold value T that will be used to segment the input image's target objects (cell nuclei) from the background. The function $\Phi(t,N)$ is expressed as

$$\Phi(t,N)=H_O(t,N)+H_B(t,N), \tag{11}$$

and T is given by

$$T=\text{maximum}(\Phi(t,N)). \tag{12}$$

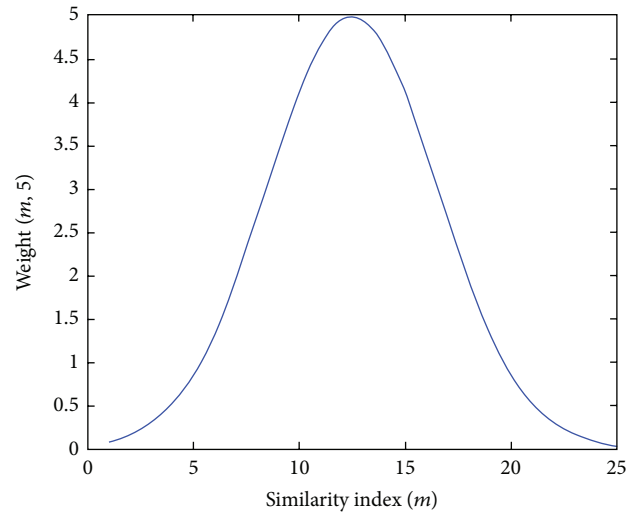
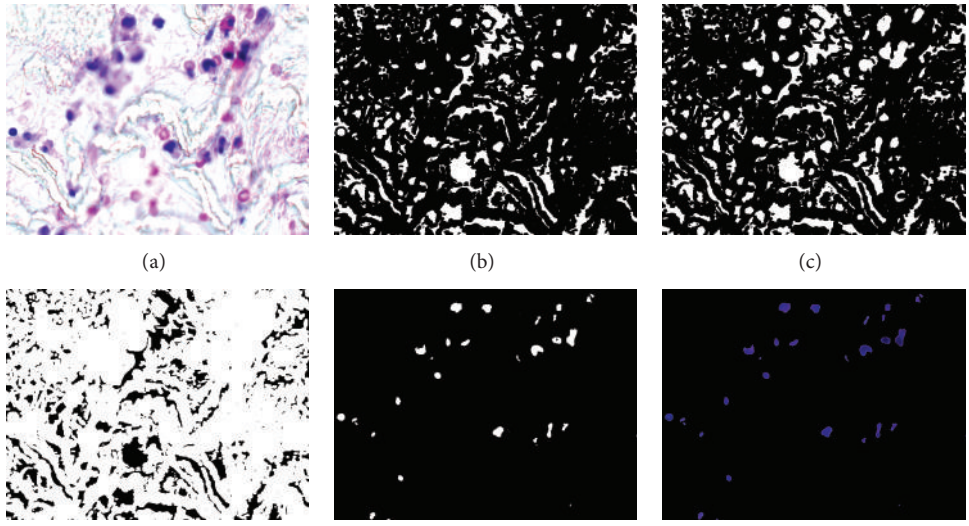


Figure 4: Graph of weight function weight $(m,5)$.

3.2.4. Thresholding-Based Segmentation. Performing the procedures described in previous section on the preprocessed images K_r, K_g, and K_b yields 3 output threshold values T_r, T_g, and T_b. The 3 threshold values are used to segment the red, green, and blue color pixel components representing cell nuclei in the images I_r, I_g, and I_b, respectively.

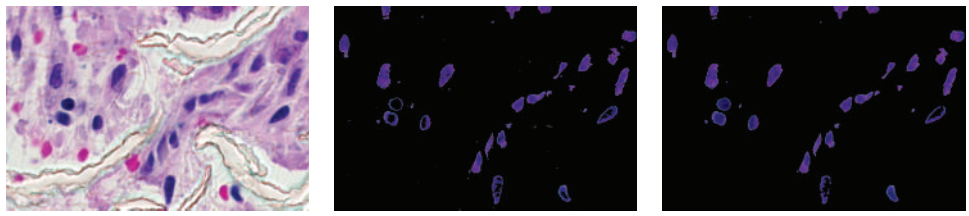
It was experimentally observed that the pixels representing cell nuclei in histologically stained images are composed of lower intensity values in the red and green color component images and higher intensity values in the blue component color image. Therefore, all pixels in I_r and I_g that have intensity values below the output threshold values T_r and T_g are considered to represent cell nuclei, whereas all pixels in I_B that have intensity values greater than T_B are considered to represent the cell nuclei. In the process of segmentation, three binary images B_r, B_g, and B_b are constructed in which all pixels that are considered as objects in I_r, I_g, and I_b, are valued as 1 at their respective locations within the binary images. The background regions are represented by pixels with value 0. The binary images B_r, B_g, and B_b, are then multiplied together to produce a binary image BW that contains only the segmented cell nuclei. In order to obtain the resulting cell nuclei segmentation in its original color, the color component images I_r, I_g, and I_b, are individually multiplied to the binary mask BW; the resulting red, green, and blue color component images are then combined to yield an image R in RGB color space which contains the results of the segmentation process. The procedures described in this section are illustrated in Figure 5.



(a) (b) (c)

(d) (e) (f)

Figure 5: (a) Input image I. (b), (c), and (d) Binary images B_r, B_g, and B_b. (e) Binary mask $BW = (B_r * B_g * B_b)$ of segmented cell nuclei. (f) Segmented cell nuclei in resulting RGB image R.



(a) (b) (c)

Figure 6: (a) Input image I. (b) Segmented image R. (c) Postprocessed image R.

3.3. Post Processing. The image R obtained from the segmentation process contains the extracted cell nuclei, but it also contains unwanted noise that occurs due to similarities in color intensities of cellular and other noncellular regions. The postprocessing procedure attempts to remove most of the unwanted noise so that the results of the cell nuclei quantification contain fewer false positives. The process consists of 2 steps, firstly, a morphological technique called fill-holes is applied on all 3 color component images of R. The fill-holes operation is useful in maintaining some structural details of cell nuclei that may have been lost during the process of segmentation.

The second step in the postprocessing of image R is median filtering. Median filtering is a nonlinear operation that is widely used to reduce salt and pepper noise in images. Median filtering is performed on each color component layer of R. After the median filtering, the image R is denoised and is ready for cell nuclei quantification. Results of the fill-holes and median filtering operations can be observed in Figure 6.

3.4. Cell Nuclei Quantification. In order to quantify the cell nuclei, the postprocessed image R is first converted to a greyscale image and then to a binary image using Otsu's thresholding method [26]. The conversion to grayscale allows the image to be represented in a bimodal fashion so as to obtain Otsu's threshold values. Each connected component of pixels representing a cell's nucleus in the resulting binary image is then counted to yield a total cell nuclei count within an image. The results obtained from the testing of the proposed algorithm are presented in the following section. Figure 7 illustrates the results of the preprocessing, entropy-based thresholding, and postprocessing steps on an input image.

4. Results

The results obtained by testing the proposed automated segmentation technique on a dataset of 21 immunohistochemically stained images are presented in Table 1.

The testing dataset consisted of 21 images belonging to a single patient that were stained using either Hematoxylin & Eosin (H&E) stain, cluster of differentiation 31 (CD-31), cluster of differentiation 68 (CD-68), or alpha-smooth muscle actin (α-SMActin). The cell nuclei from 21 test images were manually hand-counted by a pathologist, and the results that were obtained from the manual procedure were compared to the results generated by the automated segmentation technique. The qualitative and quantitative effectiveness of

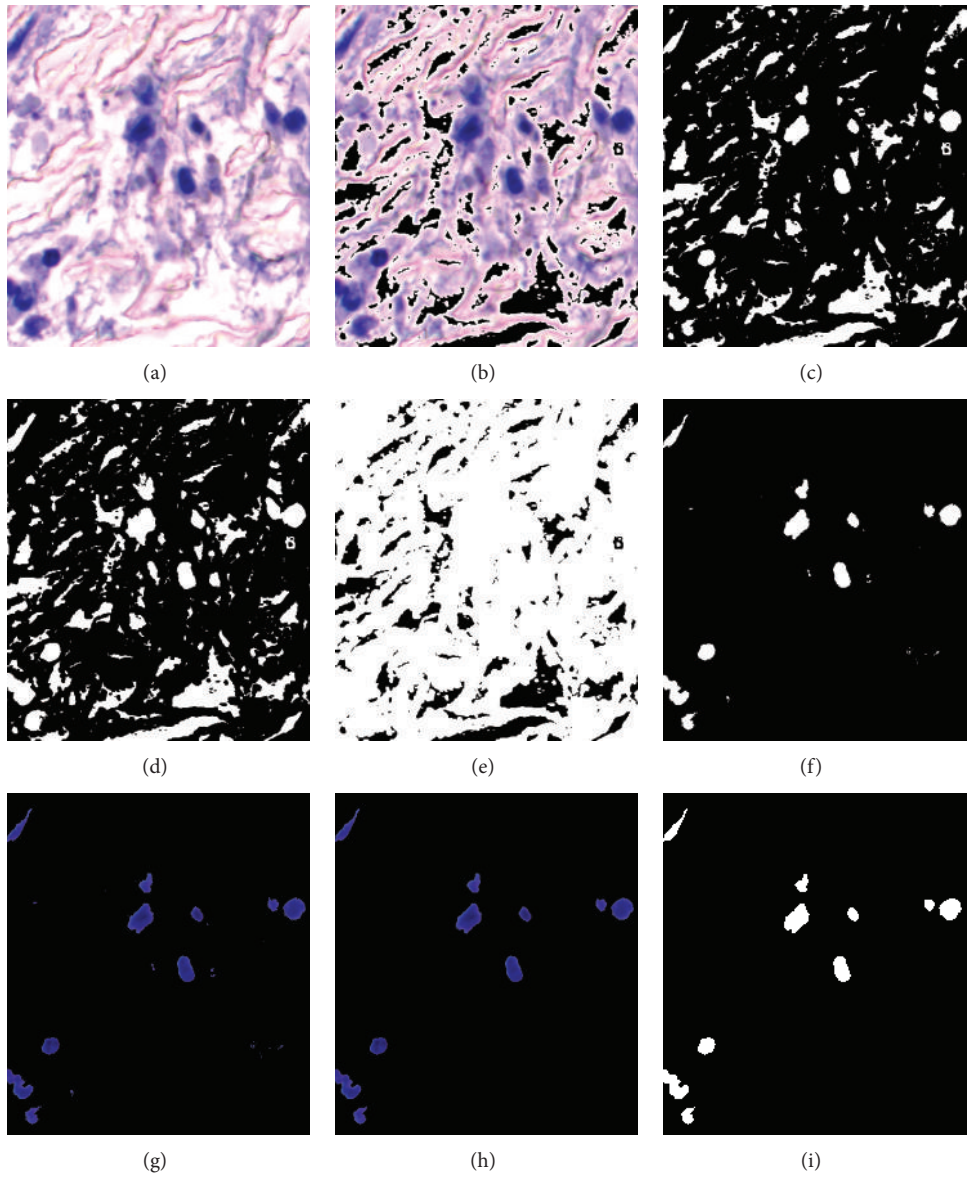


FIGURE 7: (a) Input image I. (b) Background less input image. (c) Red component thresholded at 90 to create image B_r. (d) Green component thresholded at 104 to create image B_g. (e) Blue component thresholded at 97 to create image B_b. (f) Binary image $BW = (B_r * B_g * B_b)$ of segmented cell nuclei. (g) Image R of cell nuclei with minimal noise. (h) Postprocessed image R with no noise. (i) Binary mask of output image R.

the proposed algorithm's performance is presented by means of its precision, accuracy, sensitivity, and specificity. Images acquired using 40x magnification usually contain more noise components, that is, small groups of connected pixels that do not represent cell nuclei, than the images that are acquired using a 60x magnification. This is due to the fact that the 40x images present a larger area of the tissue section in which the color of the stain is often expressed on small noncellular regions as well. The additional noise affects the precision of the segmentation technique, and this effect can be observed in Table 1 for some 40x images stained by CD-31 and CD-68. Therefore, it is ideal to use images captured at higher magnifications with the proposed algorithm, as they will obtain results with higher accuracy and precision. The true negatives in this study, that is, number of correctly identified noise components, were determined by the difference between the total number of cell nuclei quantified in an image before and after the postprocessing step. The overall accuracy and precision of the proposed segmentation algorithm based on the total number of cell nuclei identified are 95.55% and 91.27%, respectively. The high sensitivity, 96.45%, and the specificity value, 95.07%, achieved by the segmentation technique's performance suggests that the proposed method is effective at segmenting most cell nuclei while accurately identifying, distinguishing, and removing high volumes of noise from the segmented images.

Table 1: Results of automated segmentation performance on 21 test images.

Stain	ID/magnification	Manual quantification (cells)	Automated quantification (cells)	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)
SMActin	Day 5–40x	101	100	92.96	91	90.01	94.67
	Day 5–60x	59	60	91.82	88.33	89.83	93
	Day 7–40x	166	187	93.55	85.56	96.38	92.19
	Day 7–60x	87	94	93.96	89.24	95.40	93.10
	Day 14–40x	85	88	92.75	89.77	92.94	92.62
H&E	Day 5–40x	86	86	100	100	100	100
	Day 7–40x	141	138	99.05	100	97.87	100
	Day 7–60x	66	71	95.10	90.14	96.96	94.06
	Day 14–40x	105	108	98.69	97.22	100	97.6
	Day 14–60x	42	38	95.28	100	88.095	100
CD-31	Day 5–40x	76	77	99.38	98.70	100	98.85
	Day 5–60x	52	49	97.43	100	94.23	100
	Day 7–40x	140	188	94.32	77.77	97.22	93.66
	Day 7–60x	61	59	97.61	100	95.08	100
	Day 14–40x	98	104	97.97	94.23	100	97.97
	Day 14–60x	67	67	100	100	100	100
CD-68	Day 5–40x	99	96	91.36	76.19	96.96	89.39
	Day 5–60x	68	66	95.13	84.61	97.05	94.54
	Day 14–40x	143	138	98.55	100	96.50	100
	Day 14–60x	69	70	99.49	98.57	100	99.23
Total		1811	1884				
Average				95.55	91.27	96.45	95.07

In comparison to a related method for cell segmentation based on shape stability [20], the proposed method outperforms the quality of cell extraction and the precision of count for all cases with the given dataset. To ensure that the proposed system performs well on datasets that have been stained using different procedures, the system was also tested on additional immunohistochemically stained images of cancer cells hosted on the web by groups engaged in biomedical imaging research [27]. The segmentation results closely matched the accuracy and precision that were achieved in the results presented above.

5. Conclusion and Future Work

A novel translational computation system for automated cell nuclei segmentation and quantification has been proposed in this paper. Cell nuclei segmentation is a task that has several medical motivations ranging from the detection of malignant cell nuclei (tumor) in cancerous tissue images to the observation of cell nuclei for the characterization of the wound healing process within the human body. The proposed system uses an entropy-based thresholding technique to yield 3 optimal threshold values that are used to segment cell nuclei from images in the RGB color space. The entropy-based computations were based of the concepts introduced in [25]; however, the proposed algorithm introduces new methods such as the background removal preprocessing step, a noise removal postprocessing step, and a modified CCLSC histogram. The proposed technique is consistent in producing highly accurate and precise results of cell nuclei quantification; the technique overcomes the limitations of the existing time-consuming manual cell quantification methods and has great potential for use amongst pathologists. Future work will be directed towards improving the accuracy and precision of the proposed algorithm as well as towards the identification and classification of the various types of cell nuclei such as fibroblasts, and macrophages, which are segmented from the immunohistochemically stained images.

Acknowledgments

Virginia Commonwealth University Reanimation Engineering Science Center (VCURES) and the Biomedical Signal and Image Processing Lab of the Computer Science Department at Virginia Commonwealth University have supported the work presented in this paper.

References

[1] S. Y. Ji, R. Smith, T. Huynh, and K. Najarian, "A comparative analysis of multi-level computer-assisted decision making systems for traumatic injuries," *BMC Medical Informatics and Decision Making*, vol. 9, no. 2, 2009.

[2] E. Bak, K. Najarian, and J. P. Brockway, "Efficient segmentation framework of cell images in noise environments," in *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEMBCS '04)*, pp. 1802–1805, September 2004.

[3] S. Di Cataldo, E. Ficarra, A. Acquaviva, and E. Macii, "Automated segmentation of tissue images for computerized IHC analysis," *Computer Methods and Programs in Biomedicine*, vol. 100, no. 1, pp. 1–15, 2010.

[4] J. C. Sieren, J. Weydert, A. Bell et al., "An automated segmentation approach for highlighting the histological complexity of human lung cancer," *Annals of Biomedical Engineering*, vol. 38, no. 12, pp. 3581–3591, 2010.

[5] S. Wienert, D. Heim, K. Saeger et al., "Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach," *Scientific Reports*, vol. 2, p. 503, 2012.

[6] C. W. Wang, "Fast automatic quantitative cell replication with fluorescent live cell imaging," *BMC Bioinformatics*, vol. 13, p. 21, 2012.

[7] P. R. Gudla, K. Nandy, J. Collins, K. J. Meaburn, T. Misteli, and S. J. Lockett, "A high-throughput system for segmenting nuclei using multiscale techniques," *Cytometry A*, vol. 73, no. 5, pp. 451–466, 2008.

[8] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 841–852, 2010.

[9] T. Markiewicz, C. Jochymski, R. Koktysz, and W. Kozlowski, "Automatic cell recognition in immunohistochemical gastritis stains using sequential thresholding and SVM network," in *Proceedings of the 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI '08)*, pp. 971–974, May 2008.

[10] B. Ko, M. Seo, and J. Y. Nam, "Microscopic cell nuclei segmentation based on adaptive attention window," *Journal of Digital Imaging*, vol. 22, no. 3, pp. 259–274, 2009.

[11] V. R. Korde, H. Bartels, J. Barton, and J. Ranger-Moore, "Automatic segmentation of cell nuclei in bladder and skin tissue for karyometric analysis," *Analytical and Quantitative Cytology and Histology*, vol. 31, no. 2, pp. 83–89, 2009.

[12] X. Zhou, F. Li, J. Yan, and S. T. C. Wong, "A novel cell segmentation method and cell phase identification using Markov model," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 2, pp. 152–157, 2009.

[13] N. Malpica, C. O. de Solórzano, J. J. Vaquero et al., "Applying watershed algorithms to the segmentation of clustered nuclei," *Cytometry A*, vol. 28, no. 4, pp. 289–297, 1997.

[14] S. Singh, "Cancer cells detection and classification in biopsy image," *International Journal of Computer Applications*, vol. 38, no. 3, pp. 15–21, 2012.

[15] S. Singh, P. R. Gupta, and M. K. Sharma, "Breast cancer detection and classification of histopathological images," *International Journal of Engineering Science and Technology*, vol. 3, no. 5, p. 4228, 2010.

[16] H. F. Dvorak, "Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing," *New England Journal of Medicine*, vol. 315, no. 26, pp. 1650–1659, 1986.

[17] S. M. Alaish, D. A. Bettinger, O. O. Olutoye et al., "Comparison of the polyvinyl alcohol sponge and expanded polytetrafluoroethylene subcutaneous implants as models to evaluate wound healing potential in human beings," *Wound Repair and Regeneration*, vol. 3, no. 3, pp. 292–298, 1995.

[18] R. F. Diegelmann, W. J. Lindblad, and I. K. Cohen, "A subcutaneous implant for wound healing studies in humans," *Journal of Surgical Research*, vol. 40, no. 3, pp. 229–237, 1986.

[19] L. N. Jorgensen, L. Olsen, F. Kallehave et al., "The wound healing process in surgical patients evaluated by the expanded polytetrafluoroethylene and the polyvinyl alcohol sponge: a comparison with special reference to intrapatient variability," *Wound Repair and Regeneration*, vol. 3, no. 4, pp. 527–532, 1995.

[20] Z. Li and K. Najarian, "Biomedical image segmentation based on shape stability," in *Proceedings of the 14th IEEE International Conference on Image Processing (ICIP '07)*, vol. 6, pp. 281–284, September 2007.

[21] F. J. Cisneros, P. Cordero, A. Figueroa, and J. Castellanos, "Histology image segmentation," *International Journal of Information Technology and Management*, vol. 5, no. 1, pp. 67–76, 2011.

[22] N. Mirshahi, S. U. Demir, K. Ward, R. Hobson, R. Hakimzadeh, and K. Najarian, "An adaptive entropic thresholding technique for image processing and diagnostic analysis of microcirculation videos," *International Journal On Advances in Life Sciences*, vol. 2, no. 3-4, pp. 133–142, 2011.

[23] K. Najarian and R. Splinter, *Biomedical Signal and Image Processing*, CRC Press, Florida, Fla, USA, 2nd edition, 2012.

[24] C. Wilson, D. Brown, K. Najarian, E. N. Hanley, and H. E. Gruber, "Computer aided vertebral visualization and analysis: a methodology using the sand rat, a small animal model of disc degeneration," *BMC Musculoskeletal Disorders*, vol. 4, no. 1, p. 4, 2003.

[25] Y. Xiao, Z. Cao, and S. Zhong, "New entropic thresholding approach using gray-level spatial correlation histogram," *Optical Engineering*, vol. 49, no. 12, Article ID 127007, 2010.

[26] 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.

[27] Bio-Image Semantic Query User Environment (BISQUE), "Dataset of images," http://bisque.ece.ucsb.edu/client_service/browser?resource=/data_service/dataset.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 365410, 9 pages
http://dx.doi.org/10.1155/2013/365410

Research Article

The Effect of Edge Definition of Complex Networks on Protein Structure Identification

Jing Sun,[1] Runyu Jing,[1] Di Wu,[1] Tuanfei Zhu,[2] Menglong Li,[1] and Yizhou Li[1]

[1] *College of Chemistry, Sichuan University, Chengdu 610064, China*
[2] *College of Computer Science, Sichuan University, Chengdu 610064, China*

Correspondence should be addressed to Menglong Li; liml@scu.edu.cn and Yizhou Li; liyizhou_415@163.com

Received 6 December 2012; Revised 23 January 2013; Accepted 25 January 2013

Academic Editor: Guang Hu

The main objective of this study is to explore the contribution of complex network together with its different definitions of vertexes and edges to describe the structure of proteins. Protein folds into a specific conformation for its function depending on interactions between residues. Consequently, in many studies, a protein structure was treated as a complex system comprised of individual components residues, and edges were interactions between residues. What is the proper time for representing a protein structure as a network? To confirm the effect of different definitions of vertexes and edges in constructing the amino acid interaction networks, protein domains and the structural unit of proteins were described using this method. The identification performance of 2847 proteins with domain/domains proved that the structure of proteins was described well when R_{C_α} was around 5.0–7.5 Å, and the optimal cutoff value for constructing the protein structure networks was 5.0 Å (C_α-C_α distances) while the ideal community division method was community structure detection based on edge betweenness in this study.

1. Introduction

Protein structure comparison and classification are a difficult but important task since structure is a determinant for molecular interaction and function [1]. Protein folds into a specific conformation for its function depending on interactions between residues. Consequently, a protein structure can be treated as a complex system comprised of individual components residues. The method of complex network has been widely applied in various types of fields such as disease [2–4], drug target [5], drug design [6]. Network analysis facilitates the characterization of such complex system and its individual components [7, 8]. This provides novel insights into understanding the protein folding mechanism [9, 10], stability [11], function [9, 12, 13], and dynamics [14] and, more specifically, the study of protein structures. Viewing the protein structure as the an intricate network of interacting residues, metastructure analysis was proved to be an effective tool for large-scale (genome-wide) protein sequence analysis target selection for structural genomics and the identification of intrinsically unstructured (unfolded) proteins [15]. Analysis of the protein structure graphs showed that the aromatic residues along with arginine, histidine, and methionine act as strong hubs at high interaction cutoffs, which are found to play a role in bringing together different secondary structural elements in the tertiary structure of the proteins [11]. Through transforming the protein structure into residue interaction graphs, active site, ligand-binding, and evolutionary conserved residues were found to have high closeness values typically. This property will then be used to identify key protein residues [16]. Moreover, software tools were presented for the automatized generation, 2D visualization, and interactive analysis of residue interaction networks, which proved that residue networks are crucial for understanding structure-function relationships [17]. A novel web server, RING, was presented to construct physicochemically valid residue interaction networks interactively from PDB files for subsequent visualization in the Cytoscape platform [18]. The application of Cytoscape plug-ins, NetworkAnalyzer [19], and RINalyzer [17] were demonstrated for the standard and advanced analyses of network topologies [20].

In these studies, different strategies were used to define a vertex in literature: (a) only the C_α [9, 10, 15, 21–23] or C_β [21, 24] of an amino acid; (b) the center of the side

chain [11]; (c) all atoms in a residue were taken into account [16, 25]. Moreover, definition of edge also appears crucial in the construction of such networks. The characterization of protein structure is sensitive to the threshold for edges such as 5 Å (distances between two atoms from two amino acid residues) [25], 8 Å (C_α-C_α distances) [15], 8.5 Å (pairs of amino acids) [9], and a strict cutoff value of 7 Å [9, 10, 15, 21–23] based on the discovery that representing amino acids by C_α atoms may introduce bias for cutoffs below 6.8 Å [23].

Which strategy is more reasonable among all these choices? Studies have been made to find the answer. Three models were compared to prove the effects of the anisotropic nature of the side chain on the identification of the contact amino acid pairs [26]. The main objective of this study is to explore the contribution of complex network together with its different definitions of vertexes and edges to describing the structure of proteins. Automatic decomposition of protein structures into domains remains a challenging problem [27], and numbers of computer algorithms have been proposed [27–30]. Since domains can be considered as semi-independent structural units of a protein capable of folding independently [31, 32], consequently, the identification of protein domains is an efficient way to present whether a method can describe the protein structure well. In addition, the connections between the residues are dense within these structural units, which are similar to the connections between communities of the complex networks, expressing the community properties of such network well. To facilitate the understanding of such complex systems, community division was used to analyze these amino acid interaction networks. The purpose of this method is to divide the vertexes of the networks into groups, within which the connections between the vertexes are dense and the connections between which are sparser in the same time [33]. Moreover, a number of the methods based on community have been published in many fields [34–39].

In this study, protein structures were represented by complex networks, in which a vertex is a residue and an edge is an interaction between residues. Here, different cutoff values and strategies used for defining a vertex were tested. For a dataset of 2847 proteins with domain/domains, the identification performance in this study was assessed by accuracy (Acc), which was defined as the proportion of amino acids correctly identified in the certain domain regions of the query sequences according to the information of protein structures in SCOP [40]. For example, suppose the domain regions of the query sequence have 100 amino acids; if 90 of which were correctly identified as belonging to domain regions while the other 10 were misjudged as sequence regions, then the Acc will be 90%. It was observed that when the community division method was based on edge betweenness, the Acc (R_{C_α}) was stable at ~86% when R_{C_α} was around 5.0–7.5 Å, and Acc (R_{C_α}) achieved the highest value of 86.68% when R_{C_α} was 5.0 Å. In addition, when the community division method was based on random walks, the Acc (R_{C_α}) was ~81% when R_{C_α} was around 6.5–7.5 Å, and Acc (R_{C_α}) achieved the highest value of 81.87% when R_{C_α} was 7.0 Å and the step size was 10. The identification performance proved that the optimal cutoff value for constructing the protein structure networks was 5.0 Å (C_α-C_α distances), while the ideal community division method was community structure detection based on edge betweenness in this study. The results suggested that the amino acid interaction networks are an efficient method for describing the structure of proteins, and the different definitions of vertexes and edges do have important effect in this process.

Table 1: The composition of proteins contained in the dataset.

Number of domains	1	2	3	4	5	6	7
Number of proteins	1450	1077	230	66	19	3	2

2. Materials and Methods

2.1. Data Collection and Data Set Construction. The information on domains in proteins in this study were collected from ASTRAL SCOP [40] version 1.75 database. Protein domains in SCOP are grouped into species and hierarchically classified into families, superfamilies, folds, and classes [41]. This database organizes proteins hierarchically according to their families and folds, which is generally considered as the standard for protein structure classification [42]. In order to ensure the nonredundancy of the data, only these proteins with a pairwise sequence identity ≤30% were downloaded, and only those in which the structures were solved by X-ray crystallography with resolution ≤2.5 Å were kept for the clear structure of the proteins. Finally, the remaining 2847 proteins were left for this research. The compositions of the dataset were listed in Table 1.

2.2. Protein Structure Network. Protein structures can be represented as complex networks where amino acids are the nodes and their interactions are the edges [43]. In this study, each protein was considered a small self-governed network system. The structure of proteins was transformed into a complex network by taking amino acid residues as the vertexes and the interactions between the amino acid residues as edges. Various protein structure networks were constructed to investigate the protein structure and the influence of different strategies in building them.

Here, edges are defined in three ways, and from which the optimal cutoff value was finally chosen. Two amino acid residues have a connection if (a) the distance between C_α (defined as R_{C_α}) is 3–10 Å (step size of 0.5 Å, 15 different numerical values in all); (b) the distance between the centers of the side chains (defined as R_{cent}) is 3–10 Å (step size of 0.5 Å, 15 different numerical values in all); (c) the distance between any atoms of the amino acid residues (defined as R_{atm}) is 0–6 Å (step size of 0.5 Å, 13 different numerical values in all). The semidiameters of the atoms were taken into consideration. The amino acid residues interaction networks defined in this study are as shown in Figure 1, 3D structure of which is quite distinct.

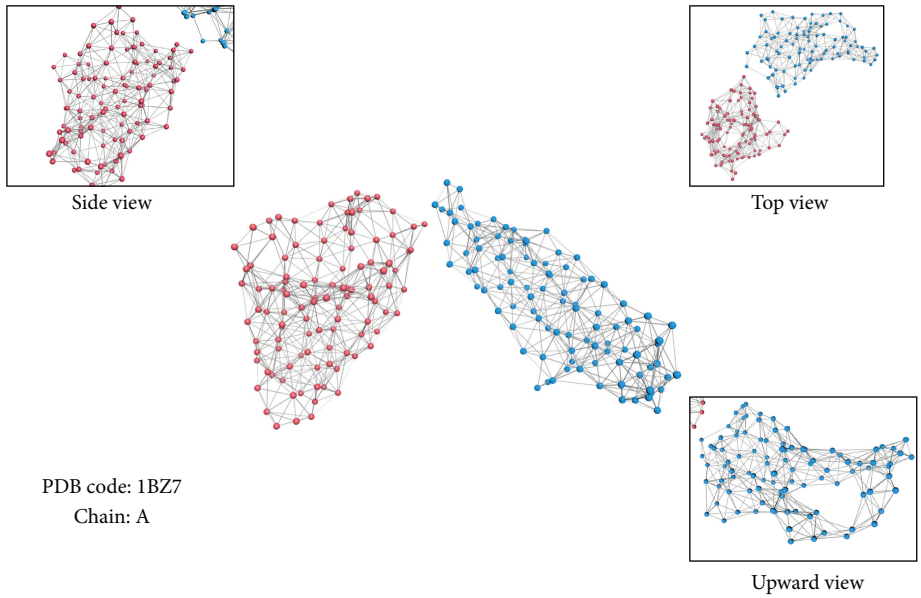


Figure 1: The amino acid residues interaction network. PDB code 1BZ7, chain A. The 3D structure of which is shown above together with its side, top and upward view. Here, the vertex is defined as C_α, and the edge is C_α-C_α distances which is set at 7.5 Å. Each point in the figure represents an amino acid in the chain, which is also the vertex of the network. Ligatures between the vertices are the edge of the network, which illustrate the interaction between the amino acids. For contrasting the figure of community division with this complex network, each vertex is colored based on its identity in SCOP. Here, reddish purple and blue represent different domain regions in this chain.

2.3. Community Division. Tools for network analysis are firmly grounded on the results in graph theory [44], including which network community structure plays an important role in organizing and understanding the complex networks. The network communities were identified as dense groups of the network, whose nodes have a much stronger influence on each other than on the rest of the network [35]. Moreover, the connections between the residues are dense within domains, which express the community properties of such network well. Based on this characteristic, in this study, the community division methods were used to divide the whole sequences into potential domain regions. Two different methods were employed here: community structure detection based on edge betweenness and community structure via short random walks, and between which the more ideal one was finally choosen.

2.3.1. Community Structure Detection Based on Edge Betweenness. Algorithms based on betweenness have been widely applied in various types of networks such as email messages, animal social networks, collaborations of jazz musicians, metabolic networks, and gene networks [33, 45–49]. For more detailed description of this method, refer to papers [45, 50]. The principle of the community structure detection based on edge betweenness is that it seems that all the shortest paths from one module to another must traverse through the edges connecting separate modules, which have high edge betweenness in that case.

As a result, this algorithm is performed by calculating the edge betweenness of the graph and removing the edge with the highest edge betweenness score gradually in order to obtain a hierarchical map. This rooted tree is the dendrogram of the graph, the leaves are the individual vertices, and the roots represent the whole graph. Finally, a numeric matrix is constructed using this algorithm.

2.3.2. Community Structure via Short Random Walks. Algorithms based on random walks have been applied in various researches of networks [50, 51]. This algorithm tries to find densely connected subgraphs which are also known as communities in a graph via short random walks. The principle of this algorithm is that short random walks are likely to stay in the same community. It takes every single node as an independent community at first, then those of which tally with certain rules were incorporated together step by step. It introduces r as a distance between the vertices, which shall be small if the two vertices are in the same community and large if they are not.

3. Results and Discussion

3.1. Community Division Based on Edge Betweenness. In this section, community division method based on edge betweenness was applied on complex networks, and the effect of different cutoff values of edges for constructing complex networks was analyzed. Then, an optimized cutoff value was identified. The flowchart of these two steps, amino acid interaction network together with community division methods, is shown in Figure 2.

For the fairness of the contrast, all complex networks constructed by different cutoff values were analyzed by community division method, which insures the most optimal

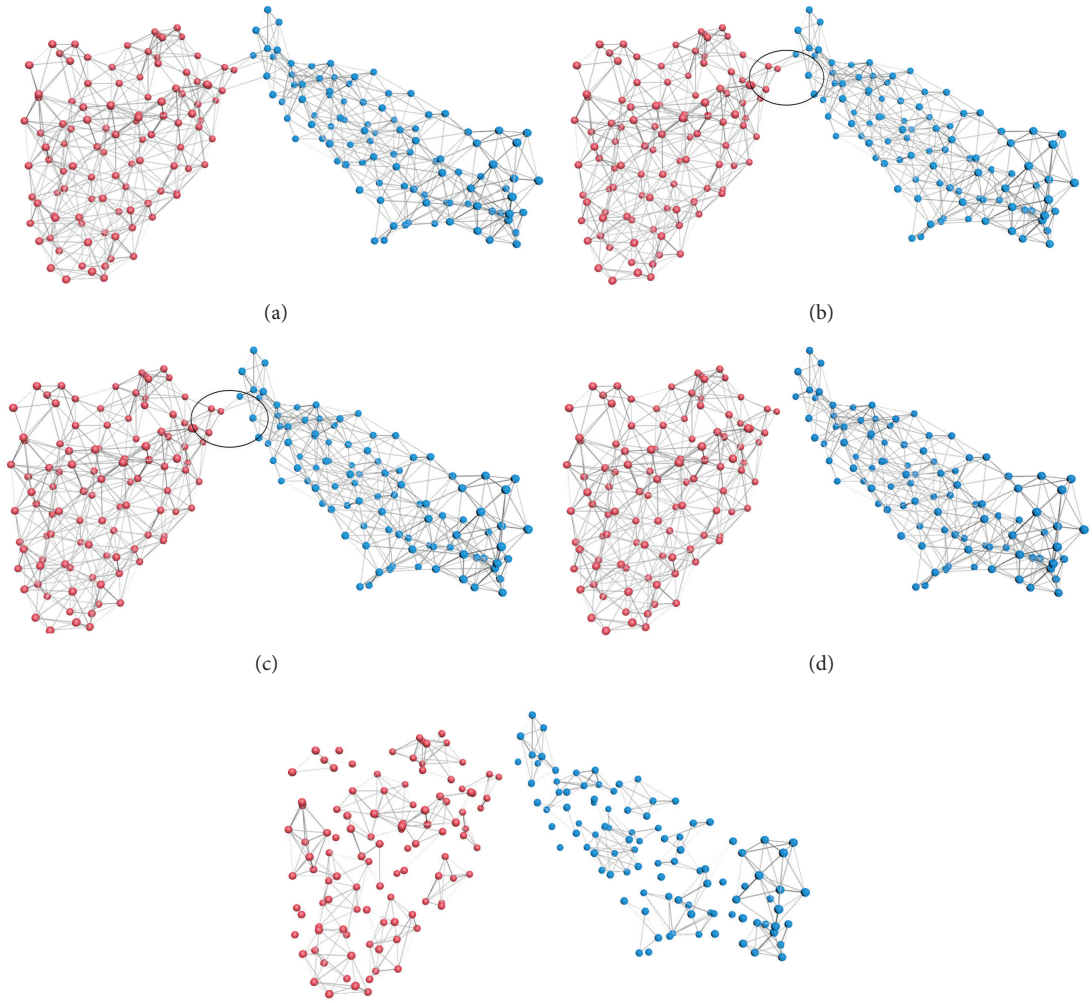


Figure 2: The flowchart of the amino acid interaction network together with community division method. PDB code 1BZ7, chain A. Each point in the figure represents an amino acid in the chain, which is also the vertex of the network. Ligatures between the vertices are the edge of the network, which illustrate the interaction between the amino acids. Here, the reddish purple and blue represent different domain regions in this chain based on the identity in SCOP. Firstly, an amino acid complex network was constructed with the vertex defined as C_α, and the edge as C_α-C_α distance which was set at 7.5 Å, as shown in (a). Secondly, community division was based on edge betweenness, and the first edge with the highest edge betweenness score was removed, as shown in (b). Thirdly, more edges were removed based on the algorithm, and (c) shows that three edges were removed. Fourthly, the community division was finished when the correct number of edges was removed, as shown in (d); two different domains have been clearly separated, and five edges were removed for this protein. Finally, if the community division is taken continually, more communities will be found in the complex network. (e) shows the result of community division for chain A of protein 1BZ7 after removing 500 edges in this complex network, and many more communities illustrate the wrong results according to the identity in SCOP.

results. In order to obtain the best prediction performance, different cutoff values were evaluated based on multidomain proteins. 15 different values (3–10 Å) of the R_{C_α} and the R_{cent} (step size of 0.5 Å) were optimized, respectively, and so were other 13 different distance values (0–6 Å) of R_{atom} (step size of 0.5 Å).

First, threshold of 7 Å, which has been reported to be an important distance parameter because all contacts are complete and legitimate (not occluded) at this distance [23], was analyzed. The results were obtained after the community division. The identification performance in this study was assessed by accuracy, which was defined as the proportion of amino acids correctly identified in the certain domain regions of the query sequences. When the R_{C_α} and the R_{cent} were 7 Å, respectively, the results are 86.21% and 85.16%, respectively.

More cutoff values were tested via different strategies of vertex. First, the average accuracies for all the proteins defined by R_{C_α} were listed in Table 2. The results indicated

Table 2: The accuracies of all proteins defined by R_{C_α} based on edge betweenness.

Threshold	Accuracy
3 Å	2.15
3.5 Å	2.17
4 Å	78.96
4.5 Å	83.42
5 Å	86.68
5.5 Å	86.45
6 Å	85.54
6.5 Å	85.76
7 Å	86.21
7.5 Å	85.92
8 Å	85.21
8.5 Å	84.75
9 Å	84.28
9.5 Å	83.71
10 Å	83.86

Table 3: The accuracies of all proteins defined by R_{cent} based on edge betweenness.

Threshold	Accuracy
3 Å	2.14
3.5 Å	2.59
4 Å	3.79
4.5 Å	7.42
5 Å	33.99
5.5 Å	78.87
6 Å	84.53
6.5 Å	85.04
7 Å	85.16
7.5 Å	85.52
8 Å	84.89
8.5 Å	84.48
9 Å	83.83
9.5 Å	83.56
10 Å	83.40

that when the method was based on the edge betweenness, Acc (R_{C_α}) achieved the highest 86.68% when R_{C_α} was 5.0 Å. When R_{C_α} was around 5.0–7.5 Å, the accuracies were around 86%, and the bias of the numerical values in this area was small (~1%). This illustrated that the cutoff values in this area reflected protein structure well. Second, the average accuracies for all the proteins defined by R_{cent} were listed in Table 3. The results indicated that Acc (R_{cent}) achieved the highest 85.52% when R_{cent} was 7.5 Å. When R_{cent} was around 6.5–8.0 Å, Acc (R_{cent}) showed relatively ideal values around 85%, which illustrated that the cutoff values in this area reflected protein structure well. However, the bias of the numerical values was evident for all the numerical values of R_{cent}. Acc (R_{cent}) were lower than 10% when R_{cent} was around 3.0–4.5 Å, which were generated by the otherness

Table 4: The accuracies of all proteins defined by R_{atom} based on edge betweenness.

Threshold	Accuracy
0 Å	85.06
0.5 Å	85.36
1.0 Å	85.58
1.5 Å	85.59
2 Å	85.06
2.5 Å	84.39
3 Å	83.73
3.5 Å	83.50
4 Å	83.95
4.5 Å	83.93
5 Å	83.51
5.5 Å	83.45
6 Å	83.31

of the size of side chains. Third, the average accuracies for all the proteins defined by R_{atom} were listed in Table 4. The results indicated that when the distance between any atoms of the amino acid residues defined as R_{atom} was taken into consideration, the superiority of the diversity of the volume of atoms should also be taken into consideration. Acc(R_{atom}) achieved the highest value of 85.59% when R_{atom} was 1.5Å. When R_{atom} was around 0.0–2.0 Å, Acc (R_{atom}) showed relatively ideal values around 85%, and the bias of the numerical values in this area was small (~0.6%). When the cutoff values were bigger than 2.0 Å, Acc (R_{atom}) decreased monotonically as R_{atom} increased. That is, overlarge R_{atom} will lead to the incorrect identification of the interactions among amino acids, which will distort the actual protein structure.

It was observed that when the community division method was based on edge betweenness, the Acc (R_{C_α}) was stable at ~86%, which illustrated that the network characterization of protein structure would not be limited by its type. Furthermore, Acc (R_{cent}) was ~1% lower than that of Acc (R_{C_α}), which was generated by the cutoff value. That is, the side chains of the amino acids have a certain space volume, and a big cutoff value signifies the space overlap of the atoms from different amino acids, which is obviously inappropriate for protein structure. In conclusion, Acc (R_{cent}) was lower than Acc (R_{C_α}) and Acc (R_{atom}), which illustrated that the space specificity of the side chains of amino acids affects the construction of the amino acids complex networks. It was observed that the highest accuracy obtained was 86.68% (R_{C_α} = 5.0 Å). That is, the optimal cutoff value was 5.0 Å (C_α-C_α distances) when the ideal community division method was based on edge betweenness.

3.2. Community Division Based on Random Walks. In this section, the community division method based on random walks was analyzed. The same cutoff values were evaluated here based on multidomain proteins, that is, 15 different numerical values (3–10 Å) of the R_{C_α} and the R_{cent} (step size of 0.5 Å) and other 13 different numerical values (0–6 Å) of

Table 5: Acc (R_{C_α}) and Acc (R_{cent}) of all proteins based on random walks under 7 Å of different step sizes.

Step size	3	4	5	6	7	8	9	10
Acc (R_{C_α})	77.37	78.56	79.84	80.21	80.93	81.23	81.43	81.93
Acc (R_{cent})	76.39	77.62	78.56	79.12	79.64	80.05	80.13	80.70

Table 6: The accuracies of all proteins defined by R_{C_α} based on random walks.

Threshold	Accuracy
3 Å	0
3.5 Å	0
4 Å	67.14
4.5 Å	69.65
5 Å	73.84
5.5 Å	79.87
6 Å	80.39
6.5 Å	81.09
7 Å	81.93
7.5 Å	81.85
8 Å	80.97
8.5 Å	80.48
9 Å	80.46
9.5 Å	79.95
10 Å	79.71

R_{atom} (using a step size of 0.5 Å). In addition, the step sizes of the community division based on random walks were also optimized here.

First, threshold of 7 Å [23] was analyzed for all the proteins. When the R_{C_α} and the R_{cent} were 7 Å, respectively, the results are listed in Table 5.

It was observed that when the community division method was based on random walks under the threshold of 7 Å via different step sizes, the highest Acc (R_{C_α}) and Acc (R_{cent}) were 81.93% and 80.70%, respectively. The numeric values of them all were ~4% lower than that for edge betweenness, which was generated by the method itself. That is, the algorithm based on the random walks attempted to find a given length called step size, which is obviously inappropriate for domains of different sizes. In large domains, a short length will not project all the amino acids in the same community.

More cutoff values were tested via different strategies of vertex. First, the average accuracies for all the proteins defined by R_{C_α} were listed in Table 6. The results indicated that Acc (R_{C_α}) achieved the highest 81.87% when R_{C_α} was 7.0 Å and the step size was 10. When R_{C_α} was around 6.5–7.5 Å, the accuracies were around 81%, and the bias of the numerical values in this area was small (~1%). This illustrated that the cutoff values in this area reflected protein structure well. However, the numeric of Acc (R_{C_α}) was ~5% lower than that for edge betweenness. Second, the average accuracies for all the proteins defined by R_{cent} were listed in Table 7. The results indicated that Acc (R_{cent}) achieved the highest value of

Table 7: The accuracies of all proteins defined by R_{cent} based on random walks.

Threshold	Accuracy
3 Å	0
3.5 Å	0
4 Å	0
4.5 Å	0
5 Å	0
5.5 Å	5.05
6 Å	59.20
6.5 Å	78.34
7 Å	80.63
7.5 Å	80.63
8 Å	80.77
8.5 Å	80.20
9 Å	79.60
9.5 Å	79.64
10 Å	79.41

Table 8: The accuracies of all proteins defined by R_{atom} based on random walks.

Threshold	Accuracy
0 Å	80.39
0.5 Å	80.58
1.0 Å	80.82
1.5 Å	80.70
2 Å	80.79
2.5 Å	80.08
3 Å	79.55
3.5 Å	79.35
4 Å	79.24
4.5 Å	78.98
5 Å	78.68
5.5 Å	78.36
6 Å	77.49

80.77% when R_{cent} was 8.0 Å and the step size was 10. When R_{cent} was around 7.0–8.5 Å, Acc (R_{cent}) showed relatively ideal values around 80%, which illustrated that the cutoff values in this area reflected protein structure well. However, the bias of the numerical values was evident for all the numerical values of R_{cent}, which were generated by the otherness of the side chains. The numeric of Acc (R_{cent}) was ~5% lower than that for edge betweenness, and Acc (R_{cent}) was as low as 0% when R_{C_α} was around 3.0–5 Å, which may be produced by the looseness of the complex networks constructed under these thresholds. Third, the average accuracifes for all the proteins defined by R_{atom} were listed in Table 8. The results indicated that when the distance between any atoms of the amino acid residues defined as R_{atom} was taken into consideration, the superiority of the diversity of the volume of atoms should also be taken into consideration. Acc (R_{atom}) achieved the highest value of 80.82% when R_{atom} was 1.0 Å and the step

Table 9: The optimal accuracies of each dataset based on edge betweenness.

Dataset	1	2	3	4	5	6	7	8
R_{C_α}	7.00 Å	5.50 Å	5.50 Å	5.00 Å	5.50 Å	5.00 Å	5.50 Å	5.50 Å
Accuracy	84.67	89.08	87.07	86.52	87.35	87.26	86.95	86.50
R_{cent}	6.50 Å	7.50 Å	7.50 Å	7.50 Å	7.50 Å	7.50 Å	7.50 Å	7.50 Å
Accuracy	82.51	86.93	86.50	85.74	86.17	86.58	85.85	85.49
R_{atom}	1.00 Å	1.00 Å	0.50 Å	1.00 Å	1.50 Å	1.00 Å	1.00 Å	1.00 Å
Accuracy	82.89	87.54	86.24	86.13	86.94	86.61	85.61	85.80

Table 10: The optimal accuracies of each dataset based on random walks.

Dataset	1	2	3	4	5	6	7	8
R_{C_α}	6.00 Å	7.50 Å	7.50 Å	7.50 Å	7.50 Å	7.00 Å	7.50 Å	7.00 Å
Step size	10	10	10	10	10	10	10	10
Accuracy	75.34	85.00	82.46	81.61	83.20	83.39	82.25	81.93
R_{cent}	7.00 Å	7.00 Å	8.00 Å	8.00 Å	8.00 Å	8.00 Å	7.50 Å	7.00 Å
Step size	10	10	10	10	9	10	10	10
Accuracy	74.62	84.95	80.97	80.89	81.84	82.67	80.61	80.79
R_{atom}	0.50 Å	1.50 Å	0.50 Å	1.00 Å	1.50 Å	1.00 Å	1.00 Å	1.00 Å
Step size	10	10	10	9	10	10	10	10
Accuracy	74.85	84.66	81.20	81.11	82.36	82.97	81.45	80.95

size was 10. When R_{atom} was around 0.0–2.5 Å, Acc (R_{atom}) showed relatively ideal values around 80%, and the bias of the numerical values in this area was small (~1%). However, the numeric of Acc (R_{atom}) was 5% lower than that for edge betweenness.

In conclusion, Acc (R_{cent}) was lower than Acc (R_{C_α}) and Acc (R_{atom}). It was observed that when the community division method was based on random walks, the numeric of the accuracy was lower than that based on edge betweenness all the while, which indicated that the ideal community division method for this research was community structure detection based on edge betweenness. Moreover, the value of Acc (R_{cent}) was the worst via both the two community division methods all along. Similar results were obtained in the study of side chain contact models; three models were compared and the isotropic sphere side chain (ISS) model was the worst in accuracy. They proved that the model which took the spatially anisotropic nature of the side chain into consideration would eliminate about 95% of the incorrectly counted contact pairs in the ISS model [26]. However, this kind of practical models do have less moderate computational cost than the popular representation model such as the use of C_α atom, which is proved to be effective for the kind of the data in this study.

3.3. The Stability Analysis of the Method. To verify the stability of the method, 8 datasets were constructed based on multidomain proteins. The first dataset was composed of 100 proteins, and every other dataset contained 100 proteins more than the previous one. That is, the 8th dataset contained 800 proteins.

The same operations were taken based on these 8 datasets. Different numerical values of R_{C_α} (3–10 Å), R_{cent} (3–10 Å), and R_{atom} (0–6 Å) were optimized based on two community division methods. The highest accuracies for each dataset were listed in Tables 9 and 10.

It was observed that when the community division method was based on edge betweenness, Acc (R_{C_α}) for each database got the highest results around ~86%–89% when R_{C_α} was ~5.00–5.50 Å, which were quite close to the result 86.68% when R_{C_α} was 5.00 Å. However, results for database one was a little bit different, 84.67% when R_{C_α} was 7.00 Å, which may be generated by the lack of statistically significant result in the small amount of the proteins. Acc (R_{cent}) for each database got the highest results around ~85%-86% when R_{cent} was 7.50 Å, which were quite close to the result 85.52% when R_{cent} was 7.50 Å. However, results for database one was a little bit different, 82.51% when R_{cent} was 6.50 Å, which may be generated by the lack of statistically significant result in the small amount of the proteins. Acc (R_{atom}) for each database got the highest results around ~82%–87% when R_{atom} was ~0.50–1.50 Å, which were quite close to the result 85.59% when R_{C_α} was 1.50 Å.

When the community division method was based on random walks, Acc (R_{C_α}) for each database got the highest results around ~81%–85% when R_{C_α} was ~7.00–7.50 Å and the step size was 10, which were quite close to the result 81.87% when R_{C_α} was 7.0 Å and the step size was 10. Acc (R_{cent}) for each database got the highest results around ~80%–84% when R_{cent} was 7.00–8.00 Å, which were quite close to the result 80.77% when R_{cent} was 8.0 Å and the step size was 10. Acc (R_{atom}) for each database got the highest results around ~80%–84% when R_{atom} was ~0.50–1.50 Å and the step size was 10, which were quite close to the result 80.82% when R_{C_α} was 1.00 Å and the step size was 10. However, results for database one was a little bit different under these three

conditions, which may be generated by the lack of statistically significant result in the small amount of the proteins.

It is observed from the results that the complex networks together with the community division methods constructed in this study were stable, which proved the creditability of the research. On the other hand, it was observed that when the community division method was based on edge betweenness, the Acc (R_{C_α}) was stable at ~86% when R_{C_α} was around 5.0–7.5 Å, and the optimal cutoff value for constructing the protein structure networks was 5.0 Å (C_α-C_α distances) in this study.

4. Conclusion

The main objective of this study is to explore the contribution of complex network together with its different definitions of vertexes and edges to describing the structure of proteins. When applying our method on a dataset of 2847 proteins with domain/domains, it was observed that when the community division method was based on random walks, the numeric of the accuracy was lower than that based on edge betweenness all the while, which indicated that the ideal community division method for this research was community structure detection based on edge betweenness. When the community division method was based on edge betweenness, the Acc (R_{C_α}) was stable at ~86% when R_{C_α} was around 5.0–7.5 Å, and Acc (R_{C_α}) achieved the highest value of 86.68% when R_{C_α} was 5.0 Å. The identification performance proved that the optimal cutoff value for constructing the protein structure networks was 5.0 Å (C_α-C_α distances), while the ideal community division method was community structure detection based on edge betweenness in this study. The results suggested that the amino acid interaction networks are an efficient method for describing the structure of proteins, and the different definitions of vertexes and edges do have important effect in this process. Distance should be taken into consideration to prevent unnecessary deviation. Moreover, the optimized network model could be further applied in future study for the number and position of protein domain prediction.

Acknowledgments

The authors would like to thank the anonymous reviewers for their patient review and constructive suggestions. This study was supported by the Natural Science Foundation of China (21175095, 20972103).

References

[1] R. C. Penner, M. Knudsen, C. Wiuf, and J. E. Andersen, "An algebro-topological description of protein domain structure," *PLoS ONE*, vol. 6, no. 5, article e19670, Article ID e19670, 2011.

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

[3] O. Magger, Y. Y. Waldman, E. Ruppin, and R. Sharan :, "Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks," *PLoS Computational Biology*, vol. 8, no. 9, article e1002690, 2012.

[4] S. Karni, H. Soreq, and R. Sharan, "A network-based method for predicting disease-causing genes," *Journal of Computational Biology*, vol. 16, no. 2, pp. 181–189, 2009.

[5] F. Cheng, C. Liu, J. Jiang et al., "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Computational Biology*, vol. 8, no. 5, article e1002503, 2012.

[6] P. Csermely, V. Ágoston, and S. Pongor, "The efficiency of multi-target drugs: the network approach might help drug design," *Trends in Pharmacological Sciences*, vol. 26, no. 4, pp. 178–182, 2005.

[7] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001.

[8] R. Albert and A. L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

[9] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus, "Small-world view of the amino acids that play a key role in protein folding," *Physical Review E*, vol. 65, no. 6, article 061910, Article ID 061910, 4 pages, 2002.

[10] N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich, "Topological determinants of protein folding," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 13, pp. 8637–8641, 2002.

[11] K. V. Brinda and S. Vishveshwara, "A network representation of protein structures: implications for protein stability," *Biophysical Journal*, vol. 89, no. 6, pp. 4159–4170, 2005.

[12] B. Thibert, D. E. Bredesen, and G. del Rio, "Improved prediction of critical residues for protein function based on network and phylogenetic analyses," *BMC Bioinformatics*, vol. 6, article 213, 2005.

[13] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular systems biology*, vol. 3, article 88, 2007.

[14] C. Böde, I. A. Kovács, M. S. Szalay, R. Palotai, T. Korcsmáros, and P. Csermely, "Network analysis of protein dynamics," *FEBS Letters*, vol. 581, no. 15, pp. 2776–2782, 2007.

[15] R. Konrat, "The protein meta-structure: a novel concept for chemical and molecular biology," *Cellular and Molecular Life Sciences*, vol. 66, no. 22, pp. 3625–3639, 2009.

[16] G. Amitai, A. Shemesh, E. Sitbon et al., "Network analysis of protein structures identifies functional residues," *Journal of Molecular Biology*, vol. 344, no. 4, pp. 1135–1146, 2004.

[17] N. T. Doncheva, K. Klein, F. S. Domingues, and M. Albrecht, "Analyzing and visualizing residue networks of protein structures," *Trends in Biochemical Sciences*, vol. 36, no. 4, pp. 179–182, 2011.

[18] A. J. M. Martin, M. Vidotto, F. Boscariol, T. Di Domenico, I. Walsh, and S. C. E. Tosatto, "RING: networking interacting residues, evolutionary information and energetics in protein structures," *Bioinformatics*, vol. 27, no. 14, pp. 2003–2005, 2011.

[19] Y. Assenov, F. Ramírez, S. E. S. E. Schelhorn, T. Lengauer, and M. Albrecht, "Computing topological parameters of biological networks," *Bioinformatics*, vol. 24, no. 2, pp. 282–284, 2008.

[20] N. T. Doncheva, Y. Assenov, F. S. Domingues, and M. Albrecht, "Topological analysis and interactive visualization of biological networks and protein structures," *Nature Protocols*, vol. 7, no. 4, pp. 670–685, 2012.

[21] A. R. Atilgan, P. Akan, and C. Baysal, "Small-world communication of residues and significance for protein dynamics," *Biophysical Journal*, vol. 86, no. 1, pp. 85–91, 2004.

[22] G. Bagler and S. Sinha, "Network properties of protein structures," *Physica A*, vol. 346, no. 1-2, pp. 27–33, 2005.

[23] C. H. Da Silveira, D. E. V. Pires, R. C. Minardi et al., "Protein cutoff scanning: a comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins," *Proteins*, vol. 74, no. 3, pp. 727–743, 2009.

[24] E. Estrada, "Universality in protein residue networks," *Biophysical Journal*, vol. 98, no. 5, pp. 890–900, 2010.

[25] L. H. Greene and V. A. Higman, "Uncovering network systems within protein structures," *Journal of Molecular Biology*, vol. 334, no. 4, pp. 781–791, 2003.

[26] W. Sun and J. He, "From isotropic to anisotropic side chain representations: comparison of three models for residue contact estimation," *PLoS ONE*, vol. 6, no. 4, Article ID e19238, 2011.

[27] J. T. Guo, D. Xu, D. Kim, and Y. Xu, "Improving the performance of DomainParser for structural domain partition using neural network," *Nucleic Acids Research*, vol. 31, no. 3, pp. 944–952, 2003.

[28] Y. Xu, D. Xu, and H. N. Gabow, "Protein domain decomposition using a graph-theoretic approach," *Bioinformatics*, vol. 16, no. 12, pp. 1091–1104, 2000.

[29] J. E. Gewehr and R. Zimmer, "SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles," *Bioinformatics*, vol. 22, no. 2, pp. 181–187, 2006.

[30] J. Cheng, M. J. Sweredoski, and P. Baldi, "DOMpro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks," *Data Mining and Knowledge Discovery*, vol. 13, no. 1, pp. 1–10, 2006.

[31] J. S. Richardson, "The anatomy and taxonomy of protein structure," *Advances in Protein Chemistry*, vol. 34, pp. 167–339, 1981.

[32] D. B. Wetlaufer, "Nucleation, rapid folding, and globular intrachain regions in proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 70, no. 3, pp. 697–701, 1973.

[33] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[34] M. Szalay-Beko, R. Palotai, B. Szappanos, I. A. Kovacs, B. Papp, and P. Csermely :, "ModuLand plug-in for Cytoscape: determination of hierarchical layers of overlapping network modules and community centrality," *Bioinformatics*, vol. 28, no. 16, pp. 2202–2204, 2012.

[35] I. A. Kovács, R. Palotai, M. S. Szalay, and P. Csermely, "Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics," *PLoS ONE*, vol. 5, no. 9, article e12528, Article ID e12528, pp. 1–14, 2010.

[36] Y. Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[37] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, 2010.

[38] A. Delmotte, E. W. Tate, S. N. Yaliraki, and M. Barahona, "Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin-myosin light chain interaction," *Physical Biology*, vol. 8, no. 5, article 055010, 2011.

[39] J. C. Delvenne, S. N. Yaliraki, and M. Barahon, "Stability of graph communities across time scales," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 29, pp. 12755–12760, 2010.

[40] S. E. Brenner, P. Koehl, and M. Levitt, "The ASTRAL compendium for protein structure and sequence analysis," *Nucleic Acids Research*, vol. 28, no. 1, pp. 254–256, 2000.

[41] L. Lo Conte, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin, "SCOP database in 2002: refinements accomodate structural genomics," *Nucleic Acids Research*, vol. 30, no. 1, pp. 264–267, 2002.

[42] R. Day, D. A. C. Beck, R. S. Armen, and V. Daggett, "A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary," *Protein Science*, vol. 12, no. 10, pp. 2150–2160, 2003.

[43] S. Lifson and C. Sander, "Antiparallel and parallel β-Strands differ in amino acid residue preferences," *Nature*, vol. 282, no. 5734, pp. 109–111, 1979.

[44] L. D. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, "Characterization of complex networks: a survey of measurements," *Advances in Physics*, vol. 56, no. 1, pp. 167–242, 2007.

[45] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, Article ID 026113, p. 1, 2004.

[46] D. M. Wilkinson and B. A. Huberman, "A method for finding communities of related genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 1, pp. 5241–5248, 2004.

[47] P. Holme, M. Huss, and H. Jeong, "Subnetwork hierarchies of biochemical pathways," *Bioinformatics*, vol. 19, no. 4, pp. 532–538, 2003.

[48] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, "E-Mail as spectroscopy: automated discovery of community structure within organizations," *Information Society*, vol. 21, no. 2, pp. 133–153, 2005.

[49] P. M. Gleiser and L. Danon, "Community structure in jazz," *Advances in Complex Systems*, vol. 6, no. 4, pp. 565–573, 2003.

[50] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Proceedings of the Computer and Information Sicences (ISCIS '05)*, vol. 3733, pp. 284–293, 2005.

[51] H. J. Zhou and R. Lipowsky, "Network brownian motion: a new method to measure vertex-vertex proximity and to identify communities and subcommunities," in *Proceedings of the Computational Science (ICCS '04)*, vol. 3038, Part 3, pp. 1062–1069, 2004.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 645702, 8 pages
http://dx.doi.org/10.1155/2013/645702

Research Article

Novel Application of a Multiscale Entropy Index as a Sensitive Tool for Detecting Subtle Vascular Abnormalities in the Aged and Diabetic

Hsien-Tsai Wu,[1] Men-Tzung Lo,[2] Guan-Hong Chen,[1] Cheuk-Kwan Sun,[3] and Jian-Jung Chen[4,5]

[1] *Department of Electrical Engineering, National Dong Hwa University, Hualien, No. 1, Section 2, Da-Hsueh Road, Shoufeng, Hualien 97401, Taiwan*
[2] *Research Center for Adaptive Data Analysis and Center for Dynamical Biomarkers and Translational Medicine, National Central University, Chungli, Taiwan*
[3] *Department of Emergency Medicine, E-Da Hospital, I-Shou University, Kaohsiung City, Taiwan*
[4] *Department of Chinese Medicine, Buddhist Tzu-Chi General Hospital Taichung Branch, Taichung, Taiwan*
[5] *School of Chinese Medicine, Tzu Chi University, Hualien, Taiwan*

Correspondence should be addressed to Hsien-Tsai Wu; dsphans@mail.ndhu.edu.tw

Received 5 November 2012; Accepted 28 December 2012

Academic Editor: Guang Hu

Although previous studies have shown the successful use of pressure-induced reactive hyperemia as a tool for the assessment of endothelial function, its sensitivity remains questionable. This study aims to investigate the feasibility and sensitivity of a novel multiscale entropy index (MEI) in detecting subtle vascular abnormalities in healthy and diabetic subjects. Basic anthropometric and hemodynamic parameters, serum lipid profiles, and glycosylated hemoglobin levels were recorded. Arterial pulse wave signals were acquired from the wrist with an air pressure sensing system (APSS), followed by MEI and dilatation index (DI) analyses. MEI succeeded in detecting significant differences among the four groups of subjects: healthy young individuals, healthy middle-aged or elderly individuals, well-controlled diabetic individuals, and poorly controlled diabetic individuals. A reduction in multiscale entropy reflected age- and diabetes-related vascular changes and may serve as a more sensitive indicator of subtle vascular abnormalities compared with DI in the setting of diabetes.

1. Introduction

Endothelial dysfunction (ED) has been documented as a sign of the imminent onset of cardiovascular disease (CVD) including atherosclerosis and CVD-related disorders (i.e., diabetes, hypertension) [1–3]. The commonly used noninvasive means of assessing ED include flow-mediated dilatation (FMD) [4–6] and reactive hyperemia peripheral arterial tonometry (RH-PAT) [7–9]. The principle underlying the measurement is the induction of transient ischemia through increased cuff pressure over the upper arm, followed by a release of pressure. The reperfusion thus produced elicits reactive hyperemia (RH) in the distal blood vessels through the release of nitric oxide (NO), which is an indicator of endothelial integrity [10–12].

Although FMD provides direct information about the changes in blood vessel diameter, it requires an experienced operator and expensive equipment. On the other hand, RH-PAT acquires arterial pulse signals of the index finger through tonometry and compares them before and after RH induction. The popularity of its clinical use, however, is also hampered by the need for experienced personnel and its costly disposable accessories. As a result, no well-designed study has investigated ED in elderly and diabetic subjects who are at high risk of CVD. The importance of early detection of ED in diabetic patients is further underscored by the finding

that ED occurs within 10 years of full-fledged diabetes. Early detection of ED and timely intervention, therefore, are of utmost importance in the prevention of diabetes and its associated complications [13–15].

This study was designed to test the sensitivity and validity of applying a novel multiscale entropy index (MEI) in evaluating the degree of ED in subjects at risk of CVD. This was performed by analyzing the dynamical complexity of arterial pulse waveform signals, obtained through the wrist before and after induction of RH from 4 different subject populations using multiscale entropy analysis of biological signals [16–19].

2. Materials and Methods

2.1. Study Population and Grouping. A total of 70 subjects were recruited from the diabetes outpatient clinic of Hualin Hospital between December 2009 and October 2010. In addition, there were 70 healthy controls recruited from a health examination program at the same hospital. The 140 study subjects were categorized into the following 4 groups: group 1, which included healthy young individuals aged 20–30 years, with no known history of CVD, glycosylated hemoglobin (HbA1c) levels of less than 6%, and fasting blood sugar levels of less than 126 mg/dL; group 2, which included healthy middle-aged or elderly individuals aged 40–70 years, with no known history of CVD, HbA1c levels of less than 6%, fasting blood sugar levels of less than 126 mg/dL, and absence of metabolic syndrome according to the ATP III report [20]; group 3, which included well-controlled diabetic individuals aged 50–80 years, with an established diagnosis of type 2 diabetes (i.e., HbA1c levels > 6.5% and fasting sugar levels > 126 mg/dL) [20], HbA1c levels between 6.5% and 8% and fasting blood sugar levels of more than 126 mg/dL at the time of the present study; group 4, which included poorly controlled diabetic individuals aged 50–80 years, who fit the criteria of diabetes with HbA1c levels of more than 8% and fasting sugar levels of more than 126 mg/dL.

2.2. Experimental Procedure. Before initiating the study, subjects were required to fill out a questionnaire on basic demographic and anthropometric data as well as information on lifestyle and personal/family history of CVD. Physicians also obtained blood samples after 8 hours of fasting for determination of serum high-density lipoprotein (HDL), low-density lipoprotein (LDL), triglyceride (TG), fasting blood sugar, and HbA1c levels. Informed consent was obtained from all subjects.

The study subjects were allowed to assume a supine position and rest in a quiet, temperature-controlled (25°C) room for 5 minutes before measurement. Blood pressure was obtained once over the left arm of the supine patients using an automated oscillometric device (BP3AG1, Microlife, Taiwan) with a cuff of appropriate size. One pressure cuff of the air pressure sensing system (APSS) was then put around the left arm, whereas the other cuff was applied on the left wrist [21, 22]. The pressure of the cuff around the wrist was maintained at 40 mmHg throughout the process of measurement, which took 17 minutes for each subject.

2.3. Dilatation Index (DI) Computation. The structure and principles of operation of the APSS have been previously reported [21]. In brief, the APSS system consists of two sets of pressure cuffs, a piezoresistive sensor, and an endothelial function measurement module board. The first set of pressure cuffs is placed over the upper arm and triggers the endothelial function, whereas the second set is placed over the wrist for data acquisition. The piezoresistive sensor, which is connected to the second set of pressure cuffs, is used to detect the pulse wave and record the arterial waveform in the system. The endothelial function measurement module board amplifies and filters the captured arterial waveform. The pressure detected by the piezoresistive sensor was thus converted into electrical signals which were then amplified and filtered to obtain the analog signals. The analog signals were digitized with an analog-to-digital converter (Model MSP430F449, Texas Instruments, TX, USA) at a sampling rate of 500 Hz and stored in a computer for later analysis [22]. The total duration of signal acquisition was 17 minutes (Figure 1), which consisted of 5 minutes of data recording at a wrist cuff pressure of 40 mmHg with the arm cuff deflated (i.e., the baseline), 3 minutes of blood flow occlusion by increasing the cuff pressure of the upper arm to 200 mmHg (i.e., the occlusion phase), and 9 minutes of data acquisition after complete deflation of the pressure cuff over the upper arm with the pressure of the wrist cuff being maintained at 40 mmHg throughout (i.e., the hyperemic phase). The amplitude of the signals during the hyperemic phase varied with the subject's age and disease status (see Figure 1). The mean amplitude of signals within a representative one-minute period between the fifth and tenth minute after the beginning of data collection was selected from the baseline and hyperemic phases, respectively, and labeled as $\mathrm{Amp}_{\mathrm{Baseline}}$ and $\mathrm{Amp}_{\mathrm{RH}}$ (see Figure 1). The dilatation index (DI) [8, 21] of the forearm blood vessel is defined as

$$\mathrm{DI}_{\mathrm{Amp}} = \frac{\mathrm{Amp}_{\mathrm{RH}}}{\mathrm{Amp}_{\mathrm{Baseline}}} \times 100\%. \quad (1)$$

In agreement with the results of previous studies [8, 9, 22], our finding (see Figure 1) showed that the value of DI decreases with advancing age and increasing severity of diabetes. In contrast with the calculation of DI, which adopted 1 minute of signals from both the baseline and reactive hyperemic phases, the present study attempted to utilize the entire 14 minutes of signals (except for the occlusion phase) in the calculation of the multiscale entropy index (MEI). This was performed to provide a sensitive tool for detecting subtle vascular abnormalities in the elderly and diabetic patients.

2.4. Multiscale Entropy Index (MEI). After deleting the 3 minutes of arterial pulse signal acquired during the occlusion phase, signals of the baseline and hyperemic phases were connected for analysis. The footpoint of each waveform was first marked, followed by the identification of the peak between two footpoints [23]. The amplitude of the waveform

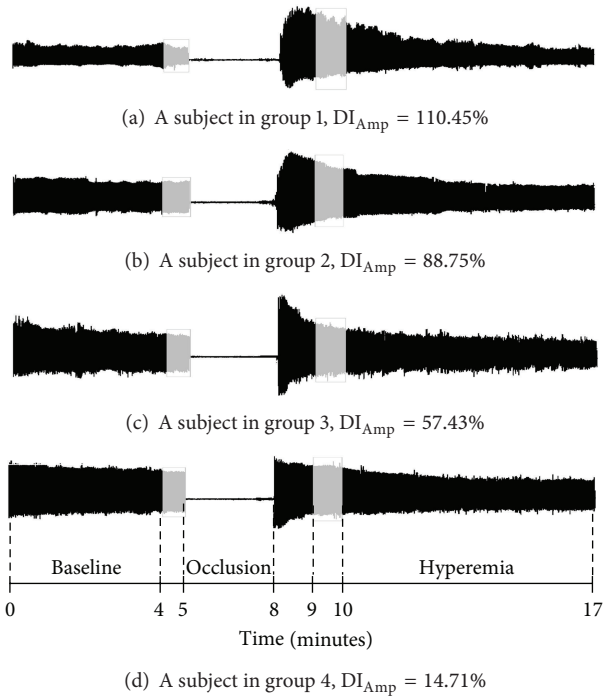


Figure 1: Representative arterial pulse signals from the 4 different groups, showing variations in the dilatation index (DI_{Amp}). Group 1: healthy young individuals; group 2: healthy middle-aged or elderly individuals; group 3: well-controlled diabetic individuals; group 4: poorly controlled diabetic individuals.

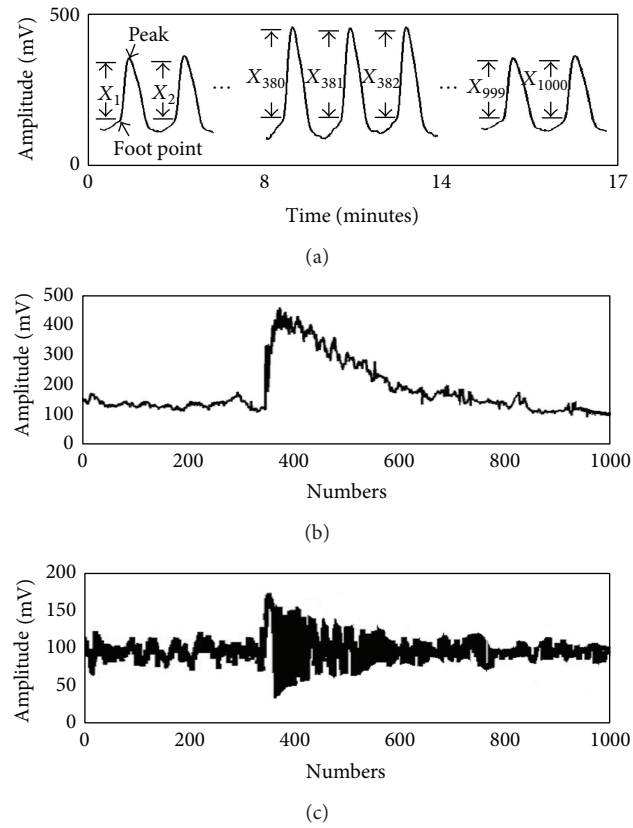


Figure 2: (a) Identification of the footpoint and peak of each arterial waveform measured from the wrist of a healthy young subject (group 1) using the air pressure sensing system (APSS), after connecting the baseline signals (5 min) to those at the hyperemic phase (9 min). (b) Plotting of the amplitudes from 1000 waveforms against time, giving a nonstationary curve. (c) Final curve after detrending using Empirical Mode Decomposition (EMD).

(X_i, $i = 1, 2, \ldots, 1000$) was defined as the vertical distance between the peak and the nearest footpoint. The amplitudes $X_1, X_2, \ldots, X_{379}$ were defined as the baseline values of the arterial pulses, whereas the amplitudes $X_{380}, X_{381}, \ldots, X_{1000}$ were defined as the hyperemic phase values of the arterial pulses (see Figure 2(a)). The values of the amplitudes thus obtained were plotted versus time (see Figure 2(b)). Since the nonstationary nature of the curve would affect the accuracy of the multiscale entropy (MSE) calculation, the curve was detrended with empirical mode decomposition (EMD), as proposed by a previous study [24–26] (see Figure 2(c)). This process yielded 1,000 amplitude points $\{X'_1, X'_2, \ldots, X'_{1000}\}$ for MSE analysis.

2.4.1. Multiscale Entropy (MSE) Computation. MSE was calculated in accordance with the procedure reported by Costa et al. [16]. Given a 1-dimensional discrete time series, $\{X'_1, X'_2, \ldots, X'_{1000}\}$, consecutive coarse-grained time series $\{y^{(\tau)}\}$, determined by the scale factor τ, can be constructed according to the equation

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} X'_i, \quad 1 \le j \le \frac{1000}{\tau}, \tag{2}$$

where τ denotes the scale factor and $1 \leqq j \leqq 1000/\tau$.In other words, coarse-grained time series for scale factor τ were acquired by taking the arithmetic average of τ neighboring original values without overlapping. The length of each coarse-grained time series is $1000/\tau$. For scale 1, the coarse-grained time series is just the original time series. Sample entropy (S_E) [27] for each of the coarse-grained time series can be obtained and plotted against the scale factor, τ.

2.4.2. Multiscale Entropy Index (MEI) Computation. The values of S_E were then obtained from a range of scale factors between 1 and 10 using the MSE data analysis method described above. The values of S_E between scale factors 1 and 5 were defined as small scale, whereas those between scale factors 6 and 10 were large scale [28, 29]. The sum of S_E values between scale factors 1 and 5 was defined as MEI_{SS}, while the sum of S_E values between scale factors 6 and 10 was defined as MEI_{LS} [28, 29], see (3) below. By defining and calculating these two indices of multiscale entropy, the complexity of signals between different time scales can be assessed and quantified. Using these 2 indices, the present study attempted to evaluate the differences in signal complexity of the hyperemic responses elicited by

temporary ischemia, an index of endothelial function, in different subject populations,

$$\mathrm{MEI_{SS}} = \sum_{\tau=1}^{5} S_E(\tau),$$
$$\mathrm{MEI_{LS}} = \sum_{\tau=6}^{10} S_E(\tau), \quad (3)$$

where $S_E(\tau)$ is the sample entropy for the respective scale factor.

2.5. Statistical Analysis. Average values were expressed as mean ±SD. Statistical Package for the Social Science (SPSS, version 14.0) was adopted. Independent sample t-test and Pearson's correlation were used for the determination of significance between different groups and assessment of correlations among different parameters, respectively. A P value < 0.05 was considered statistically significant.

3. Results

By recording serial changes in 1,000 arterial waveform amplitudes (between scale factor of 1 and 10) and analyzing their complexity (i.e., multiscale entropy) in 4 different subject populations of different ages and disease status, significant changes in S_E with different scale factors in the 4 different groups of subjects were noted (see Figure 3).

3.1. Changes in Sample Entropy, S_E, with Scale Factor. The values of S_E decreased significantly from a scale factor of 6 onward in group 1 (healthy young subjects), group 2 (healthy middle-aged or elderly subjects), group 3 (well-controlled diabetics), and group 4 (poorly controlled diabetics) (see Figure 3). No significant difference in the values of S_E at lower scale factors (i.e., 1 to 5) was noted among the 4 groups.

3.2. Comparison between Healthy Young (Group 1) and Middle-Aged or Elderly (Group 2) Subjects. Remarkable differences were noted between healthy young (group 1) and middle-aged or elderly (group 2) subjects in terms of age, body height, HbA1c ($P < 0.001$), and serum HDL and LDL levels ($P < 0.05$; Table 1). Significant differences ($P = 0.016$) in DI were also noted between group 1 (201.57% ± 43.42%) and group 2 (164.88% ± 32.33%). No notable difference in $\mathrm{MEI_{SS}}$ was noted between the 2 groups (3.43 ± 1.23 versus 2.92 ± 0.89, $P = 0.343$); however, $\mathrm{MEI_{LS}}$ was significantly higher in group 1 than in group 2 (4.22±1.41 versus 3.53±0.99, resp., $P = 0.025$).

3.3. Comparison between Healthy Middle-Aged or Elderly (Group 2) and Well-Controlled Diabetic (Group 3) Subjects. Table 1 summarizes the demographic, anthropometric, hemodynamic, and biochemical parameters, MEI, and DI between group 2 and group 3 (HbA1c < 8%) subjects, showing notably advanced age, larger waist circumference, elevated HbA1c, and fasting blood sugar levels in the latter

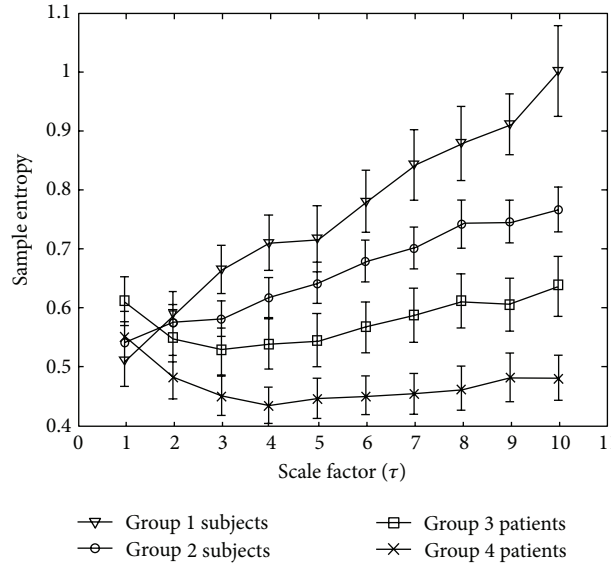


Figure 3: Changes in sample entropy (S_E) with different scale factors in the four groups of subjects. Symbols represent the mean values of entropy for each group, and bars represent the standard error ($S_E = \mathrm{SD}/\sqrt{n}$), where n is the number of subjects. Group 1: healthy young subjects; group 2: healthy middle-aged or elderly subjects; group 3: well-controlled diabetic subjects; group 4: poorly controlled diabetic subjects.

($P < 0.001$). Body weight, body mass index, and systolic blood pressure in group 3 were significantly higher than that in group 2. On the other hand, serum LDL and HDL levels in group 3 were significantly lower than that in group 2 ($P < 0.05$). Multiscale entropy analysis revealed significantly higher $\mathrm{MEI_{LS}}$ in group 2 than that in group 3 (3.53 ± 0.99 versus 3.02 ± 1.48, resp., $P = 0.037$), whereas there was no notable difference in $\mathrm{MEI_{SS}}$ between the 2 groups (2.92 ± 0.89 versus 2.78 ± 1.27 for group 2 and group 3, resp., $P = 0.452$). In terms of DI, no remarkable difference was noted between group 2 and group 3 (164.88% ± 32.33% versus 162.08% ± 35.34%, resp., $P = 0.365$). Moreover, a significant negative correlation was noted between $\mathrm{MEI_{LS}}$ and fasting blood sugar levels in the 2 groups ($R = -0.274$, $P = 0.015$) (see Figure 4(a)), whereas no notable correlation could be found between DI and fasting blood sugar levels between these groups ($R = -0.172$, $P = 0.132$) (see Figure 4(b)).

3.4. Comparison between Well-Controlled (Group 3) and Poorly Controlled Diabetic (Group 4) Subjects. Although the subjects in group 3 (HbA1c < 8%) were significantly older than those in group 4 (HbA1c > 8%), the comparison between the 2 groups revealed significantly higher HbA1c, LDL, fasting blood sugar, and triglyceride levels in group 4 (Table 1). There was no significant difference in $\mathrm{MEI_{SS}}$ between group 3 and group 4 (2.78 ± 1.27 versus 2.37 ± 0.88, resp., $P = 0.118$); however, $\mathrm{MEI_{LS}}$ was remarkably higher in the well-controlled diabetic subjects (group 3) than that in the poorly controlled diabetic subjects (group 4) (3.02 ± 1.48 versus 2.34±0.96, resp., $P = 0.024$). A notable difference in DI

Table 1: Comparison of demographic, anthropometric, hemodynamic and biochemical parameters, MEI, and DI between healthy young subjects (Group 1), healthy middle-aged or elderly subjects (Group 2), well-controlled diabetic subjects (Group 3) and poorly controlled diabetic subjects (Group 4).

Parameter	Group 1	Group 2	Group 3	Group 4
N	30	40	40	30
Age (years)	24.87 ± 2.69	$56.59 \pm 8.75^{**}$	$64.98 \pm 9.26^{++}$	$60.03 \pm 8.24^{\varepsilon}$
Body height (cm)	172.63 ± 6.86	$161.93 \pm 7.44^{**}$	160.55 ± 8.56	163.26 ± 7.16
Body weight (kg)	68.12 ± 10.99	63.31 ± 10.70	$68.09 \pm 10.28^{+}$	71.41 ± 11.93
Waist circumference (cm)	80.97 ± 9.55	82.11 ± 9.92	$93.13 \pm 9.37^{++}$	93.06 ± 11.62
BMI (kg/m^2)	22.79 ± 3.06	24.11 ± 3.59	$26.40 \pm 3.39^{+}$	26.98 ± 5.30
SBP (mmHg)	116.18 ± 12.31	118.11 ± 15.19	$128.34 \pm 17.02^{+}$	126.83 ± 17.66
DBP (mm Hg)	71.94 ± 6.18	73.94 ± 10.49	75.04 ± 10.14	74.72 ± 11.19
HbA_{1c} (%)	5.49 ± 0.25	$5.67 \pm 0.31^{**}$	$6.79 \pm 0.60^{++}$	$9.85 \pm 1.81^{\varepsilon\varepsilon}$
HDL (mg/dL)	44.81 ± 5.60	$52.94 \pm 20.64^{*}$	$42.78 \pm 16.26^{+}$	43.39 ± 14.65
LDL (mg/dL)	97.0 ± 26.83	$122.48 \pm 26.78^{*}$	$99.33 \pm 25.17^{++}$	$117.93 \pm 36.23^{\varepsilon}$
Fasting blood sugar (mg/dL)	92.69 ± 3.19	97.70 ± 15.76	$128.06 \pm 28.77^{++}$	$166.96 \pm 59.07^{\varepsilon}$
Triglyceride (mg/dL)	89.31 ± 60.14	105.09 ± 51.06	110.29 ± 41.71	$161.85 \pm 53.72^{\varepsilon}$
Creatinine (mg/dL)	0.92 ± 0.12	$0.79 \pm 0.22^{*}$	0.93 ± 0.37	1.24 ± 1.17
Microalbumin (mg/dL)	0.72 ± 0.56	0.64 ± 0.66	16.99 ± 57.99	71.68 ± 222.41
MEI_{SS}	3.43 ± 1.23	2.92 ± 0.89	2.78 ± 1.27	2.37 ± 0.88
MEI_{LS}	4.22 ± 1.41	$3.53 \pm 0.99^{*}$	$3.02 \pm 1.48^{+}$	$2.34 \pm 0.96^{\varepsilon}$
DI (%)	201.57 ± 43.42	$164.88 \pm 32.33^{*}$	162.08 ± 35.34	$132.72 \pm 36.57^{\varepsilon\varepsilon}$

Value are expressed as mean ± SD. BMI: body mass index; SBP: systolic blood pressure; DBP: diastolic blood pressure; HbA_{1c}: glycosylated hemoglobin; HDL: high density lipoprotein; LDL: low density lipoprotein; MEI_{SS}: Multiscale Entropy Index with Small Scale; MEI_{LS}: Multiscale Entropy Index with Large Scale; DI: Dilatation Index. $^{*}P < 0.05$: Group 1 versus Group 2, $^{+}P < 0.05$: Group 2 versus Group 3, $^{\varepsilon}P < 0.05$: Group 3 versus Group 4. $^{**}P < 0.001$: Group 1 versus Group 2, $^{++}P < 0.001$: Group 2 versus Group 3, $^{\varepsilon\varepsilon}P < 0.001$: Group 3 versus Group 4.

Table 2: Correlations of MEI_{LS} and DI with anthropometric, hemodynamic, and biochemical parameters.

Parameter	DI ($N = 140$)	MEI_{LS} ($N = 140$)
Age (years)	$R = -0.168$, $P = 0.062$	$R = -0.223$, $P = 0.012$
Body height (cm)	$R = 0.113$, $P = 0.144$	$R = -0.063$, $P = 0.440$
Body weight (kg)	$R = -0.078$, $P = 0.423$	$R = -0.127$, $P = 0.147$
Waist circumference (cm)	$R = -0.193$, $P = 0.043$	$R = -0.143$, $P = 0.117$
BMI (kg/m^2)	$R = -0.162$, $P = 0.043$	$R = -0.092$, $P = 0.309$
SBP (mmHg)	$R = -0.183$, $P = 0.054$	$R = -0.031$, $P = 0.735$
DBP (mmHg)	$R = -0.124$, $P = 0.195$	$R = 0.007$, $P = 0.937$
HbA_{1c} (%)	$R = -0.223$, $P = 0.013$	$R = -0.375$, $P < 0.001$
HDL (mg/dL)	$R = 0.034$, $P = 0.730$	$R = 0.240$, $P = 0.010$
LDL (mg/dL)	$R = -0.070$, $P = 0.478$	$R = -0.025$, $P = 0.791$
Fasting blood sugar (mg/dL)	$R = -0.169$, $P = 0.074$	$R = -0.344$, $P < 0.001$
Triglyceride (mg/dL)	$R = -0.165$, $P = 0.091$	$R = -0.158$, $P = 0.088$

BMI: body mass index; SBP: systolic blood pressure; DBP: diastolic blood pressure; HbA_{1c}: glycosylated hemoglobin; HDL: high density lipoprotein; LDL: low density lipoprotein.

also existed between group 3 and group 4 (162.08% ± 35.34% versus 132.72% ± 36.57%, resp., $P < 0.001$).

3.5. Correlations of MEI_{LS} and DI with Anthropometric, Hemodynamic, and Biochemical Parameters. Attempts were made to correlate values of DI and MEI_{LS} from all subjects ($N = 140$) with their anthropometric, hemodynamic, and biochemical risk factors of CVD (Table 2). The results showed that DI was negatively correlated with waist circumference, body mass index, and HbA1c levels. On the other hand, while MEI_{LS} was negatively correlated with age, HbA1c, and fasting blood sugar levels, it was positively correlated with serum HDL levels.

4. Discussion

The human body consists of physiological systems of dynamical complexity involving a myriad of interactions and feedback mechanisms [17]. Recent studies [16–19], which placed strong emphasis on the quantification of dynamical

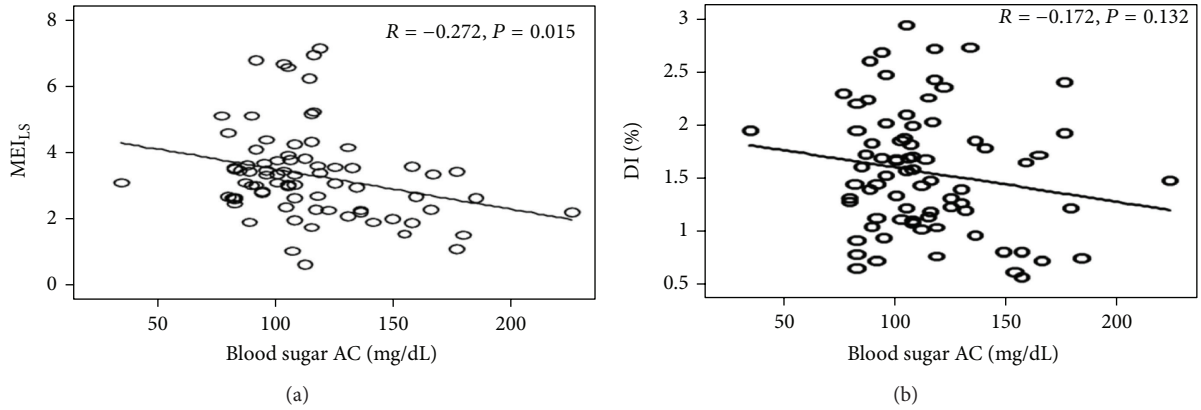


Figure 4: Correlations between (a) small-scale multiscale entropy index (MEI_{LS}) and fasting blood sugar levels; (b) dilatation index (DI) and fasting blood sugar levels in healthy middle-aged or elderly (group 2) and well-controlled diabetic (group 3) subjects.

complexity in healthy human subjects and those with cardiovascular diseases, have identified a reduction in dynamical complexity, defined by MSE, as a common characteristic of the aged and diseased subsets of the population. Previous applications of dynamical complexity analysis focused mainly on the study of *R-R* interval time series, in an attempt to investigate various cardiac diseases. For instance, compared with healthy subjects regardless of age, patients with congestive heart failure (CHF) have a higher S_E for scale 1 [16, 17]. In contrast, a lower S_E becomes apparent in subjects with CHF over scale 1. Analysis of *R-R* interval time series in normal subjects and in patients with ventricular arrhythmia and myocardial infarction revealed that S_E decreases with increasing age in both normal and diseased populations [24]. On the other hand, there is no significant difference in S_E between the healthy aged subjects and their counterparts with cardiac diseases. Moreover, healthy young subjects have the highest S_E at all scales compared with the aged and diseased groups [24].

The application of MSE in analyzing heart rate (HR) and systolic and diastolic blood pressure (BP) in 14 young patients with type 1 diabetes mellitus was first reported by Trunkvalterova et al. in 2008 [30]. MSE analysis of HR/BP signals showed a higher S_E value in the healthy subjects than that in the diabetic subjects on scale 3. Using age-matched healthy young subjects as normal controls, this study proposed that MSE is useful in detecting subtle vascular pathology in young diabetic subjects. However, the paradoxical result of MSE analysis on HR and diastolic BP in that study, which showed a higher S_E in diabetic patients compared with their healthy counterparts over scale 6, remains unexplained. The choice of a suitable physiological parameter is, therefore, essential in the successful application of MSE to the assessment of the degree of atherosclerosis and the effect of aging on vascular function.

Although the application of MSE using *R-R* interval time series in analyzing the dynamical complexity of cardiac diseases has been validated, reports on the use of MSE in assessing atherosclerotic change of blood vessels and the impact of age on the vascular system are rare.

Not only is endothelial dysfunction believed to precede microvascular changes of the cardiovascular system [1], it is also considered an indicator of atherosclerosis [1–4]. Previous studies have proposed a system of reactive RH-PAT, performed by the analysis of finger arterial pulse waves before and after applying pressure on the upper arm, in assessing vascular endothelial function. The popularity of its use, however, is restricted by the expensive equipment and the requirement of well-trained personnel for proper operation. The present study utilized APSS that we previously proposed to record the signals of arterial pulsations from the wrist before and after application of pressure on the upper arm [21]. After calculation of the DI, we attempted to assess vascular endothelial function by adopting MSE. We used it in calculating the dynamical complexity of the signals acquired from subjects belonging to different age groups and from subjects with different degrees of diabetic control, since diabetes and aging are both risk factors of atherosclerosis. In this manner, the two parameters of MEI_{LS} and MEI_{LS} were obtained and compared among the different groups.

Table 1 shows a notable difference in both MEI_{LS} and DI between healthy young (group 1) and middle-aged or elderly (group 2) subjects, whereas there was no significant difference in MEI_{SS} between the two groups. On the other hand, although DI did not differ between healthy middle-aged or elderly subjects (group 2) and well-controlled diabetic subjects (group 3), significant difference in MEI_{LS} existed between the 2 groups (Table 1). These results imply that MEI_{LS} can indicate subtle vascular changes even in well-controlled diabetic subjects, whose endothelial dysfunction is maintained at a relatively stable condition through lifestyle modification and medical control [7]. Further investigation revealed a negative correlation between fasting blood sugar levels and MEI_{LS}, whereas the correlation between fasting blood sugar levels and DI failed to reach statistical significance. In term of HbA1c levels, a better correlation was noted with MEI_{LS} ($P < 0.001$) than with DI ($P = 0.013$) (Table 2). Taken together, the findings suggest that MEI_{LS} may serve as a better indicator of subtle diabetes-associated vascular endothelial dysfunction and sugar control than DI,

indicating the possible use of MEI_{LS} as a sensitive indicator of vascular endothelial dysfunction that allows early therapeutic intervention.

When DI and MEI_{LS} were compared in terms of their correlations with the risk factors of CVD (Table 2), significant correlations were noted between DI and waist circumference ($R = -0.193$, $P = 0.043$), body mass index ($R = -0.162$, $P = 0.043$), and HbA1c ($R = -0.223$, $P = 0.013$), whereas significant correlations existed between MEI_{LS} and age ($R = -0.223$, $P = 0.012$), HbA1c ($R = -0.375$, $P < 0.001$), serum HDL ($R = 0.240$, $P = 0.010$), and fasting blood sugar levels ($R = -0.344$, $P < 0.001$). The results further suggest that MEI_{LS} may be a more sensitive indicator of endothelial dysfunction associated with aging and diabetes than DI. The superiority of MEI_{LS} over DI may be due to the fact that the latter utilizes two segments of representative 1-minute signals acquired before and after vascular occlusion, whereas the former analyzes all 14-minute signals from both the baseline and hyperemic phases using the MSE technique.

This study has unavoidable limitations. First, since the computation of MEI requires time-consuming detrending of signals and extensive MSE analysis, immediate information cannot be provided for the examinees. This problem can probably be solved by the development of appropriate software for data analysis. Second, the current study only recruited a relatively small number of subjects and focused on only a single disease. Further investigation is warranted to include a larger number of patients with diseases related to endothelial dysfunction, including stroke, angina, limb ischemia, and erectile dysfunction. Finally, the requirement for an occlusion pressure of up to 200 mmHg over the upper arm for 3 minutes may not be tolerated by some study subjects. This was the situation for 3 of our diabetic patients, who were subsequently excluded from the present study.

5. Conclusion

Using the method of MSE for nonlinear dynamical analysis of arterial pulse signals from the wrist, this study successfully detected subtle differences in dynamical complexity of the acquired signals from the young, the middle-aged or elderly, well-controlled, and poorly controlled diabetic subjects using the novel parameter MEI.

Conflict of Interests

The authors declare no conflict of interests.

Authors' Contribution

M.-T. Lo, C.-K. Sun, and J.-J. Chen equally contributed in this study compared with the corresponding author.

Acknowledgments

The authors would like to thank the Associate Editor—Professor Guang Hu—and the anonymous Reviewers for their insightful recommendations, which have significantly contributed to the improvement of this work. The authors would like to thank the volunteers involved in this study for allowing them to collect and analyze their data. The authors are grateful for the support of Texas Instruments, Taiwan, in sponsoring the low-power instrumentation amplifiers and ADC tools. This research was supported in part by the National Science Council under Grant nos. NSC 100-2221-E-259-030-MY2, NSC 101-2221-E-259-012 and the National Dong Hwa University on campus interdisciplinary integration Project no. 101T924-3. Professor M.-T. Lo was supported by NSC, Taiwan, Grant no. 100-2221-E-008-008-MY2, a joint foundation of CGH and NCU Grant nos. CNJRF-101CGH-NCU-A4 and VGHUST101-G1- 1-3. The authors also acknowledge NSC support for the Center for Dynamical Biomarkers and Translational Medicine, National Central University, Taiwan (NSC 100–2911-I-008-001).

References

[1] F. Grover-Páez and A. B. Zavalza-Gómez, "Endothelial dysfunction and cardiovascular risk factors," *Diabetes Research and Clinical Practice*, vol. 84, no. 1, pp. 1–10, 2009.

[2] P. O. Bonetti, L. O. Lerman, and A. Lerman, "Endothelial dysfunction: a marker of atherosclerotic risk," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 23, no. 2, pp. 168–175, 2003.

[3] R. M. J. Palmer, A. G. Ferrige, and S. Moncada, "Nitric oxide release accounts for the biological activity of endothelium-derived relaxing factor," *Nature*, vol. 327, no. 6122, pp. 524–526, 1987.

[4] V. Schächinger, M. B. Britten, and A. M. Zeiher, "Prognostic impact of coronary vasodilator dysfunction on adverse long-term outcome of coronary heart disease," *Circulation*, vol. 101, no. 16, pp. 1899–1906, 2000.

[5] O. T. Raitakari and D. S. Celermajer, "Flow-mediated dilatation," *British Journal of Clinical Pharmacology*, vol. 50, no. 5, pp. 397–404, 2000.

[6] M. C. Corretti, T. J. Anderson, E. J. Benjamin et al., "Guidelines for the ultrasound assessment of endothelial-dependent flow-mediated vasodilation of the brachial artery: a report of the international brachial artery reactivity task force," *Journal of the American College of Cardiology*, vol. 39, no. 2, pp. 257–265, 2002.

[7] P. O. Bonetti, G. M. Pumper, S. T. Higano, D. R. Holmes, J. T. Kuvin, and A. Lerman, "Noninvasive identification of patients with early coronary atherosclerosis by assessment of digital reactive hyperemia," *Journal of the American College of Cardiology*, vol. 44, no. 11, pp. 2137–2141, 2004.

[8] A. Nohria, M. Gerhard-Herman, M. A. Creager, S. Hurley, D. Mitra, and P. Ganz, "Role of nitric oxide in the regulation of digital pulse volume amplitude in humans," *Journal of Applied Physiology*, vol. 101, no. 2, pp. 545–548, 2006.

[9] P. O. Bonetti, G. W. Barsness, P. C. Keelan et al., "Enhanced external counterpulsation improves endothelial function in patients with symptomatic coronary artery disease," *Journal of the American College of Cardiology*, vol. 41, no. 10, pp. 1761–1768, 2003.

[10] U. Pohl, J. Holtz, R. Busse, and E. Bassenge, "Crucial role of endothelium in the vasodilator response to increased flow in vivo," *Hypertension*, vol. 8, no. 1, pp. 37–44, 1986.

[11] P. F. Davies, "Flow-mediated endothelial mechanotransduction," *Physiological Reviews*, vol. 75, no. 3, pp. 519–560, 1995.

[12] R. F. Furchgott and J. V. Zawadzki, "The obligatory role of endothelial cells in the relaxation of arterial smooth muscle by acetylcholine," *Nature*, vol. 288, no. 5789, pp. 373–376, 1980.

[13] S. Laurent, P. Boutouyrie, R. Asmar et al., "Aortic stiffness is an independent predictor of all-cause and cardiovascular mortality in hypertensive patients," *Hypertension*, vol. 37, no. 5, pp. 1236–1241, 2001.

[14] T. Münzel, T. Gori, R. M. Bruno, and S. Taddei, "Is oxidative stress a therapeutic target in cardiovascular disease?" *European Heart Journal*, vol. 31, no. 22, pp. 2741–2748, 2010.

[15] T. Nakagawa, K. Tanabe, B. P. Croker et al., "Endothelial dysfunction as a potential contributor in diabetic nephropathy," *Nature Reviews. Nephrology*, vol. 7, no. 1, pp. 36–44, 2011.

[16] M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Physical Review Letters*, vol. 89, no. 6, Article ID 068102, 4 pages, 2002.

[17] M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy analysis of biological signals," *Physical Review E*, vol. 71, no. 2, Article ID 021906, 18 pages, 2005.

[18] M. Costa, A. L. Goldberger, and C. K. Peng, "Entropy to distinguish physiological and synthetic RR time series," *Computers in Cardiology*, vol. 29, pp. 137–140, 2002.

[19] R. A. Thuraisingham and G. A. Gottwald, "On multiscale entropy analysis for physiological data," *Physica A*, vol. 366, pp. 323–332, 2006.

[20] K. G. Alberti and P. Z. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation," *Diabetic Medicine*, vol. 15, no. 7, pp. 539–553, 1998.

[21] H.-T. Wu, C.-H. Lee, and A.-B. Liu, "Assessment of endothelial function using arterial pressure signals," *Journal of Signal Processing Systems*, vol. 64, no. 2, pp. 223–232, 2011.

[22] H. T. Wu, C. H. Lee, A. B. Liu et al., "Arterial stiffness using radial arterial waveforms measured at the wrist as an indicator of diabetic control in the elderly," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 243–252, 2011.

[23] H. T. Wu, C. C. Liu, P. H. Lin et al., "Novel application of parameters in waveform contour analysis for assessing arterial stiffness in aged and atherosclerotic subjects," *Atherosclerosis*, vol. 213, no. 1, pp. 173–177, 2010.

[24] E. Tejera, J. M. Nieto-Villar, and I. Rebelo, "Unexpected heart rate variability complexity in the aging process of arrhythmic subjects," *Communications in Nonlinear Science and Numerical Simulation*, vol. 15, no. 7, pp. 1858–1863, 2010.

[25] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hubert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society A*, vol. 454, no. 1971, pp. 903–995, 1998.

[26] N. E. Huang, M. L. C. Wu, S. R. Long et al., "A confidence limit for the empirical mode decomposition and Hilbert spectral analysis," *Proceedings of the Royal Society A*, vol. 459, no. 2037, pp. 2317–2345, 2003.

[27] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate and sample entropy," *American Journal of Physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[28] H. K. Yuan, C. Lin, P. H. Tsai et al., "Acute increase of complexity in the neurocardiovascular dynamics following carotid stenting," *Acta Neurologica Scandinavica*, vol. 123, no. 3, pp. 187–192, 2011.

[29] D. Cheng, S. J. Tsai, C. J. Hong, and A. C. Yang, "Reduced physiological complexity in robust elderly adults with the APOE $\varepsilon 4$ allele," *PLoS ONE*, vol. 4, no. 11, Article ID e7733, 2009.

[30] Z. Trunkvalterova, M. Javorka, I. Tonhajzerova et al., "Reduced short-term complexity of heart rate and blood pressure dynamics in patients with diabetes mellitus type 1: multiscale entropy analysis," *Physiological Measurement*, vol. 29, no. 7, pp. 817–828, 2008.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 930362, 9 pages
http://dx.doi.org/10.1155/2013/930362

Research Article

crcTRP: A Translational Research Platform for Colorectal Cancer

Ning Deng, Ling Zheng, Fang Liu, Li Wang, and Huilong Duan

Department of Biomedical Engineering, Key Laboratory for Biomedical Engineering of Ministry of Education, Zhejiang University, Hangzhou 310027, China

Correspondence should be addressed to Huilong Duan; duanhl@zju.edu.cn

Received 8 November 2012; Accepted 2 January 2013

Academic Editor: Bairong Shen

Colorectal cancer is a leading cause of cancer mortality in both developed and developing countries. Transforming basic research results into clinical practice is one of the key tasks of translational research, which will greatly improve the diagnosis and treatments of colorectal cancer. In this paper, a translational research platform for colorectal cancer, named crcTRP, is introduced. crcTRP serves the colorectal cancer translational research by providing various types of biomedical information related with colorectal cancer to the community. The information, including clinical data, epidemiology data, individual omics data, and public omics data, was collected through a multisource biomedical information collection solution and then integrated in a clinic-omics database, which was constructed with EAV-ER model for flexibility and efficiency. A preliminary exploration of conducting translational research on crcTRP was implemented and worked out a set of clinic-genomic relations, linking clinical data with genomic data. These relations have also been applied to crcTRP to make it more conductive for cancer translational research.

1. Introduction

Nowadays, cancer is still one of the major diseases that endanger human life. As American Cancer Society reported, a total of 1,638,910 new cancer cases and 577,190 deaths from cancer were projected to occur in the United States in 2012 [1]. Among all kinds of cancer, colorectal cancer is the second leading cause of cancer death in the United States and the fifth leading cause in China [2, 3]. Though researches focusing on the molecular mechanism of colorectal cancer have made great progress, effective clinical measurements for early diagnoses and treatments are still very scarce due to the wide gap between laboratory research and clinical practice. Therefore, it is of great significance to conduct translational research, which can promote the transforming of basic research findings on colorectal cancer into clinical practice [4], for reducing the mortality of colorectal cancer.

Obtaining required data efficiently and conveniently is a necessity for the smooth conduct of translational research. However, biomedical data are usually stored in heterogeneous databases with different terminologies, since they are generin several domains including clinical practice, epidemiology survey, and laboratory research by different institutions. Organizing and managing various types of biomedical data efficiently and then sharing these data through a public platform can help researchers surmount difficulties of data acquisition, which will contribute a lot to colorectal cancer translational research.

In this study, we developed a translational research platform, named crcTRP, aiming at accelerating translational research for colorectal cancer. crcTRP is consisted of a multisource biomedical information collection solution, a clinic-omics database, and a web portal. The multisource biomedical information collection solution focuses on acquiring different types of biomedical information related with colorectal cancer from different data sources. The clinic-omics database is developed for integrating various types of data obtained through the information collection solution reasonably and effectively. While the web portal aims at sharing the whole

information we acquired among physicians, molecular biologists, and other researchers related with colorectal cancer translational research. Using information provided by crcTRP, we then generated a set of clinic-genomic relations based on UMLS [5], which is a typical example for illustrating what contributions can be made by crcTRP in translational research. These relations were then been applied to crcTRP to make it better for translational research.

2. Methods

First, a multisource biomedical information collection solution was proposed to collect various biomedical data related to colorectal cancer, including clinical data, epidemiology data, individual omics data, and public omics data. Second, a clinic-omics database was constructed to organize and manage the obtained dataset. Third, a web portal was designed to share the obtained information among various researchers related with colorectal cancer translational research. Finally, we proposed a data mining method to map clinic-genomic relations from crcTRP.

2.1. Acquisition of Multisource Biomedical Information. The acquisition of multisource biomedical information includes clinical information collection based on SNOMED CT [6], epidemiology data collection based on in-house web application, individual omics data collection based on MIAME [7], and public biomedical data collection based on OMIM [8].

2.1.1. Clinical Data Acquisition Based on SNOMED CT. With the rapid progress of health information technology, Electronic Medical Records systems (EMR systems) have been widely used in the hospitals. This way, it is possible for us to collect clinical data electronically. We designed a data acquisition interface to collect deidentified clinical data from these EMR systems. Private information such as name, date of birth, and medical record ID is removed so that the patient cannot be directly identified. The same clinical concept may have different expressions in different information systems, which will reduce the efficiency of data utilization. So we coded the clinical information standardizedly using SNOMED CT. As a compositional concept system, SNOMED CT provides a compositional syntax for building new biomedical concepts. This feature makes SNOMED CT present complex clinical concepts about colorectal cancer appropriately. Concepts already included in SNOMED CT were precoordinated, while concepts not included were postcoordinated. For example, "severe pain in the left abdomen" cannot be expressed using concepts in SNOMED CT directly, but it can be postcoordinated as:

21522001 |abdominal pain|:
272741003 |laterality| = 77710000 |left|,
246113005 |severity| = 24484000 |severe|.

2.1.2. Epidemiology Data Acquisition Based on In-House Web Application. Epidemiology data can help us gain a comprehensive understanding of high-risk population's dietary habits, personal medical history, family disease history and other factors which may cause the disease. We developed a web-based epidemiology data acquisition system for colorectal cancer. Designed with Browser/Server architecture, this system can be loaded to several mobile platforms. Moreover, by using advanced features of HTML5 including novel client-side storage method and offline web applications [9], this system can work on mobile devices smoothly even in places without network connection. Once the system has been loaded to a mobile device, researchers can collect epidemiology data using the mobile device at bedsides, communities, or other areas with highincidence of colorectal cancer (usually remote mountain areas) without concern about the internet connection.

2.1.3. Individual Omics Data Acquisition Based on MIAME. Since translational informatics is ready to revolutionize human health and healthcare using large-scale measurements on individuals [10], massive individual omics data will emerge as expected. Biospecimens are the physical sources of individual omics data. Well-annotated biospecimens will aid scientists to validate their research and correlate their findings with the associated pathological and clinical annotations. Collecting specimen through a relatively standardized procedure can obtain more useful and valid annotation information. Therefore, we developed an information managing system for biospecimen and individual omics data to collect personalized genomic data as well as their annotation information. Our management procedure for specimens was inspired by caTissue [11], a biospecimen informatics system of caBIG [12], while the collected individual omics data were designed to satisfy the MIAME guidelines, which outline the minimum information about a microarray experiment.

2.1.4. Public Biomedical Data Acquisition Based on OMIM. With the rapid development of high throughput technology, massive amounts of molecular biology data such as nucleic acid data and protein data have emerged. Moreover, owing to the sharing culture in this field, data are now increasingly publicly available [13]. However, molecular biology data in different types are often stored in different database, which makes it inconvenient for researchers to utilize these data comprehensively. We proposed a scheme for extracting information about colorectal cancer from these databases through OMIM. OMIM, a comprehensive, authoritative compendium of human genes and genetic phenotypes, is freely available and updated frequently. First, we searched for genes related to colorectal cancer using the keywords combination "(colon or colorectal or colonic or rectal) and (cancer or carcinoma or adenoma)" in OMIM. We limited the search fields as title and allelic variants, considering that content provided by title and allelic variants field associates with the corresponding theme most closely [14]. Second, we searched for protein IDs and article IDs using the obtained gene IDs in files provided by NCBI [15], which are about the associations between genes and other biomedical data. Finally, we extracted gene, protein, and literature information related to colorectal cancer from public biomedical databases including NCBI Gene [16],

RefSeq [17], PubMed [18], and Swiss-Prot [19] using IDs obtained in the first two steps.

2.2. Construction of a Clinic-Omics Database. Data collected through the multisource biomedical information collection solution can be categorized into individual information and public information. Individual information means information centering on patients, including clinical data, epidemiology data, and individual omics data. Public information refers to information collected from pubic biomedical database, including gene, protein, and literature information. Public information centers on omics data.

To manage all the data in a database efficiently, we analyzed the characteristics of these data. With the rapid development of molecular biology, new attributes of biomarkers or new biomarkers will be found undoubtedly, which results in the dynamic nature of individual omics data. Meanwhile, it is impossible for a certain patient to have all the items of clinical data, suggesting that clinical data such as symptoms and examinations of colorectal cancer are sparse. For example, not all patients need an MRI examination, causing the corresponding data item recording the test to result in the conventional database being null in most cases. On the contrary, epidemiology data and public information we acquired are relatively stable. Traditional ER model is not suitable for the dynamic and sparse nature of clinical omics data because of the fixed database schema. The EAV model, recording data based on entity-attribute-value, is more flexible and has the advantage of being able to store dynamic and sparse data efficiently [20]. However, when it comes to attribute-centered query, the most frequent query mode in translational research, EAV model is less efficient than the conventional ER model. Therefore, we introduced EAV-ER mixed model, which combines advantages of the above two models, to construct the clinic-omics database. EAV tables are designed for dynamic and sparse data while relatively stable data are stored in conventional tables.

Data relationships in the clinic-omics database are illustrated as Figure 1. Individual information is organized using Patient ID, while public information is organized using Gene ID. Individual information connects with public information by Gene ID of individual omics data.

2.3. Design of a Web Portal. Sharing comprehensive biomedical information between clinical physicians and basic researchers does not mean providing data access to them simply. Reasonable information distribution layout should be taken into consideration for researchers to find their expected data quickly. We designed a web portal with framework shown in Figure 2 to satisfy this requirement.

Homepage gives an overview of the web portal and links to news reporting the latest research progress in colorectal cancer translational research. The information query engine of crcTRP's web portal consists of standard query and advanced query, to satisfy different query requirements. Besides, since different researchers may focus on different information items, query results are designed to be configurable. Researchers can download data they interested in for future use. Integrated view of patient information provides a summary view of the selected patient's whole information including clinical information, epidemiology survey results and personalized molecular biological information. Public biomedical information library offers researchers with systematic knowledge of colorectal cancer. Information in the library can be accessed through two different ways. One is clicking a certain gene in the gene list. The other is clicking a certain gene in the chromosome map, which depicts the distribution of genes highly relative with colorectal cancer in all chromosomes. Genes in the chromosome map are extracted from OMIM using method described in Section 2.1.

2.4. Clinic-Genomic Information Mapping Based on UMLS. Relations between clinical data and genomic data may build a bridge between clinical practice and basic research, which is one of the key tasks of translational bioinformatics [21]. Various kinds of information acquired using the multisource biomedical information solution can be used for the potential clinic-genomic relation mining. The UMLS, or Unified Medical Language System, is a set of files and software that brings together many health and biomedical vocabularies and standards. UMLS covers a wealth of biomedical concepts from both clinical domains and molecular biology domains. Various relations among these concepts are also recorded in UMLS, laying foundation for us to implement the clinic-genomic information mapping. Installing UMLS locally will yield a series of RRF (Rich Release Format) files, including MRCONSO.RRF, MRREL.RRF, MRCOC.RRF, and MRSTY.RRF. MRCONSO.RRF lists out all concepts; MRREL.RRF contains information about the relationship between two concepts; MRCOC.RRF records cooccurring concepts; and MRSTY.RRF contains the semantic information on the concepts [22]. Two methods, namely, direct information mapping and indirect information mapping via disease, were proposed for information mapping based on UMLS. Mapping methods are shown in Figure 3.

Figure 3(a) illustrates direct information mapping method. First, search MRCONSO.RRF with source concept to get CUI (Concept Unique Identifier) of the source concept. Second, search MRREL.RRF and MRCOC.RRF to obtain concepts related with source concept using obtained CUI. Finally, search MRSTY.RRF file to obtain the semantic type of concepts found in the second step and pick out concepts with the desired semantic type. To this end, source concept is mapped to the picked out concepts. This method can be used for various kinds of information mapping. As for clinic-genomic information mapping, genomic concepts used here, called G, are genes extracted from OMIM using method described in Section 2.1. Clinical concepts used here, called C, are clinical items selected from clinical information acquired also using method described in Section 2.1. Since the clinical information we collected has been coded using SNOMED CT, concepts in C are lots of SNOMED CT codes.

Two kinds of direct information mapping, mapping from genomic to clinic or in turn, are distinguished from each other by mapping direction. When source concept is genomic

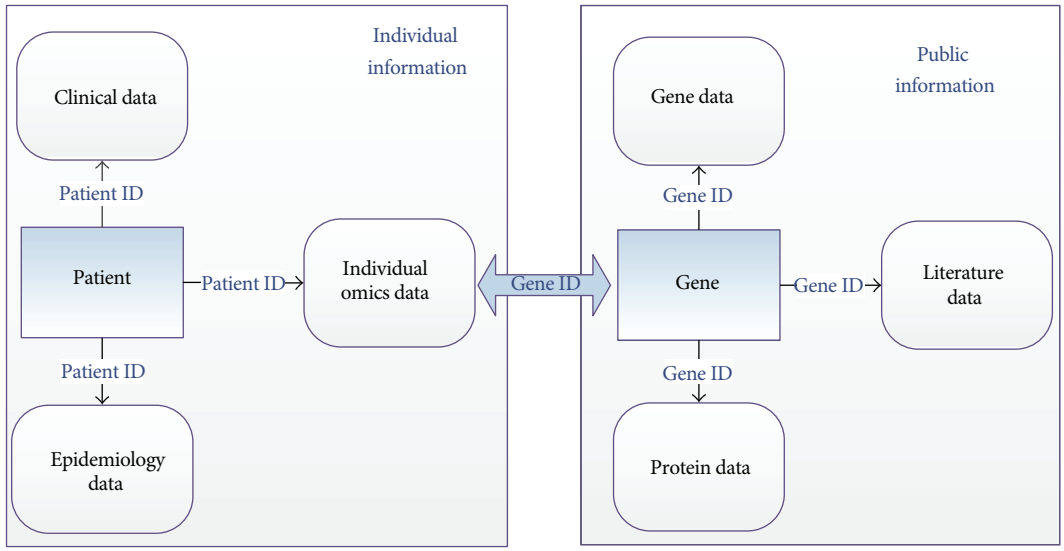


Figure 1: Data Relationships of the Clinic-omics Database.

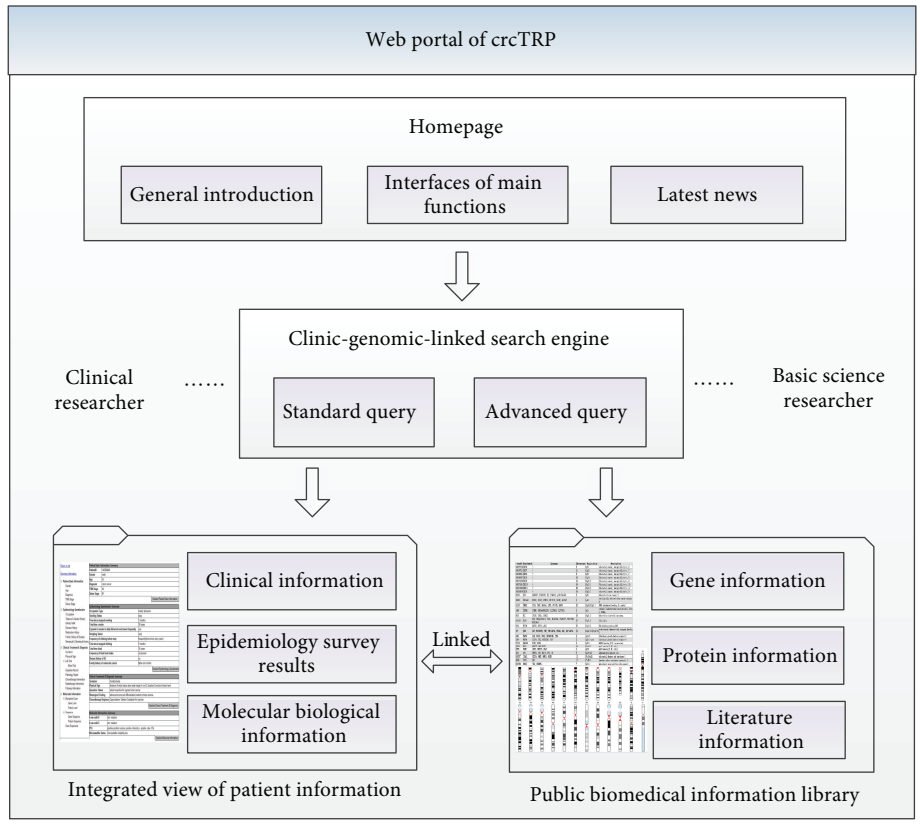
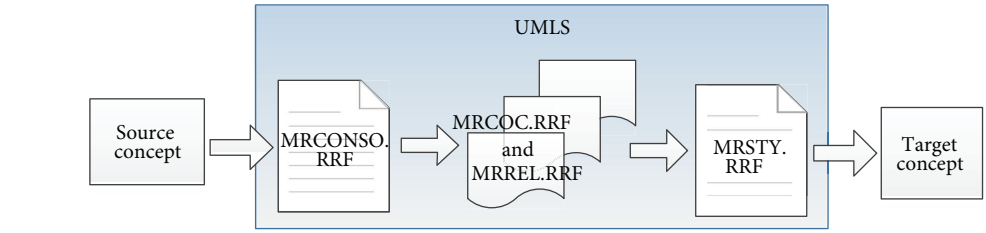


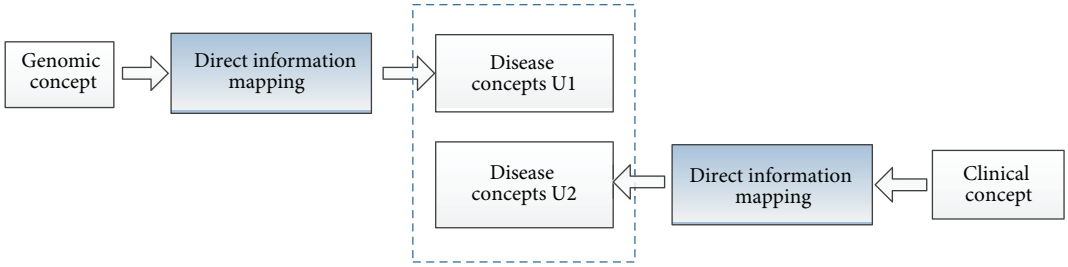
Figure 2: Framework of crcTRP's Web portal.

concept from G, target concepts of direct information mapping aslo included in C are mapped to the genomic concept. Similarly, when source concept is clinical concept from C, target concepts of direct information mapping also included in G are mapped to the clinical concept.

Genomic information reflects disease from the micro side, indicating the mechanism of disease. Meanwhile, Clinical information reflects disease from the macro side, recording symptoms or manifestations of disease. Therefore, disease concepts can be used to relate clinical data with genomic data. The procedure of indirect mapping via disease concepts is shown in Figure 3(b). First, disease concepts mapped to genomic concept, named as U1, are obtained using direct mapping method by picking out concepts with disease



(a) Direct information mapping based on UMLS



(b) Indirect information mapping via disease

Figure 3: Clinic-genomic information mapping based on UMLS.

semantic type in the last step of direct information mapping. Second, disease concepts mapped to clinical concept, named as U2, are also obtained using the same method. Then, if a disease concept related with colorectal cancer presents in both U1 and U2, the genomic concept is thought to be mapped to the clinical concept.

3. Results and Discussions

A translational research platform for colorectal cancer, named crcTRP, was developed and lots of clinic-genomic relations were found out based on information provided by the platform. Several kinds of biomedical data were collected using a multisource biomedical information collection solution and integrated in a clinic-omics database. crcTRP serves for colorectal cancer translational research by sharing these data through a unified web portal. Those clinic-genomic relationships were out of the preliminary exploration of the capabilities of the platform and have been used to relate integrated view of patient's comprehensive information with public biomedical information library on the web portal of crcTRP, which enriched crcTRP in return. Much more potential research fruits are expected to be gained by users of crcTRP.

3.1. crcTRP. crcTRP consists of three parts as follows.

3.1.1. A Multisource Biomedical Information Collection Solution. Based on methods described in Section 2.1, the information collection solution is able to collect clinical data, epidemiology data, individual omics data as well as public omics data. We have collected comprehensive data of more than 150 patients with colorectal cancer at present. Besides, a total of 384 data items related with clinical and epidemiological information have been coded using SNOMED CT. Epidemiology information collected using our epidemiology questionnaire system covers patients' dietary habits, personal medical history, family disease history, and other risk factors of cancer. A total of 62 genes related with colorectal cancer were extracted from OMIM, including BRAF [23] and APC [24], two important biomarkers of colorectal cancer. In addition, 54 proteins and 2006 articles related to colorectal cancer were extracted from public biomedical database by using the 62 genes.

3.1.2. A Clinic-Omics Database. A database with the capability of integrating clinical and omics data was constructed. Data acquired using the multisource information collection solution were already stored in this database, which illustrated the integration capabilities of the database well. With EAV-ER model, the database is flexible enough to adapt to the dynamic and sparse nature of biomedical data while guaranteeing the efficiency of attribute-centered query. For example, biomaterial and biomarker are entities and a biomaterial may have several biomarkers. We constructed two tables named "Biomaterial" and "Gene_Biomarker" to stand for biomaterial and biomarker, respectively. Obviously, "Biomaterial" and "Gene_Biomarker" form a typical ER model. Gene sequence and mutation status are two common attributes of biomarker. In table "Gene_Biomarker," column "Sequence" and "Is_Mutation" are used to record the corresponding attributes according to the ER mode, which is beneficial to the attributed-centered query. However, with rapid development of molecular biotechnology, new attributes will be found undoubtedly. In order to record those new attributes without modifying database architecture, we designed another two tables, "New_Gene_Biomarker" and "New_Gene_Biomarker_Attribute." "New_Gene_Biomarker" was designed to be an EAV table, including three columns to record entity, attribute and value respectively. New

attribute can be expressed via referring the primary key of table "New_Gene_Biomarker_Attribute" by "New_Gene_Biomarker." This way, both ER model and EAV model are employed, forming the mixed EAV-ER model, to achieve the best performance as much as possible. The clinic-omics database was implemented in SQL Server 2008 with a total of 174 data tables, of which 19 for clinical data, 50 for epidemiology data, 99 for individual omics data, and 6 for public biomedical information.

3.1.3. A Web Portal. The web portal was developed using ASP.NET in Microsoft Visual Studio 2010 and has been deployed in IIS server. It can be accessed via http://60.191.25.26:8088/. This web portal is useful for both clinical researchers and molecular biologists. For clinical researchers, they can learn the molecular mechanism of colorectal cancer, which may lead to better understanding about the diagnosis, therapy, and prognosis of colorectal cancer. For molecular biologists, they can study the function and phenotype of genes and molecular pathways. Figure 4 shows some screenshots of the web portal (red annotations and blue arrows will be explained later). Figure 4(a) is the integrated view of a certain patient's comprehensive information. Information covering by the integrated view includes patient basic information, epidemiology information, clinical information, and molecular information. Tree view on the left can bring researchers to the right part quickly. Figure 4(b) shows a gene list, which is one of the entrances of public information library. Clicking any one record on the list will come to the detail information of the selected gene, as shown in Figure 4(c). Detailed information integrates gene, protein, literature and other information, giving a systematic knowledge to researchers.

3.2. A Set of Clinic-Genomic Relations about Colorectal Cancer. A total of 50 clinical information items and 62 genes selected from the clinic-omics database, forming C and G mentioned in Section 2.4, were used for the information mapping. Using direct mapping method, we detected 170 relations between clinical data and genomic data, 142 from genomic to clinic and 28 from clinic to genomic. Using indirect mapping method, we detected 611 relations. Combining all these mapping results, we generated a total of 781 candidate clinic-genomic relations for colorectal cancer, without consideration about the overlap between relations generated from direct method and those from indirect method.

Appearing in the same article, a clinical concept and a genomic concept can be considered as relating to each other in some ways. Based on this idea, we proposed a relation validation method to select more closely linked relations by searching published articles. PubMed [18] is the most widely used biomedical literature retrieval system. Most articles in PubMed are assigned to Medical Subject Headings (MeSH) [25], a controlled vocabulary thesaurus of indexing terms arranged in a hierarchy, which provides a consistent way to find citations, even when authors use different terms for the same concept [26]. Furthermore, using MeSH has been shown to improve the efficiency of search, which means retrieving fewer irrelevant citations [26].

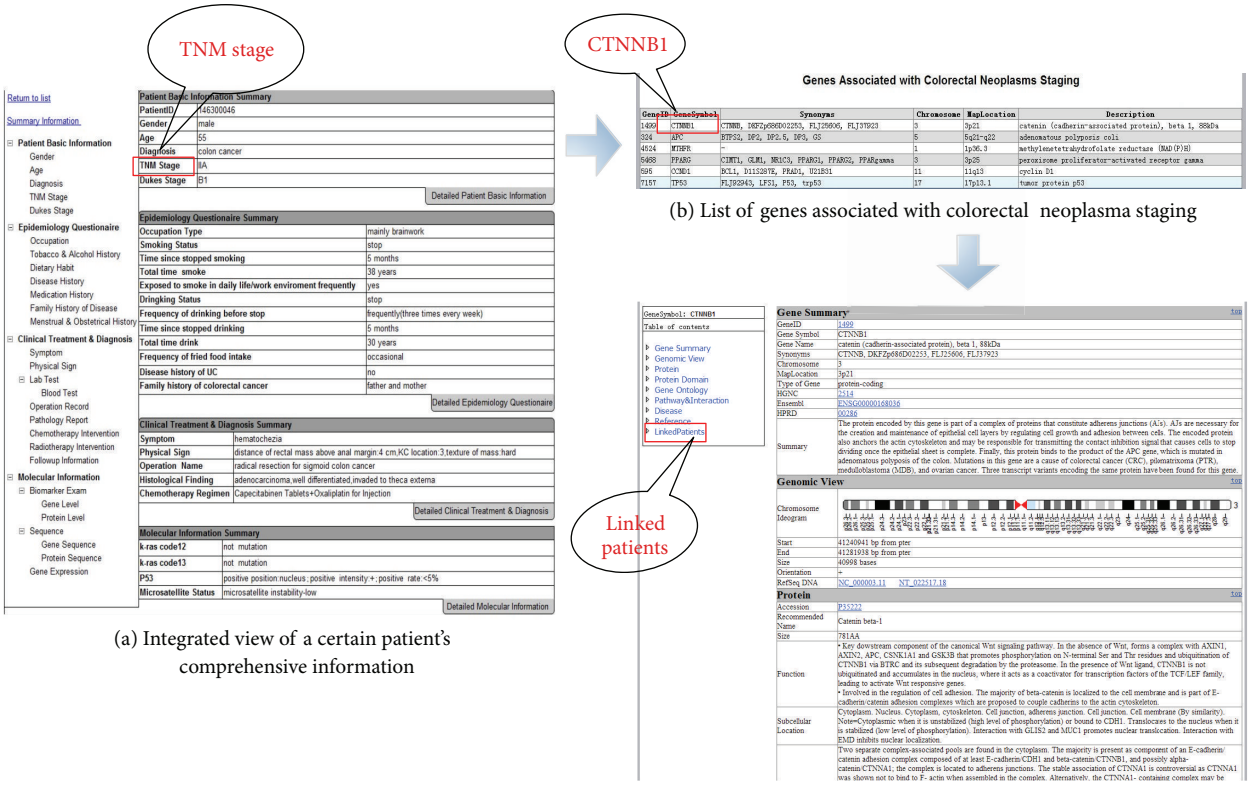
Table 1: Representative clinic-genomic relations about colorectal cancer.

Clinical data items	Genes
Abdominal pain	FOS, HFE, NRAS, TP53
Blood glucose	BAX, CCND1, CTNNB1, FGFR3, FOS, PPARG, SRG, TLR2, TP53
Carcinoembryonic antigen	APC, BAX, CCND1, CEACAM5, CEACAM7, CEACAM1, CTNNB1, DCC, EGFR, ERBB2, MLH1, MSH2, PSG2, SRC, TLR2, TP53
CA19-9 antigen	CTNNB1, NRAS, TP53, SRC
Colorectal neoplasms staging	**APC, CCND1, CTNNB1, MTHFR, PPARG, TP53**
Crohn's disease	APC, BAX, CCND1, CTNNB1, DCC, MTHFR, NRAS, PPARG, TLR2, TP53
Diabetes mellitus	APC, BAX, CCND1, CHEK2, CTNNB1, DCC, FGFR3, MTHFR, NRAS, PPARG, PTPN12, SRC, TLR2, TP53
Dyspepsia	PPARG
Intestinal obstruction	APC, CCND1, MSH2, PPARG, TLR2
Lymphatic metastases	MCC, TP53, APC, BAX, CCND1, CTNNB1

Taking the above factors into consideration, specific scheme of our relation validation method is to search PubMed online database based on MeSH ontology. The validation procedure is shown in Figure 5. Given that two concepts, X and Y, form a relation pairwise (X, Y). To validate this relation, MeSH terms of X and Y, denoted as X' and Y', respectively, are extracted by searching MRCONSO.RRF from UMLS using CUI of X and Y. Then, an online literature search is performed through NCBI E-utilities, a program provided by PubMed. The search keywords are defined as "X'[MeSH Terms] AND Y'[MeSH Terms] AND Colorectal neoplasms [MeSH Terms]." Term of "Colorectal neoplasms" guarantees the hit papers are related with colorectal cancer. If there is at least one article meeting the searching condition, the relation pairwise (X, Y) is thought to have passed the validation. While those relations having not passed this validation are treated as less closely linked relations at current cognitive level of human.

After validation, a total of 249 relations remained. Representative relations are presented in Table 1. Clinical data items listed in Table 1 cover most types of the focused clinical information related to treatments for colorectal cancer. These types of clinical information include symptoms, blood test results, pathology states, and diseases highly related with colorectal cancer.

We have successfully utilized these relations on crcTRP to annotate the collected clinical and omics dataset. Noticing the bold text in Table 1, clinical data item "Colorectal Neoplasm Staging" is related with six genes, including APC,



(a) Integrated view of a certain patient's comprehensive information

(b) List of genes associated with colorectal neoplasma staging

(c) Detail information of CTNNB1

Figure 4: Representative screenshots of crcTRP web portal.

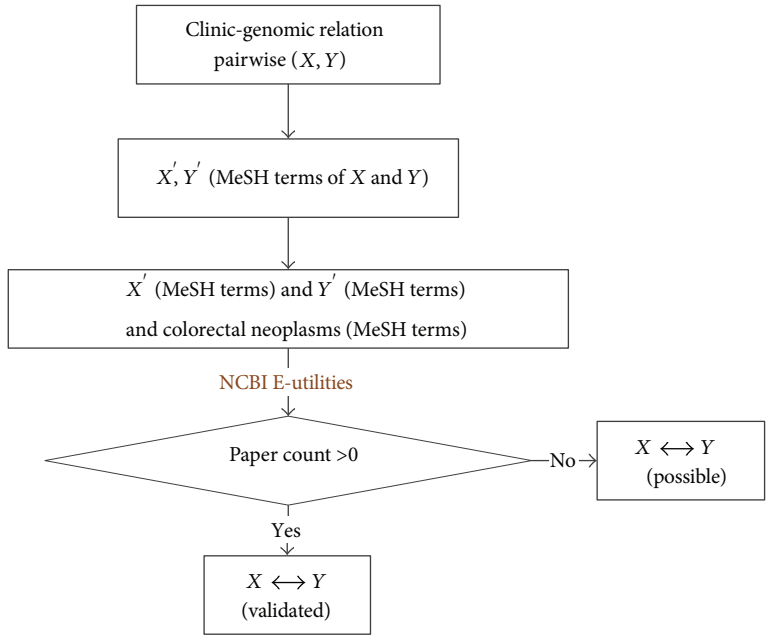


Figure 5: Procedure of PubMed based relation validation.

CCND1, CTNNB1, MTHFR, PPARG, and TP53. Clicking the "TNM staging," another expression of "Colorectal Neoplasm Staging," on Figure 4(a), browser will jump to the gene list page, which lists the six genes associated with "Colorectal Neoplasm Staging," shown as Figure 4(b). CTNNB1, the first record in the gene list, has been proved to play roles in colorectal cancer [27]. Specifically, tumor CTNNB1 status has substantial modifying effects on the beneficial prognostic role of postdiagnosis physical activity [27]. Clicking CTNNB1 on the gene list will bring out the detailed information of CTNNB1, shown as Figure 4(c). Meanwhile, IDs of patients who have the clinical items related to gene CTNNB1 are listed in the "Linked Patient" part of the gene detail information page. In this way, both clinical researchers and molecular biologists may benefit from these relations. On one hand, clinical researchers may gain a clearer understanding of the mechanism of disease by viewing biology molecular information related with a certain clinical item. On the other hand, molecular biologists may discover new roles of genes from these relations or get new ideas about next research projects through patient information related with a certain gene.

More work can be done to make better meaningful use of these relations. For example, quantifying these relations and then visualizing the quantified relations using various visualization means will offer researchers with a more intuitive understanding of these relations, which may lead to another discovery.

4. Conclusions

Colorectal cancer is the fifth leading cause of cancer mortality in China. Integrating existing clinical, genomic, and proteomic information and sharing these information through a unified platform will expedite the transformation of basic research results into clinical practice, and therefore promote the development of innovative treatment strategies. The main contributions of our research are concluded as follows.

First, we developed a platform named crcTRP serving for colorectal cancer translational research. crcTRP is the first platform offering various biomedical data to researchers working for colorectal cancer translational research through a unified web portal.

Second, we brought out a set of clinic-genomic relations based on crcTRP and then put these relations into use on crcTRP to serve for translational research. Not only these relations themselves are very significant for bridging clinical data with genomic data, but also the idea of how to take advantage of crcTRP is worth learning.

We view crcTRP as a good starting point. Much more work remains to be done to optimize crcTRP or take advantages of crcTRP. For example, data analysis tools can be developed and published on the web portal of crcTRP for online use. Moreover, data provided by crcTRP are excellent resource for data mining and data analysis, especially for relation mining owing to the abundant data types. At present, we are conducting a research about clinic-genomic relation mining based on large repositories of gene expression data. Clinical data provided by crcTRP aid us greatly in this ongoing research.

Acknowledgments

This work was supported by the National High Technology Research and Development Programs of China (863 Programs, no. 2012AA02A601, no. 2012AA020201), and the National Natural Science Foundation of China, no. 31100592.

References

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

[2] L. H. Kushi, T. Byers, C. Doyle et al., "American Cancer Society guidelines on nutrition and physical activity for cancer prevention: reducing the risk of cancer with healthy food choices and physical activity," *Ca: A Cancer Journal for Clinicians*, vol. 56, pp. 254–281, 2007.

[3] P. Zhao, M. Dai, W. Chen, and N. Li, "Cancer Trends in China," *Japanese Journal of Clinical Oncology*, vol. 40, no. 4, pp. 281–285, 2010.

[4] I. N. Sarkar, "Biomedical informatics and translational medicine," *Journal of Translational Medicine*, vol. 8, article 22, 2010.

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

[6] T. S. De Silva, D. Ma, G. Paterson, K. C. Sikdar, and B. Cochrane, "Systematized nomenclature of medicine clinical terms (SNOMED CT) to represent computed tomography procedures," *Computer Methods and Programs in Biomedicine*, vol. 101, no. 3, pp. 324–329, 2011.

[7] A. Brazma, P. Hingamp, J. Quackenbush et al., "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data," *Nature Genetics*, vol. 29, no. 4, pp. 365–371, 2001.

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

[9] "HTML5, A vocabulary and associated APIs for HTML and XHTML," 2012, http://www.w3.org/TR/html5/.

[10] N. H. Shah, "Translational bioinformatics embraces big data," *Yearbook of Medical Informatics*, vol. 7, no. 1, pp. 130–134, 2012.

[11] "caTissue Suite," 2012, https://wiki.nci.nih.gov/display/caTissue/caTissue+Suite.

[12] K. K. Kakazu, L. W. Cheung, and W. Lynne, "The Cancer Biomedical Informatics Grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research," *Hawaii medical journal*, vol. 63, no. 9, pp. 273–275, 2004.

[13] A. J. Butte, "Translational bioinformatics: coming of age," *Journal of the American Medical Informatics Association*, vol. 15, no. 6, pp. 709–714, 2008.

[14] OMIM, 2012, http:// www.ncbi.nlm.nih.gov/omim.

[15] GENE, 2012, http://www.ncbi.nlm.nih.gov/gene.

[16] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez gene: gene-centered information at NCBI," *Nucleic Acids Research*, vol. 39, no. 1, pp. D52–D57, 2011.

[17] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Research*, vol. 33, pp. D501–D504, 2005.

[18] PubMed, 2012, http://www.ncbi.nlm.nih.gov/pubmed/.

[19] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000," *Nucleic Acids Research*, vol. 28, no. 1, pp. 45–48, 2000.

[20] L. Marenco, P. Nadkarni, E. Skoufos, G. Shepherd, and P. Miller, "Neuronal database integration: the Senselab EAV data model," *AMIA Annual Symposium Proceedings*, pp. 102–106, 1999.

[21] J. Nakaya, "The translational research informatics (TRI)," *International Journal of Computer Science and Network Security*, vol. 6, no. 7, pp. 117–121, 2006.

[22] "Metathesaurus—Rich Release Format (RRF)," 2012, http://www.ncbi.nlm.nih.gov/books/NBK9685/.

[23] H. Davies, G. R. Bignell, C. Cox et al., "Mutations of the BRAF gene in human cancer," *Nature*, vol. 417, no. 6892, pp. 949–954, 2002.

[24] S. C. Abraham, B. Nobukawa, F. M. Giardiello, S. R. Hamilton, and T. T. Wu, "Fundic gland polyps in familial adenomatous polyposis: neoplasms with frequent somatic adenomatous polyposis coli gene alterations," *American Journal of Pathology*, vol. 157, no. 3, pp. 747–754, 2000.

[25] C. E. Lipscomb, "Medical Subject Headings (MeSH)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, pp. 265–266, 2000.

[26] R. R. Richter and T. M. Austin, "Using MeSH (medical subject headings) to enhance PubMed search strategies for evidence-based practice in physical therapy," *Physical Therapy*, vol. 92, no. 1, pp. 124–132, 2012.

[27] T. Morikawa, A. Kuchiba, M. Yamauchi et al., "Association of CTNNB1 (β-Catenin) alterations, body mass index, and physical activity with survival in patients with colorectal cancer," *Journal of the American Medical Association*, vol. 305, no. 16, pp. 1685–1694, 2011.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 909525, 8 pages
http://dx.doi.org/10.1155/2013/909525

Research Article

Molecular Signature of Cancer at Gene Level or Pathway Level? Case Studies of Colorectal Cancer and Prostate Cancer Microarray Data

Jiajia Chen,[1,2] Ying Wang,[1,3] Bairong Shen,[1] and Daqing Zhang[1]

[1] *Center for Systems Biology, Soochow University, Jiangsu, Suzhou 215006, China*
[2] *Department of Chemistry and Biological Engineering, Suzhou University of Science and Technology, Jiangsu, Suzhou 215011, China*
[3] *Laboratory of Gene and Viral Therapy, Eastern Hepatobiliary Surgical Hospital, Second Military Medical University, Shanghai 200438, China*

Correspondence should be addressed to Daqing Zhang; szdaq@126.com

Received 2 November 2012; Accepted 23 December 2012

Academic Editor: Tianhai Tian

With recent advances in microarray technology, there has been a flourish in genome-scale identification of molecular signatures for cancer. However, the differentially expressed genes obtained by different laboratories are highly divergent. The present discrepancy at gene level indicates a need for a novel strategy to obtain more robust signatures for cancer. In this paper we hypothesize that (1) the expression signatures of different cancer microarray datasets are more similar at pathway level than at gene level; (2) the comparability of the cancer molecular mechanisms of different individuals is related to their genetic similarities. In support of the hypotheses, we summarized theoretical and experimental evidences, and conducted case studies on colorectal and prostate cancer microarray datasets. Based on the above assumption, we propose that reliable cancer signatures should be investigated in the context of biological pathways, within a cohort of genetically homogeneous population. It is hoped that the hypotheses can guide future research in cancer mechanism and signature discovery.

1. Introduction

Microarray technology has evolved rapidly in the past several years as a powerful tool for large-scale gene expression profiling [1]. By monitoring changes in gene expression patterns, microarray technology is widely utilized in search of molecular signatures for many medical conditions including cancer. However, evidence is mounting that differentially expressed gene (DEG) lists detected from different studies for the same disease are often inconsistent [2, 3]. One might attribute the inconsistency to the variation in microarray platforms, experimental samples, normalization and analysis methods, and inherent biological uncertainty. Yet this discordance remains even in technical replicate tests using identical samples as in the case of Ein-Dor et al. [4]. Therefore, signature identification at the level of differential genes has been challenged about its robustness and reliability. In light of the inconsistency between DEG lists obtained from different datasets, we propose herein two hypotheses: (1) the expression signatures of different cancer microarray datasets are more similar at pathway level than at gene level; (2) the comparability of the cancer molecular mechanisms of different individuals is related to their genetic similarities. The hypotheses are subsequently verified by case studies of colorectal cancer and prostate cancer microarray datasets, respectively. Hopefully, the hypotheses would explain the inconsistency of the DEG lists derived from multiple experiments and provide novel methods for discovering robust and specific biomarkers of cancer.

2. Materials and Methods

2.1. Data Collection. We collected 5 gene expression profiling datasets on colorectal cancer and 10 datasets on prostate cancer from public gene expression data repositories, for example, Gene Expression Omnibus (GEO), Oncomine

Table 1: Colorectal cancer gene expression datasets used in the meta-analysis.

Dataset	Platform	Total genes	Total samples	Experimental design		Statistical method
				Normal	Tumor	
Hong	HGU133	54675	22	10	12	t-test
Sabates-Bellver	HGU133	54675	64	32	32	Mann-Whitney test
Galamb1	HGU133	54675	30	11	19	SAM
Galamb2	HGU133	54675	38	8	30	PAM
Graudens	cDNA	23232	30	12	18	z-statistics

SAM: significance analysis of microarrays; PAM: prediction analysis of microarrays.

[5] and Supplementary Materials from published literatures. The detailed information of the datasets was summarized in Table 1 for colorectal cancer and Supplementary Table 1 (see Supplementary Material available online at http://dx.doi.org/10.1155/2013/909525) for prostate cancer. These data were collected from two types of platforms, that is, cDNA two-channel arrays and Affymetrix microarray platforms including Human 6800 Affy gene chips, HG-U95A and HG-U133 series. Each dataset was named after the first author of the original literature. Only profiles of normal and cancer tissues were extracted for further analysis.

2.2. Preprocessing of Raw Data. The images of the cDNA array were processed using GenePix Pro 5.0.1.24 software. Background correction was performed by subtracting the median background intensities from the median foreground intensities of all spots in both channels. The raw datasets measured with Affymetrix chips were analysed via MAS5.0 algorithm in R platform. To eliminate the systematic error from heterogeneous datasets before the identification of signatures, we performed Locally Weighted Scatter Plot Smoothing (LOWESS) method for within-chip normalization of cDNA array's dataset and Median Absolute Deviation (MAD) method for between-chip normalization of all datasets. In addition, data was filtered to eliminate bad spots, and the filter criterion was defined as 60% absence across all of the samples. All of the data of preprocessing procedures were performed in R programming environment.

2.3. Determination of the Differentially Expressed Outlier Genes. Cancer Outlier Profile Analysis (COPA) method was performed for detecting genes that were differentially expressed between cancer and normal samples. We used COPA package by MacDonald and Ghosh [6] in R platform. According to the COPA package guidelines, the data was centered and scaled on a rowwise basis using median average difference. The rows of microarray expression data matrix were genes, and the columns were samples. The COPA function calculates a "COPA" score from a set of microarrays. As a preliminary step the function used a percentile for prefiltering the data. The number of outlier samples for each gene was calculated, and all genes with a number of outlier samples less than the percentile (default 95th) were removed from further consideration. A threshold cutoff for "outlier" status was set as 1.7 and applied to all genes.

2.4. Functional Enrichment of Outlier Genes. The significant outlier genes were subsequently mapped to functional databases, for example, GSEA [7], KEGG [8], and GeneGO (GeneGO, Inc.) for the pathway enrichment analysis. GSEA analysis and KEGG pathway analysis were performed using Gene Set Enrichment Analysis (GSEA) tool [7] and Onto-Express [9, 10], respectively. GSEA tool used a collection of gene sets from molecular signatures database (MSigDB), which was divided into five major collections. In our work, we used C2 curated gene sets. Enriched GeneGO pathways were detected by MetaCore (GeneGO, Inc) [11] software. P-value was used to evaluate the statistical significance of each candidate pathway. In MetaCore, the statistics significance (P-value) was calculated by using hypergeometric distribution. False Discovery Rate (FDR) adjustment was applied for multiple test correction.

2.5. Pairwise Overlapping Comparison at Gene/Pathway Level. The overlapping percentage between two datasets is calculated as follows:

$$\text{Overlapping percentage} = \frac{m}{n_1 + n_2 - m} \times 100\%, \quad (1)$$

where n_1 is the number of all the data in dataset 1, n_2 is the number of all the data in dataset 2, and m is the number of overlapping data between two datasets.

3. Results

3.1. Outlier Detection Using Novel Statistic Method. Table 1 listed the statistical methods for identifying differentially expressed genes by the original articles. Most of the prevailing analytical methods like t-test, SAM, and z-statistic considered the average value of gene intensities in the cancer samples. These statistical methods, however, would fail to find "outlier genes" which are only involved in subsets of the cancer samples. Despite their scarcity, outlier genes are nontrivial and may present a hallmark of potential oncogenes. These conventional methods are not suitable for detecting such subset-specific oncogene expression profiles as proposed by Tomlins et al. [12] and Lian [13]. Through applications to public cancer microarray datasets in our previous study [14], we have demonstrated that some newly developed statistics showed superior performance than traditional t-statistics in outlier detection. We herein applied Cancer Outlier Profile Analysis (COPA), a novel significant genes analysis method

Table 2: The number of pathway/gene sets enriched by differentially expressed gene for five colorectal cancer datasets.

Dataset	Number of enriched pathways in GeneGO	Number of enriched gene sets in GSEA
Hong	71	154
Sabates-Bellver	50	303
Galamb1	78	91
Galamb2	36	128
Graudens	149	172

proposed by Tomlins et al. [12], to meta-analyze multiple cancer datasets.

3.2. Signatures Are More Similar at Pathway Level across Multiple Colorectal Cancer Datasets. In order to verify our first hypothesis, we performed meta-analysis of 5 colorectal cancer gene expression profiling datasets from independent laboratories [15–19].

After COPA analysis, we identified 3258 genes differentially expressed between normal colorectal and colorectal tumor samples. The searches in the Entrez PubMed database showed that only 450 out of 3258 (13.8%) identified genes by COPA method were associated with colorectal cancer.

The number of overexpressed genes was obviously discrepant across all groups because of the different samples, arrays, and platforms. To decrease the discrepancy, we tried to understand the cancer molecular mechanism at systems biological level. We then mapped the DEGs identified by COPA using Gene Set Enrichment Analysis (GSEA) and MetaCore software for pathway enrichment analysis, respectively. Totally we found 262 enriched pathways in GeneGO's database with a P value threshold of 0.05; the detailed list of the pathways are provided in Supplementary Table 2. In addition, we performed the gene sets enrichment analysis in GSEA by using C2 curated file, which includes 1892 gene sets/pathway annotation. 111 outlier gene sets with NOM P-value <0.05 and FDR < 0.05 were also found and listed in Supplementary Table 3. The numbers of significant GeneGO pathways or GSEA gene sets enriched by the differentially expressed gene for 5 colorectal cancer datasets were listed in Table 2.

We performed pairwise comparison between 5 datasets in terms of DEGs, GSEA's enriched gene sets, and GeneGO's enriched pathways, respectively. For 5 different datasets, 10 pairs of datasets are available for comparison. Figure 1 showed the pairwise overlapping percentage at different observation levels. A significantly higher overlap at pathway level than at gene level is observed with 70% of the dataset pairs by GeneGO and 60% of the dataset pairs by GSEA. This observation supports our first hypothesis that the overlapping percentage at the pathway level is higher than that at the gene level.

Moreover, we found 4 GeneGO pathways that were shared by 4 datasets. These pathways were considered to be most overlapped and listed in Table 3. Among them, ECM remodeling, chemokines, and adhesion pathways, belonging

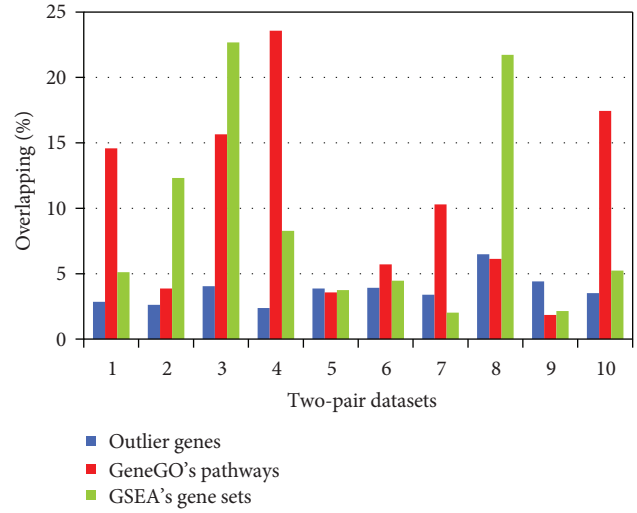


Figure 1: Pairwise overlapping percentage of 5 datasets among differentially expressed genes, enriched gene sets in GSEA, and enriched pathways in GeneGO database. The x-axis represented all the two-pair combination of 5 datasets. The y-axis represented the overlapping percentage.

to cell adhesion category, were previously reported to play a role in colorectal cancer. The other two pathways, integrin outside-in signalling pathway and L-selenoamino acids incorporation in proteins during translation pathway, have not been reported as colorectal cancer associated pathways. The network objects in both of the pathways, however, have been widely reported in colorectal cancer. Integrins are heterodimeric adhesion receptors, and most of them recognize ECM proteins. A major function of integrin signaling is to link ECM proteins to intracellular actin filaments via interactions of integrins with actin-binding proteins. Therefore, the correlation between integrin signaling and ECM pathway may play an active role in colorectal cancer. We infer that these two pathways might be putative novel colorectal cancer related pathways which could provide crucial guidance for biological scientists. Their roles in colorectal cancer need further experimental validation in the future.

We performed paired t-test to decide whether the different overlapping percentages observed between different levels are significant. The P-values for the difference between outlier genes and GeneGO's enriched pathways were 0.01354 by paired t-test and 0.02441 by Wilcoxon test. The P-values for the difference between outlier genes and GSEA gene sets were 0.028 by paired t-test and 0.08 by Wilcoxon test, respectively. The P-values indicate that the overlapping percentages at gene set or pathway level are significantly higher than that at individual gene level. We thus came to the conclusion that the expression signatures of independent datasets at higher functional level are significantly more consistent than that at gene level.

3.3. The Prostate Cancer Outlier Gene Enriched Pathways Show a Regional Distribution Feature. In support of the second hypothesis, we performed a regional analysis of 10

Table 3: The top 4 most overlapped GeneGO's pathways shared by 4 datasets.

GeneGO ontology	Pathway name	Pubmed citation count
Translation	(L)-selenoamino acids incorporation in proteins during translation	0
Cytoskeleton remodeling	Integrin outside-in signaling	0
Cell adhesion	ECM remodeling	64
Cell adhesion	Chemokines and adhesion	1117

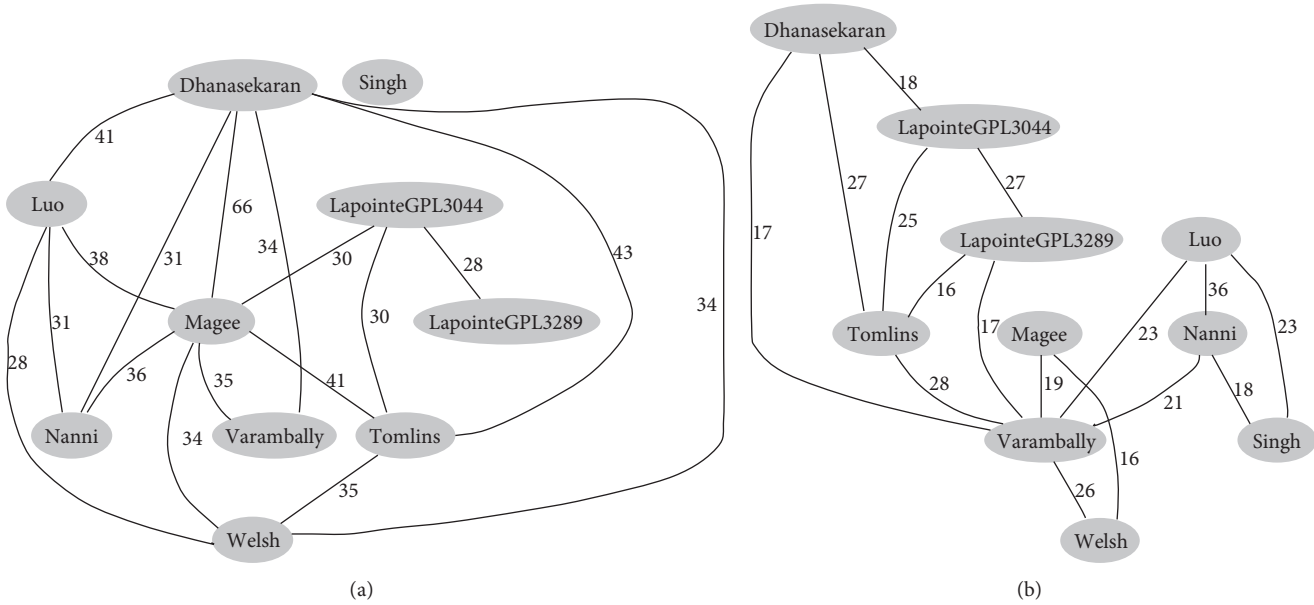


Figure 2: A simple network that associates datasets according to their similarity distances. The distances were calculated based on the overlapping percentage of the enriched pathways identified by (a) GeneGO and (b) KEGG. The lines between two datasets mean that their overlapping is more than two-thirds of the all. Each circle represented a dataset, and the overlapping percentage was shown on the lines.

publicly available prostate cancer gene-expression datasets from different locations [20–28].

We first conducted KEGG and GeneGO pathway enrichment analysis on these datasets, followed by a pairwise comparison of pathway overlapping percentage among them. Only the significantly enriched pathways with previous evidence of prostate cancer association were adopted for the comparison. Text mining was performed to make sure that there was at least one published paper describing the function of these pathways in prostate cancer.

Based on pathway overlapping analysis, we calculated the distance matrices between these datasets and generated a network to display their association. Five common distances, that is, Euclidean distance, Pearson correlational distance, Manhattan distance, Kendall's tau correlational distance, and Hamming distance were used to measure the similarity of these datasets. Based on these distances, a network graph was generated to display the association of these datasets. Figures 2(a) and 2(b) illustrate the association of the datasets based on GeneGO pathways and KEGG pathways, respectively.

Figure 2 revealed an essential regional distribution feature of significant pathways across multiple datasets. It is obvious from the graph that the distance between two Lapointed [29] datasets is the closest among all the datasets. Datasets by Dhanasekaran et al. [20], Tomlins et al. [25], and Magee et al. [23] feature a high pathway overlap which could be reflected by distances, indicating their similarities. The datasets from Singh et al. [26], Luo et al. [22], Welsh et al. [24], and Nanni et al. [27] diverge less from each other than those from the other six datasets.

We then investigated the regional sources of the tissue specimens for each dataset, as listed in Table 4. Samples of Dhanasekaran et al. [20] and Tomlins et al. [25] were obtained from the same place; those of Magee et al. [23] were close to them. Samples of Singh et al. [26], Welsh et al. [24] and Luo et al. [22, 30] came from adjacent states in America. Although the samples of Lapointe et al. [21] were not given a specific location, the author informed us their two experiment datasets were taken from patients of the same population. Apparently, there is obvious concordance between dataset similarity and sample source distribution.

Considering the influence by different microarray platforms, we compared the total unique genes of each dataset in order to testify that the significant pathway distribution feature is caused by different data sources rather than different experimental platforms. As implied in Figure 3, the similarities of the experimental platforms, here the overlapping proportion of the nonredundant probes used in different platforms, are not correlated to the regional distribution. Therefore, the regional distribution of cancer signature at pathway level is independent of the experimental platforms.

Table 4: Tissue specimen sources of each prostate cancer expression dataset.

Datasets	Tissue specimens sources	Locations
Dhanasekaran	University of Michigan Specialized Program of Research Excellence in Prostate Cancer (SPORE) tumor bank	America, Michigan (MI)
Lapointe	Stanford University; Karolinska Institute; Johns Hopkins University	America, California (CA); Sweden, just outside Stockholm; America, Maryland (MD);
Tomlins	University of Michigan	America, Michigan (MI)
Luo	Johns Hopkins Hospital	America, Maryland (MD)
Magee	Washington University School of Medicine; University of Washington Medical Center	America, Missouri (MO); America, Washington (WA);
Welsh	University of Virginia (UVA)	America, Virginia (VA)
Varambally	University of Michigan Prostate Cancer Specialized Program of Research Excellence (SPORE) Tissue Core	America, Michigan (MI)
Singh	Brigham and Women's Hospital	America, Massachusetts (MA)
Nanni	Regina Elena Cancer Institute	Italy, Rome

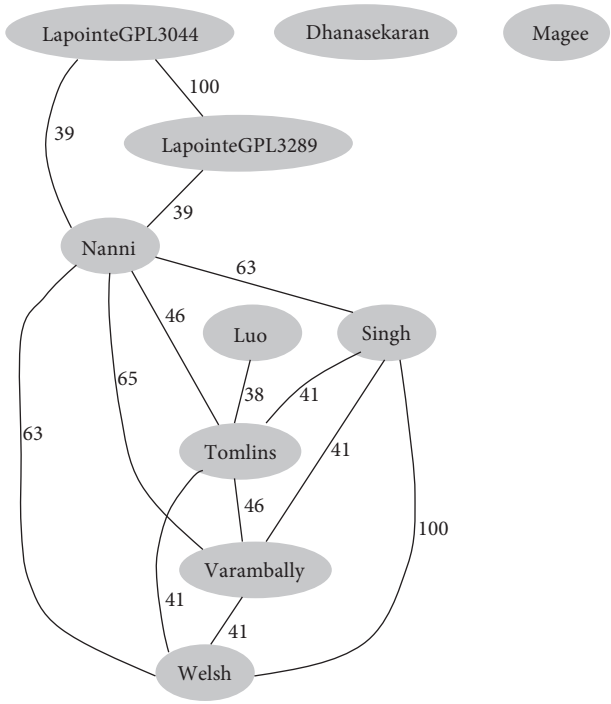


Figure 3: A simple network that associates datasets according to the similarity in microarray platforms. The distances represent the overlapping proportion of the probes used in different platforms.

4. Discussion

4.1. Comparison of DEGs between Different Experiments Revealed Little Overlap. The application of DNA microarrays for the investigation of cancer has led to numerous microarray studies that examined the same clinical conditions. Nevertheless, experiments from different groups have given dissimilar results when DEG lists are directly compared. The disparity was demonstrated in this study, where a meta-analysis of 5 colorectal cancer microarray expression datasets from 4 independent laboratories was performed. We calculated the pairwise overlapping proportion of DEGs between any two datasets, only to find that the overlap between the two lists was disappointingly small (~5%).

Such inconsistency has been observed in gene expression profiling of various types of cancer. For example, in two prominent studies that aimed to predict survival of breast cancer patients [31, 32], both groups claimed to have generated gene lists with predictive power, but only 17 genes appeared on both lists. In another attempt to predict the 5-year metastasis of breast cancer, van't Veer et al. [31] and Wang et al. [33] reported a list of gene sets with good prediction performance, respectively. But the predictive success of their studies was frustrated by the fact that the sets of metastasis-related genes identified by these two independent studies had only 3 overlapping genes. More recently our colleagues [3] meta-analyzed 10 independent microarray datasets associated with prostate cancer, but the resulting set of DEGs had only ~20% overlap between each datasets.

The most straightforward explanation of this lack of agreement is the variation in microarray platforms, experimental samples, normalization, and analysis methods. The open question is, however, whether the inconsistency can be attributed only to these trivial reasons?

To address the issue, Ein-Dor et al. [4] sought to remove all the technical differences mentioned above by analyzing a single breast cancer dataset [31] with a single method. By randomly generating training datasets, they demonstrated that the same analysis could have obtained many equally predictive gene lists and that two such lists share, typically, only a small number of genes. This finding indicates that low consistency occurs even in technical replicate tests using identical samples. The reason for this inconsistency or instability would be that (1) the number of DEGs is large whereas the number of samples is limited; (2) the resulting set of DEGs fluctuates according to the subset of patients used for gene selection.

4.2. Identifying Robust Molecular Signatures at Functional Modules Level or Pathway Level. In this study we evaluated the consistency of signatures across 5 colorectal cancer datasets produced by different platforms. Although the DEG lists selected had only ~5% overlaps, their enriched pathways were still consistent. Consistency analysis at different levels provides solid evidence that cancer signatures at pathway level diminish the discrepancies observed in direct comparisons of DEGs and are more consistent across multiple datasets than at gene level.

As the understanding of tumor biology deepens, it is well recognized that carcinogenesis is characterized with coordinated molecular changes. Functionally correlated genes often display coordinated expression to accomplish their roles; one would therefore expect that the inconsistent DEG lists across independent experiments are functionally more consistent. In other words, the discrepancies of DEGs would be less pronounced when they are mapped to functional groups or biological pathways.

Following this line, some previous studies have shifted their focus from individual genes to the biologically related groups of genes in the analysis of cancer microarray data. For example, in order to investigate the robustness of biological themes, Hosack et al. [34] applied the Expression Analysis Systematic Explorer (EASE) to determine the biological theme for DEG lists generated by various gene selection methods. Their research provided strong evidence that biological themes are stable to varying methods of gene selection. Zhu et al. [35] developed a novel tool for identifying cancer signatures at functional modules level. Its applications to two cancer types demonstrated that the functional modules enjoy explicit relevance to cancer biology. Recently, Yang et al. [36] proposed semantic similarity measure for DEG lists detected under varied statistical thresholds and from different studies. They reported that gene lists could be functionally consistent according to their semantic similarity. In addition, Gorlov et al. [37] conducted functional annotation analysis of the prostate cancer genes identified by two different methods. They observed a considerable overlap between biological functions identified by varied methods.

In recent years, pathway analysis has received a great deal of attention in the study of cancer microarray data [7, 34]. Pathway analysis typically correlates the identified DEGs with predefined pathway databases. It is reported that pathway analysis applied to differential gene lists detected under varied statistical methods yielded common results [38]. This discovery was validated in our previous study by Wang et al. [3], who evaluated the consistency of signature across 10 prostate cancer datasets produced by different platforms. Although the datasets share disappointingly few DEGs, their DEG-enriched pathways were still consistent.

4.3. Searching for Common Signatures among a Cohort of Genetic Homogeneous Population. As for the second hypothesis we assume that the individuals bearing similar genetic/environmental factors tend to share more common pathways. However, the information on the genetic/environmental characteristics of the patient samples is generally lacking. We believe it should be statistically reasonable to take the geographical location of the sample resources as the measurement of the similarities of their genetic/environmental factors. According to the similarity of outlier enriched pathways found by GeneGO and KEGG, we are able to classify 10 different prostate cancer related datasets into several groups. The datasets from same or adjacent geographical locations tend to reside within the same group. In other words, we observed an essential regional distribution feature of significant pathways across multiple datasets. In this sense molecular signatures from the geographically adjacent tissue specimens would be more consistent than those generated from geographically isolated samples. This observation is basically in accordance with our hypothesis that the comparability of the cancer molecular mechanisms of different individuals is related to their genetic similarities.

Cancer represents a heterogeneous disease, which reflects the interaction of a myriad of etiological and genetic contributions [39]. Therefore the gene expression profiles of cancer patients are diverse, depending on factors such as genetic information, environment effect, and personal behaviors. The role of genetic and environmental factors in modulating gene expression variation in humans has been extensively investigated. Most of the previous studies on cancer microarray profiling, however, ignored the interindividual variation in gene expression. It is likely that differences in expression that appear to be related with the disease may in fact represent random genetic variation. This situation will further introduce false discoveries and reduce the overall reproducibility of DEG detection. This concern was mentioned by Michiels et al. [40], who investigated the stability of seven published datasets to predict prognosis of cancer patients. It was observed that the predictive gene lists reported by the various groups were highly unstable and depended strongly on the subset of samples chosen for training.

It is assessed that, to achieve a typical overlap of 50% between two predictive lists of genes, the expression profiles of several thousands of patients would be needed [41]. Unfortunately, obtaining such a large number of samples is currently impractical due to limited tissue availability and financial constraints. A more practical approach would be to search for common signatures among a genetically homogeneous human population other than those among a mixed population. Although different individuals may have different regulatory mechanisms and discrepant cancer associated pathways, we assume that the individuals bearing similar genetic and environmental factors tend to share more common pathways.

Thus it would be reasonable to group patients into well-defined small subgroups on the basis of each person's unique genetic and environmental information. In this way, the individual difference of cancer mechanism is accounted when we analyze cancer expression data from different resources. This kind of investigation will help to find population-specific cancer pathways and facilitate personalized medicine.

5. Conclusions

Based on previous observations, we proposed herein two novel points of view for the cancer signatures identification. The pathway-based approach suggested in this paper would hopefully improve the comparability of different microarray datasets and, therefore, may lead to more valid and reliable biological interpretation of microarray results. Moreover, the generation of the population-specific cancer signatures would help to deliver effective therapy to patients most likely to benefit from such treatment and enable "personalized medicine." With increasing amount of cancer datasets available, the challenge in the future is to collect more cancer datasets from independent populations to prove our hypotheses.

Conflict of Interests

The authors declare they have no direct financial relation with the trademarks mentioned in this paper that might lead to a conflict of interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (91230117, 31170795) Grants, the Specialized Research Fund for the Doctoral Program of Higher Education of China (20113201110015), the International S&T Cooperation Program of Suzhou (SH201120), and the National High Technology Research and Development Program of China (863 program, Grant No. 2012AA02A601).

References

[1] J. Dopazo, E. Zanders, I. Dragoni, G. Amphlett, and F. Falciani, "Methods and approaches in the analysis of gene expression data," *Journal of Immunological Methods*, vol. 250, no. 1-2, pp. 93–112, 2001.

[2] M. Zhang, C. Yao, Z. Guo et al., "Apparently low reproducibility of true differential expression discoveries in microarray studies," *Bioinformatics*, vol. 24, no. 18, pp. 2057–2063, 2008.

[3] Y. Wang, J. Chen, Q. Li et al., "Identifying novel prostate cancer associated pathways based on integrative microarray data analysis," *Computational Biology and Chemistry*, vol. 35, no. 3, pp. 151–158, 2011.

[4] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: is there a unique set?" *Bioinformatics*, vol. 21, no. 2, pp. 171–178, 2005.

[5] D. R. Rhodes, S. Kalyana-Sundaram, V. Mahavisno et al., "Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles," *Neoplasia*, vol. 9, no. 2, pp. 166–180, 2007.

[6] J. W. MacDonald and D. Ghosh, "COPA—cancer outlier profile analysis," *Bioinformatics*, vol. 22, no. 23, pp. 2950–2951, 2006.

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

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

[9] S. Drăghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz, "Global functional profiling of gene expression," *Genomics*, vol. 81, no. 2, pp. 98–104, 2003.

[10] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nature Genetics*, vol. 22, no. 3, pp. 281–285, 1999.

[11] S. Ekins, A. Bugrim, L. Brovold et al., "Algorithms for network analysis in systems-ADME/Tox using the MetaCore and MetaDrug platforms," *Xenobiotica*, vol. 36, no. 10-11, pp. 877–901, 2006.

[12] S. A. Tomlins, D. R. Rhodes, S. Perner et al., "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer," *Science*, vol. 310, no. 5748, pp. 644–648, 2005.

[13] H. Lian, "MOST: detecting cancer differential gene expression," *Biostatistics*, vol. 9, no. 3, pp. 411–418, 2008.

[14] Y. Tang, J. Chen, C. Luo, A. Kaipia, and B. Shen, "MicroRNA expression analysis reveals significant biological pathways in human prostate cancer," in *Proceedings of the 5th IEEE International Conference on Systems Biology (ISB '11)*, pp. 203–210, IEEE Computer Society, Zhuhai, China, September 2011.

[15] E. Graudens, V. Boulanger, C. Mollard et al., "Deciphering cellular states of innate tumor drug responses," *Genome Biology*, vol. 7, no. 3, article R19, 2006.

[16] O. Galamb, B. Györffy, F. Sipos et al., "Inflammation, adenoma and cancer: objective classification of colon biopsy specimens with gene expression signature," *Disease Markers*, vol. 25, no. 1, pp. 1–16, 2008.

[17] O. Galamb, F. Sipos, N. Solymosi et al., "Diagnostic mRNA expression patterns of inflamed, benign, and malignant colorectal biopsy specimen and their correlation with peripheral blood results," *Cancer Epidemiology Biomarkers and Prevention*, vol. 17, no. 10, pp. 2835–2845, 2008.

[18] Y. Hong, K. S. Ho, K. W. Eu, and P. Y. Cheah, "A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis," *Clinical Cancer Research*, vol. 13, no. 4, pp. 1107–1114, 2007.

[19] J. Sabates-Bellver, L. G. Van der Flier, M. de Palo et al., "Transcriptome profile of human colorectal adenomas," *Molecular Cancer Research*, vol. 5, no. 12, pp. 1263–1275, 2007.

[20] S. M. Dhanasekaran, T. R. Barrette, D. Ghosh et al., "Delineation of prognostic biomarkers in prostate cancer," *Nature*, vol. 412, no. 6849, pp. 822–826, 2001.

[21] J. Lapointe, C. Li, J. P. Higgins et al., "Gene expression profiling identifies clinically relevant subtypes of prostate cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 3, pp. 811–816, 2004.

[22] J. Luo, D. J. Duggan, Y. Chen et al., "Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling," *Cancer Research*, vol. 61, no. 12, pp. 4683–4688, 2001.

[23] J. A. Magee, T. Araki, S. Patil et al., "Expression profiling reveals hepsin overexpression in prostate cancer," *Cancer Research*, vol. 61, no. 15, pp. 5692–5696, 2001.

[24] J. B. Welsh, L. M. Sapinoso, A. I. Su et al., "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer," *Cancer Research*, vol. 61, no. 16, pp. 5974–5978, 2001.

[25] S. A. Tomlins, R. Mehra, D. R. Rhodes et al., "Integrative molecular concept modeling of prostate cancer progression," *Nature Genetics*, vol. 39, no. 1, pp. 41–51, 2007.

[26] D. Singh, P. G. Febbo, K. Ross et al., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[27] S. Nanni, C. Priolo, A. Grasselli et al., "Epithelial-restricted gene profile of primary cultures from human prostate tumors: a molecular approach to predict clinical behavior of prostate cancer," *Molecular Cancer Research*, vol. 4, no. 2, pp. 79–92, 2006.

[28] S. Varambally, J. Yu, B. Laxman et al., "Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression," *Cancer Cell*, vol. 8, no. 5, pp. 393–406, 2005.

[29] J. Lapointe, C. Li, J. P. Higgins et al., "Gene expression profiling identifies clinically relevant subtypes of prostate cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 3, pp. 811–816, 2004.

[30] J. Luo, D. J. Duggan, Y. Chen et al., "Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling," *Cancer Research*, vol. 61, no. 12, pp. 4683–4688, 2001.

[31] L. J. van't Veer, H. Dai, M. J. Van de Vijver et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[32] T. Sørlie, C. M. Perou, R. Tibshirani et al., "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 19, pp. 10869–10874, 2001.

[33] Y. Wang, J. G. M. Klijn, Y. Zhang et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[34] D. A. Hosack, G. Dennis Jr., B. T. Sherman, H. C. Lane, and R. A. Lempicki, "Identifying biological themes within lists of genes with EASE," *Genome Biology*, vol. 4, no. 10, article R70, 2003.

[35] J. Zhu, J. Wang, Z. Guo et al., "GO-2D: identifying 2-dimensional cellular-localized functional modules in Gene Ontology," *BMC Genomics*, vol. 8, article 30, 2007.

[36] D. Yang, Y. Li, H. Xiao et al., "Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories," *Bioinformatics*, vol. 24, no. 2, pp. 265–271, 2008.

[37] I. P. Gorlov, G. E. Gallick, O. Y. Gorlova, C. Amos, and C. J. Logothetis, "GWAS meets microarray: are the results of genome-wide association studies and gene-expression profiling consistent? Prostate cancer as an example," *PLoS One*, vol. 4, no. 8, Article ID e6511, 2009.

[38] T. Manoli, N. Gretz, H. J. Gröne, M. Kenzelmann, R. Eils, and B. Brors, "Group testing for pathway analysis improves comparability of different microarray datasets," *Bioinformatics*, vol. 22, no. 20, pp. 2500–2506, 2006.

[39] J. Chen, Y. Wang, D. Guo, and B. Shen, "A systems biology perspective on rational design of peptide vaccine against virus infections," *Current Topics in Medicinal Chemistry*, vol. 12, no. 12, pp. 1310–1319, 2012.

[40] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488–492, 2005.

[41] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 15, pp. 5923–5928, 2006.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 197892, 8 pages
http://dx.doi.org/10.1155/2013/197892

Research Article

A Modified Amino Acid Network Model Contains Similar and Dissimilar Weight

Xiong Jiao,[1] Lifeng Yang,[2] Meiwen An,[1] and Weiyi Chen[1]

[1] *Institute of Applied Mechanics and Biomedical Engineering, Taiyuan University of Technology, Taiyuan 030024, China*
[2] *College of Computer Science and Technology (College of Software), Taiyuan University of Technology, Taiyuan 030024, China*

Correspondence should be addressed to Xiong Jiao; jiaox.china@gmail.com

Received 7 November 2012; Revised 22 December 2012; Accepted 23 December 2012

Academic Editor: Guang Hu

For a more detailed description of the interaction between residues, this paper proposes an amino acid network model, which contains two types of weight—similar weight and dissimilar weight. The weight of the link is based on a self-consistent statistical contact potential between different types of amino acids. In this model, we can get a more reasonable representation of the distance between residues. Furthermore, with the network parameter, average shortest path length, we can get a more accurate reflection of the molecular size. This amino acid network is a "small-world" network, and the network parameter is sensitive to the conformation change of protein. For some disease-related proteins, the highly central residues of the amino acid network are highly correlated with the hot spots. In the compound with the related drug, these residues either interacted directly with the drug or with the residue which is in contact with the drug.

1. Introduction

In living cells, proteins are very important molecules, and they participate in almost all of the cell functions. During these biological activities, the structure of some proteins shows an obviously conformational flexibility. For a correct and fast implementation of the biological functions through the conformation change, there needs a motor coordination for the residues in different parts of the protein. In this process, a fast communication mechanism is vital for the information sharing between residues about these concerted actions. In fact, this information exchange is achieved through the interaction between residues. But when we put all these residues and the interactions between them together, the protein becomes a very complicated system.

On the other hand, from the viewpoint of complex network [1, 2], a protein molecule can be treated as a complex network. In this network, each residue can be simplified as a node, and the interaction between different residues is treated as the link. With this useful tool—complex network, some new research ideas and methods are applied to the study of the structure-function relationship, and some phenomenon can be explained through the analyzing of this network. Such related work as the identification of the "key residues" through the network parameter—betweenness [3]. In the measuring process of the topology of the protein contact network, the result shows that the kinetic ability for folding is determined by the topological properties of the protein conformation [4]. Through the biological networks, the rigidity and flexibility of protein structure can be analyzed. Furthermore, with this approach, the cytoskeletal tensegrity can be discussed [5]. The network model also has been wildly used in the drug design and drug discovery [6].

In the amino acid network, each residue is simplified to a single point, and this point is used as the network node. Generally, the carbon alpha is selected as the network node. In some other network models, a point between the carbon alpha and the carbon beta is used to as the network node. The links between these nodes are determined by the distance between them. If the distance between two nodes is less than a cut-off value, then there will exist a link between these two nodes. This cut-off is usually set to 7.0 angstrom [7] or set to 8.5 angstrom [3].

There is another type of amino acid network model. In this model, each residue is also simplified to a node. But the link between two nodes is based on the atom contact between these two residues. A cut-off value—4.5 angstrom [8], or 5.0 angstrom [9], is used as a criterion for the contacts between atoms. If there is an atom contacts between two residues, these two nodes will be connected by a link. For different amino acid network models, the criterion to dictate residue contacts has been reviewed and analyzed [10]. In this paper, the Miyazawa-Jernigan potential is used to construct the link weight, so the side chain center is used to represent the node, and the cut-off value used by Miyazawa and Jernigan is also used in this work [11, 12].

In the weighted amino acid network, in which the link is based on a contact between different residues, the weight of the link can be drawn from the contact probability between different residues [3], or the weight can be drawn from a statistical residue contact potential [11–13]. With the contact potential as the link weight, a weight elastic network model is used to calculate the protein structure dynamics [13]. For the network model based on atom contact, the weight of the link can be deduced from the number of atom contacts between nodes. Furthermore, when the diversity of amino acids is taken into account, these weights can be modified by a normalization factor [8].

For the weight of the link, it can be classified into two types. One is the similar weight and the other is the dissimilar weight [14]. For the similar weight, the value describes the similarity between two nodes. A higher value means that the two nodes are more similar, and the distance between them will be shorter. As for the dissimilar weight, a higher weight value, corresponding to a longer distance between the two nodes, means that the difference between these two nodes are more distinct.

For the weighted amino acid network, the related research work is underway, and many questions needed to be explored, such as which parameter should be selected as the weight and how to assign the weight to the link with a more reasonable mode. In our previous work, we proposed a weight amino acid model [15], but only one type of weight—similar weight is used in the previous model, so we cannot get a more detailed description of the interaction situation between residues.

This paper will modify the previous model with two types of weight, and the weight used in this paper is based on a self-consistent statistical contact energy between residues [12]. In this paper, firstly, the construction methods of the weighted network are compared. Then, for the 197 proteins with low homology, the weighted amino acid networks are constructed and the statistic characteristics of the parameters of these networks are studied, including the average clustering coefficient (C) and the average shortest path length (L). Thirdly, with this weighted network, in order to get a relation between the change of the network parameter and the change of the protein conformation, we studied the changes of the average shortest path length for the small protein CI2 on its high temperature unfolding pathway. The last, take the FKBP-FK506 as an example, we show the application of amino acid network in the drug design.

2. Theory and Method

In this weighted amino acid network, for each amino acid, the geometrical center of the side chain is chosen to represent the network node. The link between a pair of nodes is determined by the distance between them. If the distance between residues i and j (marked with r_{ij}) is less than the cut-off (r_c), there will be a link between them. In this paper, the cut-off is 6.5 angstrom. Thereby, the adjacency matrix element of the unweighted amino acid network can be expressed as follows:

$$a_{ij} = \begin{cases} 1, & i \neq j,\ r_{ij} < r_c, \\ 0, & i = j \text{ or } r_{ij} \geq r_c. \end{cases} \tag{1}$$

Based on the contact potential between residues, the weighted network can be constructed. In the previous model, we use another set of contact potential. All the items of the contact potential are less than zero, and the calculation of the repulsive interaction between residues is very complex.

In this work, we adopt a self-consistent interresidue contact potential to construct the weight of the link. In this contact potential, if two residues are attracted in most cases, the potential between them will get a negative value, and if they are repulsed generally, the potential will be a positive value. With this contact potential, the adjacency matrix element of the weighted amino acid network can be expressed as

$$a_{ij}^{w} = \begin{cases} a_{ij} w_{ij}, & j \neq i \pm 1, \\ 0, & j = i \pm 1. \end{cases} \tag{2}$$

In this definition, we take the contact potential between residues i and j as the link weight, marked as w_{ij}. The value of w_{ij} is related to the types of the residues i and j. For the covalent bond between residues i and $i \pm 1$, the link weight is assumed as zero.

In this amino acid network, if the two nodes are attracted, the potential between them is a negative real number, so, the link between them will get a negative weight. If the attraction between these two nodes become stronger, the absolute value of the weight will become a bigger one. Then, the negative weight can be treated as a similar weight. For the same reason, if the two nodes are repulsed, the potential between them corresponds to a positive real number, and the link between them will get a positive weight. When the repulsion between these two nodes become stronger, the link will get a bigger positive weight value. So, the positive weight can be treated as a dissimilar weight.

Thus, based on the weighted adjacency matrix, the distance matrix can be constructed and the definition of its element can be written as follows. We labeled this definition as definition 1:

$$d_{ij} = \begin{cases} 0, & i = j, \\ \infty, & a_{ij} = 0,\ i \neq j, \\ \dfrac{1}{\left(1 - w_{ij}\right)}, & w_{ij} < 0,\ a_{ij} = 1, \\ 1, & w_{ij} = 0,\ a_{ij} = 1, \\ 1 + w_{ij}, & w_{ij} > 0,\ a_{ij} = 1. \end{cases} \tag{3}$$

When the interaction between two residues is an attractive interaction, the corresponding link weight is a similar one. In this distance definition, a reciprocal function of the weight is used to represent the distance between a pair of attracting nodes. For a stronger attractive interaction between residues, the actual distance between them is shorter than others. And because the weight for attraction is negative, a bigger absolute value corresponds to a shorter distance, as defined in the distance matrix.

At the same time, if the interaction between residues is repulsive, the corresponding link weight is a dissimilar one. The distance definition between them is a linear combination function of the weight. A stronger repulsive interaction, corresponding to a longer actual distance between them, will get a bigger distance value from the distance matrix.

The network model used in this work is an undirected model, and the link is just to represent the existence of the interaction between these two residues. The status of the two ends of a link is equal. So, for the weighted network and the unweighted one, the adjacency matrixes are all symmetric matrixes. In the distance matrix, the similar weight and the dissimilar weight are coexistent in the same distance matrix, and the distance matrix is also a symmetric one.

For a comparison between different definitions, if we do not make a difference between the similar weight and the dissimilar one, and just the dissimilar weight is used in this model, the distance matrix can be defined as below. We labeled it as definition 2:

$$d_{ij} = \begin{cases} 0, & i = j, \\ \infty, & a_{ij} = 0,\ i \neq j, \\ 1 + \dfrac{w_{ij}}{2.19}, & a_{ij} = 1. \end{cases} \tag{4}$$

On the other hand, we can convert the dissimilar weight to the similar weight. The distance between nodes can be defined as below, and it is labeled as definition 3:

$$d_{ij} = \begin{cases} 0, & i = j, \\ \infty, & a_{ij} = 0,\ i \neq j, \\ \dfrac{1}{\left(1 - w_{ij}\right)}, & a_{ij} = 1. \end{cases} \tag{5}$$

Additionally, a new network parameter—strength—is introduced into the weighted amino acid network. The strength of node i can be written as [16, 17]

$$S_i = \sum_{j=1}^{N} \left| a_{ij}^{w} \right|, \tag{6}$$

where N is the number of network nodes and a_{ij}^{w} is an element of the weighted adjacency matrix.

The clustering coefficient of the weighted network can be calculated using the next expression [16, 17]:

$$C_i = \frac{1}{S_i \left(K_i - 1\right)} \sum_{j,h} a_{ij} a_{ih} a_{jh} \frac{\left|w_{ij}\right| + \left|w_{ih}\right|}{2}, \tag{7}$$

where S_i is the strength of the node i, and K_i is its degree. The means of a_{ij} and w_{ij} are the same as that of the expression (2).

The betweenness of node u can be defined as below [18]:

$$B_u = \sum_{i,j} \frac{\sum_{l \in S_{ij}} \delta_l^u}{\left|S_{ij}\right|}. \tag{8}$$

The denominator is the number of shortest paths between i and j, and the numerator is the number of shortest paths between i and j through node u. Betweenness is a useful measure of the node's importance to the network. In order to reflect the significance of betweenness for different nodes, the Z-score is introduced, and the definition of Z-score for B_u is as follows [19]:

$$Z_u = \frac{B_u - \overline{B}}{\sigma}, \tag{9}$$

where B_u is the betweenness of residue u, $\overline{B}$ is the average value of the betweenness of all protein residues, and σ is the standard deviation of these betweenness values.

3. Results and Discussion

3.1. Comparison between Different Definitions of the Distance. For the contract potential used to construct the weighted network in this paper, the value ranges from −1.19 to 0.76. The corresponding distance for varying weights, get from the three different definitions of distance matrix, is shown in Figure 1(a). From this figure, we can see that, when the interaction between residues is a repulsive interaction, if the link weight is a similar weight, the distance got from the distance definition 3 will increase sharply. But based on common sense, it is unreasonable.

On the other hand, in the statistical calculation process of this self-consistent statistical contact potential between different types of amino acids, the cut-off is 6.5 angstrom, and this cut-off is still being used in the contact definition between residues in this paper. So, the distance between a pair of network nodes should be less than 6.5 angstrom. In a statistic calculation process of the actual distance between nodes, the result shows that this actual distance ranges from 3.88 to 6.5 angstrom. The ration of the maximum with the minimum is about 1.7. In the definition 3, due to the sharply increasing of the distance, this ratio is about 9. But for definition 1 and 2, this ration is about 3. So, it can be concluded that in the definition 3, it is not a reasonable assumption that the positive weight be treated as a similar weight.

In our previous work, there is only one type of weight—similar weight. This definition should be revised as follows: a link with a positive weight should be assigned a dissimilar weight, as the rule of definition 2.

At the same time, in the statistic calculation process of the actual distance between nodes, as mentioned above, the result shows that the a great majority of the distances is about 5 angstrom, and most of the interactions between these nodes are an attractive one. So, the middle part of

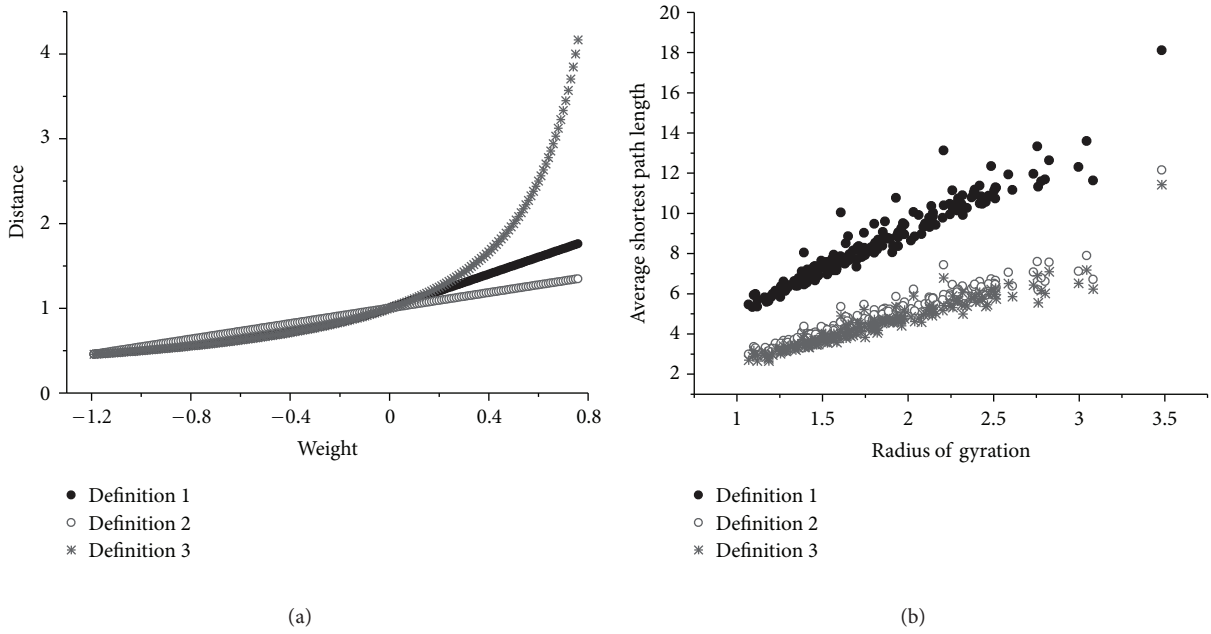


Figure 1: The comparison between the three definitions of the distance matrix. (a) The relation between the weight and the distance. (b) The comparison of the correlation between the average shortest path length and the radius gyration.

the weight-distance curve should be a nearly horizontal line. For the negative weight, the curve of definition 3 is more horizontal than that of definition 2. This phenomenon shows that when the link weight is a negative value, the similar weight assumption is more suitable to reflect the truth.

Based on the above discussion, we can see that the similar weight assumption is reasonable for a negative weight, and the dissimilar weight assumption is suitable for a positive value. Put all these together, we can get definition 1, and the following calculation of distance will use the definition showed in (3).

With a set of 197 proteins selected from the Protein Data Bank (PDB), the weighted amino acid networks are constructed. These proteins include the four structure types: α, β, $\alpha+\beta$, and $\alpha-\beta$. The resolution of these selected proteins is better than 1.8 Å and the sequence identity is less than 20%. The sizes of proteins vary from 51 to 779 residues. The distance matrix is calculated with definition 1.

Radius of gyration is a useful parameter to indicate the size of a molecule. With the network model, the average shortest path length can also be used as an indicator of the molecular size. For the data set, we calculate the radius of gyration for each protein with GROMACS [20]. At the same time, we can get the average shortest path length from the weighted amino acid network. The relation between the average shortest path length with the radius of gyration is shown in Figure 1(b). The correlation coefficient for the path length from definition 1 with the radius of gyration is 0.96. This correlation coefficient is 0.95 for definition 2 and 0.79 for definition 3. Definition 1 gets the best result.

3.2. The Small-World Characteristic of the Amino Acid Network. The "small-world" property is a very important character for complex networks, and the "small-world network" is ubiquitous in the real life, such as the neural networks [21, 22] and the gene network [23, 24]. A vivid example of the "small-world network" is the "six degrees separation" [21, 25]. In a small-world network, most nodes are not connected directly by a link. But due to the short-cut between nodes, most nodes can be reached from every other through a small number of steps. With the increasing of the nodes number, the shortest-path distance between nodes grows sufficiently slowly, and it can be expressed as a function of the logarithm of the number of nodes in the network.

For a complex network and a random network, if they have the same node numbers and the same link numbers, when some condition be satisfied, the complex network can be thought that it holds the "small-world" property. These conditions include two items, the first one is that the average clustering coefficient C of the complex network is far more than that of the random network, and the second condition is that the average shortest path length L is about equal to that of the random network. These conditions can be showed as the following expression [21]:

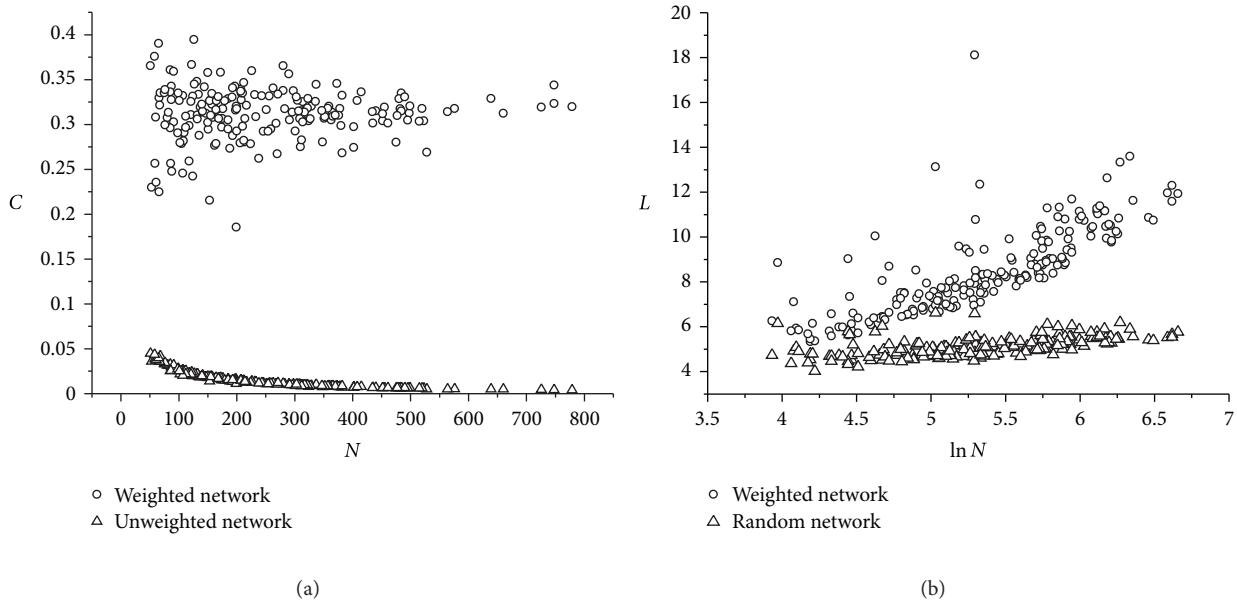
$$C \gg C_r, \qquad L \geq L_r. \tag{10}$$

In this inequality, C_r and L_r are the network parameter of the random network. C_r is the average clustering coefficient and L_r is the average shortest path length. C_r and L_r can be calculated with the following expressions [21]:

$$C_r \approx \frac{\langle K \rangle}{N}, \qquad L_r \approx \frac{\ln N}{\ln \langle K \rangle}. \tag{11}$$

In this expression, N is the node number and $\langle K \rangle$ is the average degree of the random network.

In the "small-world" network, most of the nodes can be reached fast from every other through the "short-cut"



(a) (b)

Figure 2: For the 197 proteins and the corresponding random networks with the same size, the comparison of network parameter. (a) The clustering coefficient of the weighted amino acid network and that of the random network with the same size; (b) the average shortest path length of the weighted amino acid network and that of the random network with the same size.

between residues. So, the average clustering coefficient of the network will get a relatively large value, and the average shortest path length (also be called: characteristic path length) will keep as small as that of a random network.

For the 197 proteins, we constructed the weighted network and calculated the average clustering coefficients and the average shortest path lengths with the distance matrix definition 1. Figures 2(a) and 2(b) showed these results. At the same time, for the random networks with the same size, these two parameters are calculated and the results are also shown in Figures 2(a) and 2(b). From these two figures, we can see that the weighted amino acid networks, contain similar and dissimilar weight for the link, present an obvious "small-world" property. From other works, we have known that the amino acid network is a "small-world" network, so, these results prove that the distinction introduced between similar and dissimilar weights is reasonable, and the construction method of the weighted network also is rational.

In the amino acid network, very few residues can get a high degree value. They usually lie in the core of the globular protein and act as the hubs of the networks [8, 26]. There are more interactions between these hub residues with other residues, so these hub residues play a vital role to the stability of whole protein structures [7, 8, 27]. In some other work, in order to embody the influence of the local environment, the distribution of residue clusters has been analyzed, and the outcome is a log-normal distribution [28].

3.3. The Change of Average Shortest Path Length with the Conformation Change. For exploring the changes of network parameters with the changes of the protein conformations, the protein CI2 (PDB code: 3CI2) was selected as a research object.

With the MD program GROMACS 3.3 [20], the molecular dynamic (MD) simulation was performed at 498 K for 11.2 ns. The force field parameters used in this simulation were taken from GROMOS96 43a1 and the SPC/E water model was used. After the simulation, this protein will become unfolded, and most secondary structures will be depolymerized. However, the protein still keeps a random coil state. With this MD trajectory data, we extract the structures with an interval of 100 ps and then construct the weighted amino acid networks. On this unfolding pathway, along with the conformational changes, the change rule of the average shortest path length (short as: L) was analyzed. This change of L is used to represent the conformation change, and the results are shown in Figure 3.

On the unfolding pathway, when the structure becomes looser, the average shortest path lengths of the weighted amino acid network become longer. Under a high temperature, with the unfolding of the protein, the hydrophobic core will be destroyed. In this process, the hydrophobic-hydrophobic link, which is important to the stability of the protein structures, will be broken. These hydrophobic-hydrophobic links all have a negative weight, and the distance of these links is less than 1. Therefore, while the hydrophobic core derogates, the shortest path length will rise more obviously. From Figure 3, we can see that the average shortest path length from definition 1 is more sensitive to the conformation change than that of the other two definitions.

3.4. The Application of the Amino Acid Network in Drug Design. In the process of drug action, many drugs take the related protein as their target. The structure and the dynamic of this target protein hold a very important role to the

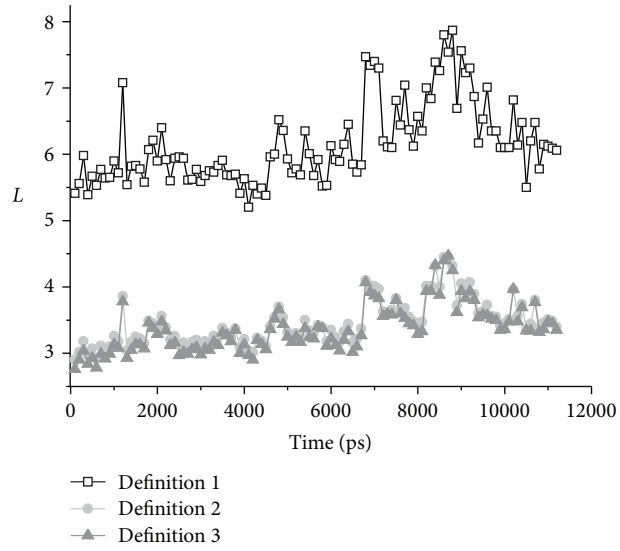


Figure 3: For the three definitions, the average shortest path length of the weighted amino acid networks of protein CI2 on the unfolding pathway.

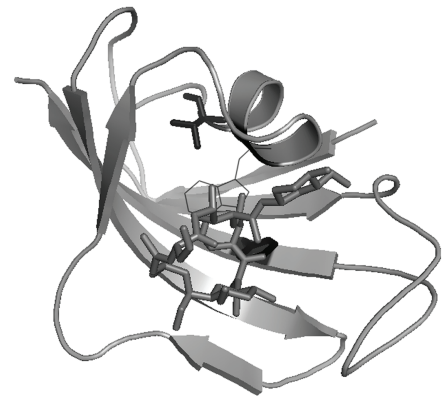


Figure 4: The structure of FKBP-FK506 complex (PDB ID: 1FKF). In this figure, the FK506 is shown with a stick model, and the FKBP is shown with a cartoon model. The Val^{63} and Phe^{99} are shown with stick model in black color. The Trp^{59} is shown with lines model.

therapeutic effect of the drug. The residues located at the binding sites are crucial to the binding and the stability of the complex. These residues often are tightly packed and can provide a major part of the decrease of the binding free energy. They are often called as hot spots, and the central nodes of the amino acid network usually can be predicted as the hot spots [19, 29, 30]. With the support vector machine technology, a model is proposed for the prediction of the binding sites of heme protein [29]. This model contains three types of information: the first is the sequence information, the second is the geometry information of the structure, and the last one is based on some amino acid network parameter. Some scoring function based on the amino acid network also has been proposed for the protein docking [31–33]. Here, we take the immunosuppressant drug (FK506) binding protein—FKBP [34] as an example to show the application of amino acid network in the drug design.

FKBP, or FK506 binding protein (PDB ID: 1FKF), is a immunophilins protein, which is involved in the immune response pathway and is used as a target for the immunosuppressant drug (FK506). Through the binding of FKBP with FK506, the signal transduction in T cells will be blocked, and then the normal immune system reaction will be interfered [35, 36].

Figure 4 shows the structure of the complex of FKBP with FK506.

With the structure, we can draw the detailed information about the complex that the binding sites contain which parts of the drug and which parts of the protein. We can find, as the structure showed above, that the α helix and the β sheet of the FKBP form a cavity, and the FK506 is binding with FKBP in this shallow cavity. For this structure, we construct its amino acid network and then calculate the related network parameter (betweenness) with the corresponding Z-score. In this work, only the Z-score value, which is greater than or equal to 3.0, is considered as a significant one, and the corresponding node will be discussed in the following parts [19]. For 1FKF, the calculating results show that there only two nodes get a higher Z-score value: Val^{63} and Phe^{99}. At the same time, the contacts between the FKBP and FK506 are calculated. For FK506 holds a bigger volume than a residue, so, we use the atom contact between FK506 and FKBP. The Phe^{99} has ten atom contacts with the FK506, and these contacts are mainly due to the side chain of Phe^{99}, which participates in assembling of the binding cavity with other residues. For Val^{63}, although there is no direct interaction with FK506, it has nine atom contacts with Trp^{59}, and Trp^{59} is interacted with FK506 through 20 atom contacts. The nodes with high Z-score value, for 1FKF, are either corresponding to the hot spot or to the residue which has a direct interaction with the ligand [19].

We also take the complex of fkbp12 with rapamycin (PDB ID: 1C9H [37] and 1FKB [38]) to calculate the Z-score value for every node of the amino acid network and to determine the contacts between the drugs with FKBP. The results also show that the node with high Z-score value either interacted directly with the drug or with nodes which is contacted directly with the drug. For these three proteins, the region from Phe^{99} to Val^{101} all contain a binding site with the drug. One is the Phe^{99} for 1FKF and 1FKB, and Val^{101} for 1C9H. On the other hand, when FK506 is binding to FKBP, we can find that the change of FKBP's structure is undersized, but the structural change of FK506 is large. So, we can deduce that the binding sites of FKBP with the related drug are spatial conserved. This useful information is helpful for the design of some new drugs, which has a better curative effect or less toxic than the FK506.

4. Conclusion

A modified weighted amino acid network based on a self-consistent contact potential is proposed in this paper. This model contains two types of weight, one is the similar weight and the other is the dissimilar weight. By the analysis of the influence of different definitions of the distance based on the weights, it is revealed that the distance definition contains two types of weights is more reasonable. The average shortest path length has a significant linear correlation with the radius gyration of the molecule. For a set of 197 proteins, through the analysis of the network parameters of the weighted amino acid networks, it is found that the weighted amino acid network holds an obvious "small-world" property. Additionally, with the protein CI2 as an example, through the analysis of the changes of the weighted network parameters on the unfolding pathway, it is observed that the shortest path length of the weighted network will rise increasingly when the protein is unfolding. The highly central residues of the amino acid network play a key role in the binding of protein with drug. These central nodes either interacted directly with the drug or contacted with a residue which is interacted directly with the drug. In other words, for the interaction path between these central residues with the drug, at most, there is an interval between them.

This modified weighted network, which contains two types of weights, is more reasonable than the previous model. This work is helpful for the studies of the structure-function relationship and also is beneficial to the drug design.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 31070828, 31271005, and 11032008), the China Postdoctoral Science Foundation funded project (Grant no. 2012T50247, 20100471587), Program for the Innovative Talents of Higher Learning Institutions of Shanxi, and Natural Science Foundation of Shanxi (Grant no. 2009021018-2). An earlier version of this paper was presented at the International Conference on ICISE 2010.

References

[1] R. Albert and A. L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

[2] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[3] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus, "Small-world view of the amino acids that play a key role in protein folding," *Physical Review E*, vol. 65, no. 6, Article ID 061910, 4 pages, 2002.

[4] N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich, "Topological determinants of protein folding," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 13, pp. 8637–8641, 2002.

[5] M. E. Gáspár and P. Csermely, "Rigidity and flexibility of biological networks," *Briefings in Functional Genomics*, vol. 11, no. 6, pp. 443–456, 2012.

[6] P. Csermely, T. Korcsmáros, H. J. M. Kiss et al., "Structure and dynamics of molecular networks: a novel paradigm of drug discovery. A comprehensive review," http://arxiv.org/abs/1210.0330.

[7] A. R. Atilgan, P. Akan, and C. Baysal, "Small-world communication of residues and significance for protein dynamics," *Biophysical Journal*, vol. 86, no. 1 I, pp. 85–91, 2004.

[8] K. V. Brinda and S. Vishveshwara, "A network representation of protein structures: implications for protein stability," *Biophysical Journal*, vol. 89, no. 6, pp. 4159–4170, 2005.

[9] L. H. Greene and V. A. Higman, "Uncovering network systems within protein structures," *Journal of Molecular Biology*, vol. 334, no. 4, pp. 781–791, 2003.

[10] W. Sun and J. He, "From isotropic to anisotropic side chain representations: comparison of three models for residue contact estimation," *PLoS ONE*, vol. 6, no. 4, article e19238, 2011.

[11] S. Miyazawa and R. L. Jernigan, "Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading," *Journal of Molecular Biology*, vol. 256, no. 3, pp. 623–644, 1996.

[12] S. Miyazawa and R. L. Jernigan, "Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues," *Proteins: Structure, Function, and Bioinformatics*, vol. 34, no. 1, pp. 49–68, 1999.

[13] W. Sun, "Normal mode analysis of protein structure dynamics based on residue contact energy," in *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW '11)*, 2011.

[14] L. D. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, "Characterization of complex networks: a survey of measurements," *Advances in Physics*, vol. 56, no. 1, pp. 167–242, 2007.

[15] X. Jiao, S. Chang, C. H. Li, W. Z. Chen, and C. X. Wang, "Construction and application of the weighted amino acid network based on energy," *Physical Review E*, vol. 75, no. 5, Article ID 051903, 2007.

[16] M. Aftabuddin and S. Kundu, "Weighted and unweighted network of amino acids within protein," *Physica A: Statistical Mechanics and Its Applications*, vol. 369, no. 2, pp. 895–904, 2006.

[17] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, pp. 3747–3752, 2004.

[18] K. I. Goh, E. Oh, B. Kahng, and D. Kim, "Betweenness centrality correlation in social networks," *Physical Review E*, vol. 67, no. 1, Article ID 017101, 4 pages, 2003.

[19] A. del Sol and P. O'Meara, "Small-world network approach to identify key residues in protein-protein interaction," *Proteins: Structure, Function and Genetics*, vol. 58, no. 3, pp. 672–682, 2005.

[20] E. Lindahl, B. Hess, and D. van der Spoel, "GROMACS 3.0: a package for molecular simulation and trajectory analysis," *Journal of Molecular Modeling*, vol. 7, no. 8, pp. 306–317, 2001.

[21] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[22] C. J. Stam, B. F. Jones, G. Nolte, M. Breakspear, and P. Scheltens, "Small-world networks and functional connectivity in Alzheimer's disease," *Cerebral Cortex*, vol. 17, no. 1, pp. 92–99, 2007.

[23] P. Mendes, W. Sha, and K. Ye, "Artificial gene networks for objective comparison of analysis algorithms," *Bioinformatics*, vol. 19, supplement 2, pp. ii122–ii129, 2003.

[24] P. Brazhnik, A. de la Fuente, and P. Mendes, "Gene networks: how to put the function in genomics," *Trends in Biotechnology*, vol. 20, no. 11, pp. 467–472, 2002.

[25] D. J. Watts, *Six Degrees: The Science of a Connected Age*, WW Norton, New York, NY, USA, 2004.

[26] U. K. Muppirala and Z. Li, "A simple approach for protein structure discrimination based on the network pattern of conserved hydrophobic residues," *Protein Engineering, Design and Selection*, vol. 19, no. 6, pp. 265–275, 2006.

[27] E. Estrada, "Universality in protein residue networks," *Biophysical Journal*, vol. 98, no. 5, pp. 890–900, 2010.

[28] W. Sun and J. He, "Understanding on the residue contact network using the log-normal cluster model and the multilevel wheel diagram," *Biopolymers*, vol. 93, no. 10, pp. 904–916, 2010.

[29] R. Liu and J. Hu, "Computational prediction of heme-binding residues by exploiting residue interaction network," *PloS One*, vol. 6, no. 10, article e25560, 2011.

[30] S. Grosdidier and J. Fernandez-Recio, "Protein-protein docking and hot-spot prediction for drug discovery," *Current Pharmaceutical Design*, vol. 18, no. 30, pp. 4607–4618, 2012.

[31] C. Pons, F. Glaser, and J. Fernandez-Recio, "Prediction of protein-binding areas by small-world residue networks and application to docking," *BMC Bioinformatics*, vol. 12, no. 1, p. 378, 2011.

[32] X. Jiao and S. Chang, "Scoring function based on weighted residue network," *International Journal of Molecular Sciences*, vol. 12, no. 12, pp. 8773–8786, 2011.

[33] S. Chang, X. Jiao, C. H. Li, X. Q. Gong, W. Z. Chen, and C. X. Wang, "Amino acid network and its scoring application in protein-protein docking," *Biophysical Chemistry*, vol. 134, no. 3, pp. 111–118, 2008.

[34] G. D. van Duyne, R. F. Standaert, P. A. Karplus, S. L. Schreiber, and J. Clardy, "Atomic structure of FKBP-FK506, an immunophilin-immunosuppressant complex," *Science*, vol. 252, no. 5007, pp. 839–842, 1991.

[35] T. Wang, P. K. Donahoe, and A. S. Zervos, "Specific interaction of type I receptors of the TGF-β family with the immunophilin FKBP-12," *Science*, vol. 265, no. 5172, pp. 674–676, 1994.

[36] J. Liu, J. D. Farmer, W. S. Lane, J. Friedman, I. Weissman, and S. L. Schreiber, "Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP-FK506 complexes," *Cell*, vol. 66, no. 4, pp. 807–815, 1991.

[37] C. C. S. Deivanayagam, M. Carson, A. Thotakura, S. V. L. Narayana, and R. S. Chodavarapu, "Structure of FKBP12.6 in complex with rapamycin," *Acta Crystallographica Section D: Biological Crystallography*, vol. 56, no. 3, pp. 266–271, 2000.

[38] G. D. van Duyne, R. F. Standaert, S. L. Schreiber, and J. Clardy, "Atomic structure of the rapamycin human immunophillin FKBP-12 complex," *Journal of the American Chemical Society*, vol. 113, no. 19, pp. 7433–7434, 1991.

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2012, Article ID 568950, 10 pages
doi:10.1155/2012/568950

Research Article

Identification and Functional Annotation of Genome-Wide ER-Regulated Genes in Breast Cancer Based on ChIP-Seq Data

Min Ding,[1,2] Haiyun Wang,[2] Jiajia Chen,[3] Bairong Shen,[3] and Zhonghua Xu[4]

[1] *Department of Viral and Gene Therapy, Eastern Hepatobiliary Surgery Hospital, Second Military Medical University, Shanghai 200438, China*
[2] *School of Life science and Technology, Tongji University, Shanghai 200092, China*
[3] *Center for Systems Biology, Soochow University, Suzhou Jiangsu 215006, China*
[4] *Department of Cardiothoracic Surgery, Second Affiliated Hospital of Soochow University, Suzhou Jiangsu 215004, China*

Correspondence should be addressed to Zhonghua Xu, drxuzh@sohu.com

Received 1 November 2012; Accepted 18 December 2012

Academic Editor: Hong-Bin Shen

Estrogen receptor (ER) is a crucial molecule symbol of breast cancer. Molecular interactions between ER complexes and DNA regulate the expression of genes responsible for cancer cell phenotypes. However, the positions and mechanisms of the ER binding with downstream gene targets are far from being fully understood. ChIP-Seq is an important assay for the genome-wide study of protein-DNA interactions. In this paper, we explored the genome-wide chromatin localization of ER-DNA binding regions by analyzing ChIP-Seq data from MCF-7 breast cancer cell line. By integrating three peak detection algorithms and two datasets, we localized 933 ER binding sites, 92% among which were located far away from promoters, suggesting long-range control by ER. Moreover, 489 genes in the vicinity of ER binding sites were identified as estrogen response elements by comparison with expression data. In addition, 836 single nucleotide polymorphisms (SNPs) in or near 157 ER-regulated genes were found in the vicinity of ER binding sites. Furthermore, we annotated the function of the nearest-neighbor genes of these binding sites using Gene Ontology (GO), KEGG, and GeneGo pathway databases. The results revealed novel ER-regulated genes pathways for further experimental validation. ER was found to affect every developed stage of breast cancer by regulating genes related to the development, progression, and metastasis. This study provides a deeper understanding of the regulatory mechanisms of ER and its associated genes.

1. Introduction

Breast cancer is a complex disease with high occurrence. It involves a wide range of pathological entities with diverse clinical courses. Gene and protein expression have been extensively profiled in different subtypes of breast cancer [1]. Growth of human breast cells is closely regulated by hormone receptors. Estrogen receptor (ER), a hormonal transcription factor, plays a critical role in the development of breast cancer. Combined with estrogen, it regulates the expression of multiple genes. Studies have found that ER-positive and ER-negative breast cancers are fundamentally different [2]. The outcome of hormone receptor positive tumors is better than hormone receptor negative tumors [3]. Thus, the identification of ER target genes may reveal critical biomarkers for cancer aggressiveness and is therefore crucial to understanding the global molecular mechanisms of ER in breast cancer. To identify direct target genes of ER, it is necessary to map the ER binding sites across the genome. ChIP-Seq is an effective technology for the genome-wide localization of histone modification and transcription factor binding sites. It enables researchers to fully understand many biological processes and disease states, including transcriptional regulation of ES cells, tissue samples, and cancer cells.

Several previous studies have been dedicated to ER-regulated genes and their function in breast cancer cell line [4, 5]. However, most studies lacked the comprehensive and genome-wide view and failed to perform an integrated analysis. In this study, we combined ChIP-Seq and microarray

Table 1: The CHIP-Seq datasets.

Dataset	Platform	Cell line	Sample information
GSE19013	Illumina	MCF-7	Ethanol treated
			E2-treated
GSE14664	Illumina	MCF-7	ER_minus_ligand
			ER_E2

datasets to analyze the ER-regulated genes in the MCF-7 breast cancer cell line. The molecular mechanisms of ER were fully studied, including binding sites, motif, regulated genes, related single nucleotide polymorphisms (SNPs) and functional annotation. The process of this analysis was illustrated in Figure 1.

2. Materials and Methods

2.1. Datasets. The breast cancer associated ChIP-Seq datasets were extracted from Gene Expression Omnibus (GEO): GSE19013 [6] and GSE14664 [7]. Both datasets can be used to survey genome-wide binding of estrogen receptor (ER) in the MCF-7 breast cancer cell line. Control sample was incorporated for the genomic peak finding of ER. (See Table 1 for details.)

2.2. Chip-Seq Analysis. Bowtie [8] was selected to align sequence tags to human genome. Bowtie is an ultrafast and best short-read aligner. It is suitable for sets of short reads where many reads have at least one good and valid alignment, many reads with relatively high quality, and the number of alignment reported per read is small (closed to 1). ChIP-seq datasets we used were satisfied these criteria. In the analysis, tags were selected using the criterion that alignments had no more than 2 mismatches in the first 35 bases on the high quality end of the read, and the sum of the quality values at all mismatched positions could not exceed 70.

Peak detection algorithm is crucial to the analysis of ChIP-Seq dataset. Currently, several tools are available to identify genome-wide binding sites of transcription factors, such as FindPeaks [9], F-Seq [10], CisGenome [11], MACS [12], SISSRs [13], and QuEST [14]. These different methods have their own advantages and disadvantages, although they act in a similar manner. Table 2 showed an overview of the characteristics of these algorithms. ChIP-Seq data has regional biases because of sequencing and mapping biases, chromatin structure, and genome copy number variations [15]. It is believed that more robust ChIP-Seq peak predictions can be obtained by matching control samples [12]. In order to get more stable result, three tools, CisGenome, MACS, and QuEST, were used to identify the binding sites of ER in this study. All the three tools systematically used control samples to guide peak finding and calculate the FDR (False Discovery Rate) value of peaks.

Additionally, MEME program [16] was employed for de novo motif search, keeping default options (minimum width: 6, maximum width: 50, motifs to find: 3, and minimum sites: ≥ 2). For each site, statistical significance (P value) gives the probability of a random string having the same match score or higher. And a criterion of P-value < 0.01 was used here.

2.3. Expression and SNP Analysis. Expression analysis was performed using the same package [17, 18]. Differentially expressed genes were selected based on the q-value less than 1%.

Using the table SNP (131) (dbSNP build 131) [19] in UCSC (http://genome.ucsc.edu/), we identified SNPs near the ER binding sites. The SNPs with at least one mapping in the regions were selected.

2.4. Functional Annotation. Three functional annotation systems, the Gene Ontology (GO) categories [20], canonical KEGG Pathway Maps [21], and commercial software MetaCore-GeneGo Pathway Maps, were used to perform the enrichment analysis for gene function.

Enrichment of GO categories was determined with the Gene Ontology Tree Machine (GOTM) [22], using Hypergeometric test, Multiple test adjustment (BH), and a P-value cut-off of 0.01. WebGestalt (WEB-based GEne SeT AnaLysis Toolkit) [23] (http://bioinfo.vanderbilt.edu/webgestalt/option.php) was used for enrichment of KEGG Pathway. Hypergeometric test, Multiple test adjustment (BH), and a P-value cut-off of 0.01 were also used as criterion. MetaCore-GeneGo is a commercial software which offers gene expression pathway analysis and bioinformatics solutions for systems biology research and development. Hypergeometric intersection was used to estimate P-value, the lower P-value means higher relevance. P-value < 0.01 and FDR < 0.05 were used as criterion.

3. Results and Discussion

3.1. ChIP-Seq Analysis Mapped ER Binding Sites across the Human Genome. Using ChIP-Seq datasets, we identified the global ER binding sites. Sequence tags were firstly aligned to human genome assembly (UCSC, hg19) using Bowtie. Three ChIP-Seq peak calling programs, CisGenome, MACS, and QuEST, were selected to identify the enriched binding peaks. Using a false discovery rate of 0.01, 933 ER binding peaks were revealed by all the three tools in both datasets (Table 3). There were differences among the predicted results using different methods in both two datasets (Figure 2). The calculated FDR value was not only related to different methods, but also influenced by datasets. The overlapped binding sites seemed to be more robust, with 84.9% having FDR value less than 0.005 in all methods and datasets. These binding sites were used for the following analysis. Firstly, we compared these binding sites with two published studies by Welboren et al. [7] and Hu et al. [6]. Our results showed a substantial overlap with the two studies (77.8 and 78.5%, resp.). Also, 719 binding sites, which were shared by all three studies, were likely to be more reliable. The presence of consensus sequence motifs in the ER binding sites was also examined. De novo motif search using the MEME program

Table 2: An overview of the characteristics of different Chip-Seq peak detection algorithm.

Algorithm	Profile	Background model	Control sample	Use control to compute FDR
F-Seq	Kernel density estimation (KDE)		√	
FindPeaks	Aggregation of overlapped tags	Monte Carlo		
SISSRs	Window scan	Poisson	√	
QuEST	Kernel density estimation (KDE)		√	√
MACS	Tags shifted then window scan	dynamic Poisson	√	√
CisGenome	Strand-specific window scan	Negative binomial	√	√

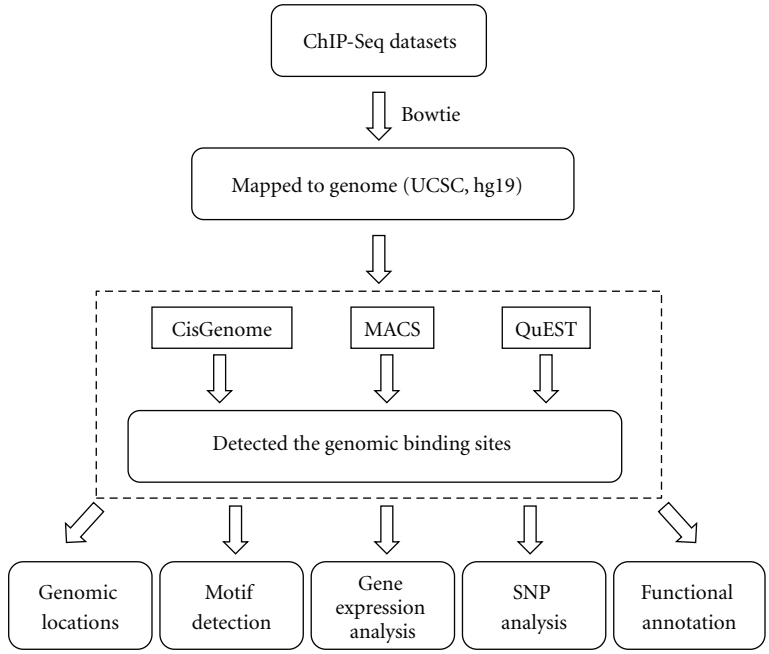


Figure 1: The ChIP-Seq data analyzing pipeline.

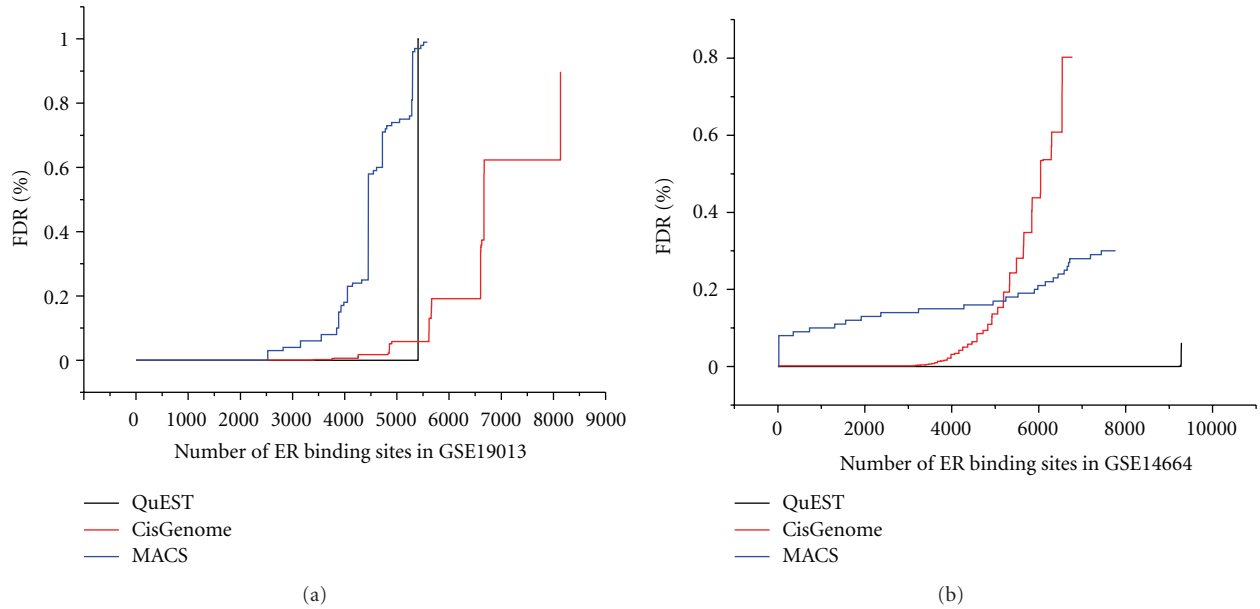


Figure 2: Comparison of QuEST, CisGenome, and MACS predicted result. (a) The FDR value in the dataset of GSE19013. (b) The FDR value in the dataset of GSE14664.

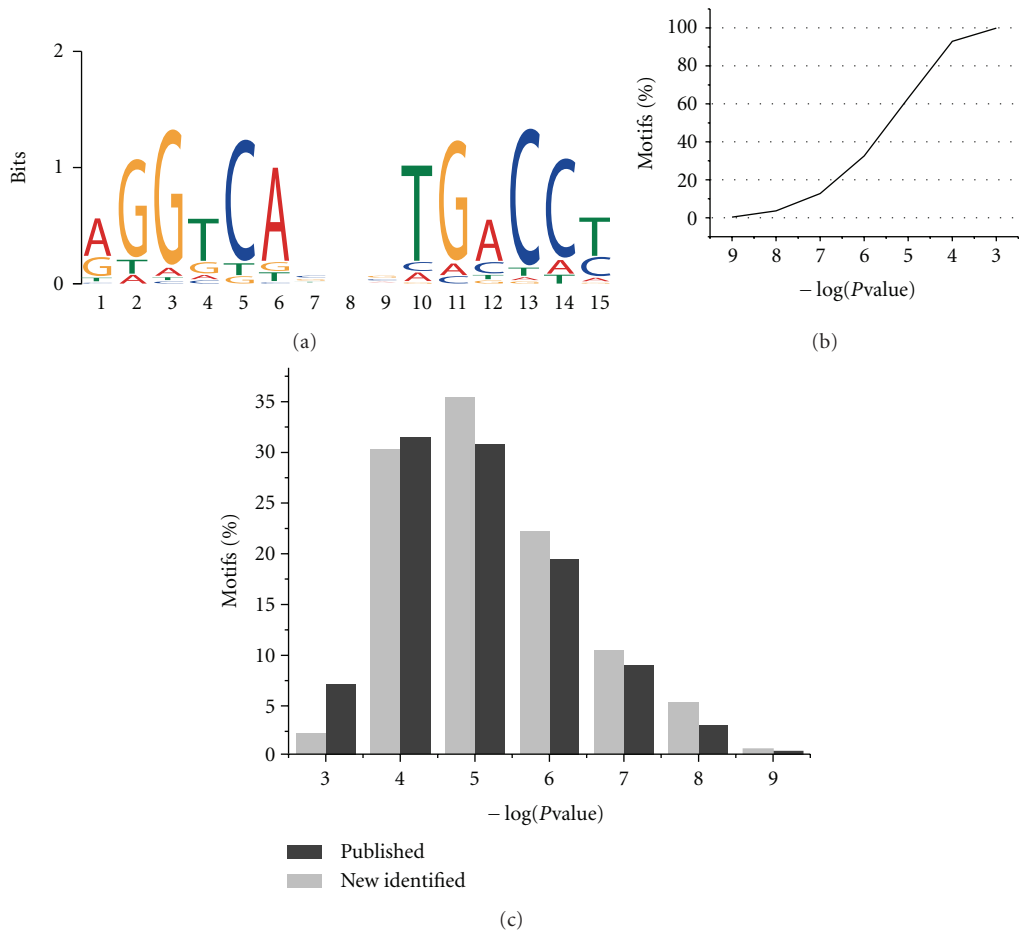


Figure 3: The genomic binding sites of ER. (a) The consensus motif identified in the ERE binding sites. De novo motif search was performed using the MEME program. (b) The percentage of occurrences of ERE motifs in ER binding sites. (c) Comparison of the occurrences of ERE motifs between published and newly identified binding sites.

Table 3: Number of ER binding sites identified by three ChIP-Seq peak calling programs (FDR < 0.01).

Dataset	Number of ER binding sites			Number of overlaped sites	
	CisGenome	MACS	QuEST		
GSE19013	8137	5583	5418	2019	933
GSE14664	6773	7765	9280	5061	

[16] identified a refined ERE motif that was markedly similar to the canonical ERE (Figure 3(a)). Almost all of the ER binding sites contained one or more ERE motif (P-value < 0.01) (Figure 3(b)). Both published and newly identified binding sites contained at least one ERE motif (Figure 3(c)).

Furthermore, we examined the location of ER enrichment sites relativer to the nearest-neighbor genes. The result was shown in Figure 4(a). Only 8% (72) of the peaks occured within gene promoters (defined here as within 5 kb upstream of 5′ to TSS). Also, 34% (317) of the peaks resided in intragenic sites, including 1% (10) in the 3′UTR, 9% (81) in the 5′UTR, 2% (20) in the exon, and 22% (206) in the intron. The occupancy of enhancer (>5 kb away 5′ to TSS) was 35% (332). According to Figure 4(b), the peaks occurred most frequently between −10 kb to −100 kb, +10 kb to +100 kb, with +10 kb to +100 kb being the highest. A further insight into the peaks within +10 kb to +100 kb showed that peaks were preferably located within the regions spanning from +10 kb to +40 kb (Figure 4(c)).

3.2. Using Gene Expression Data to Confirm the ER Binding Sites. In order to determine the specific gene responses corresponding to ER in MCF-7 cells, we compared the nearest-neighbor genes of ER binding sites to the published studies examining differentially expressed genes between ER+ and ER− breast tumors. We used the 3 studies in Table 4 for the gene expression analysis. Differentially expressed genes were selected based on a q-value cut-off of less

Table 4: Breast cancer gene expression dataset and differently expressed genes number (q-value < 1%).

Author	Journal	Array type	Sample N ER+	Sample N ER−	Differently expressed genes	
					Upregulated	Downregulated
Graham et al. [24]	Clin Cancer Res	Affy	15	15	709	333
Wang et al. [25]	Lancet	Affy	209	77	2081	2537
Lu et al. [26]	Breast Cancer Res Treat	Affy	76	53	5136	5445
		All			5692	6101

than 1% using a stringent statistical analysis method. We identified 5692 and 6101 up- and downregulated genes. When combined with the nearest-neighbor genes of ER binding sites, 289 up-regulated genes and 198 down-regulated genes were associated with the ER binding sites (see additional file 1, Supplementary Material available online at doi:10.1155/2012/568950). Among these genes, 33 upregulated genes and 11 downregulated genes were also identified by published ChIP-PET analysis [27].

Our analysis found that more binding sites were associated with ER up-regulated genes (60%) compared to down-regulated genes (40%), indicating that ER was more frequently involved in the direct regulation of up-regulated genes. We also examined the location of ER binding sites in up-regulated and down-regulated genes. As shown in Figure 5, both the up- and down-regulated genes occurred most frequently between −10 kb to −100 kb, +10 kb to +100 kb, which verified the long-range control mode of ER factor.

3.3. SNPs Occurred near the ER Binding Sites. Current studies have shown that the breast cancer risks are associated with commonly occurring single nucleotide polymorphisms (SNPs) [28–32]. The table SNP (131) (dbSNP build 131) in UCSC (http://genome.ucsc.edu/) was used to identify SNPs near the ER binding sites. A total of 2694 SNP loci were found and subsequently annotated using dbSNP in NCBI.

Compared with the differently expressed gene set in the vicinity of ER binding sites, 836 SNPs in or near 157 ER-regulated genes were identified (see additional file 2). Most of the SNPs (94.5%) were located in intron and untranslated regions. Only 5.5% were located in the regions of near-gene, coding-synon, missense, and frameshift. These SNPs might have close relationship with breast cancer.

3.4. Functional Annotation of ER Binding Sites. To identify the biological processes and pathways altered by ER, we employed three functional annotation systems, the Gene Ontology (GO) categories [20], canonical KEGG Pathway Maps [21], and commercial software MetaCore-GeneGo Pathway Maps, to perform the enrichment analysis for gene function.

To gain an overview of the biological processes in which the nearest-neighbor genes of ER binding sites reside, we firstly performed gene set enrichment analysis using Gene Ontology database. Statistically significant (Hypergeometric test, P-value < 0.01) enriched GO terms were identified using the web tool GOTM (Gene Ontology Tree Machine) [22]. The Gene Ontology Directed Acyclic Graph for the nearest-neighbor genes generated by GOTM was presented in Figure 6. The terms with red color were significantly enriched. In terms of biological process, negative regulation of biological process and cellular process, cellular component movement, and regulation of localization and locomotion, structure and system development were significantly enriched. Furthermore, whether differently expressed or not, genes were mostly associated with biological regulation and metabolic process in biological process terms, protein binding in molecular function terms, and membrane in cellular component terms (each term included more than 100 genes). Gene functions for all the nearest-neighbor genes were summarized in Table 5.

The KEGG Pathway database (posted on May 23, 2011) was used to identify functional modules regulated by ER. Seventeen significantly enriched pathways (P-value < 0.01) were revealed (Table 6). In these pathways, most genes were also differentially expressed between ER+ and ER− tumors. Pathways in cancer, focal adhesion, axon guidance, regulation of actin cytoskeleton, and MAPK signaling pathway ranked among the most enriched pathways. The top enriched maps, such as focal adhesion pathway and MAPK signaling pathway, were reported to be related with ER in breast cancer. High expression of focal adhesion kinase had been reported to be related to cancer progression of breast. And tumors with high expression of focal adhesion kinase lack ER and PR [33]. It was also reported that hyperactivation of MAPK could repress the ER expression in breast tumors [34]. Pathways in cancer were the top enriched KEGG pathway. The abnormal expression of some genes occurred in several types of cancer [35–37]. Axon guidance pathway played important roles in cancers. Axon guidance molecules might control the development, migration, and invasion of cancer cells [38]. Regulation of actin cytoskeleton was related to cancer cell migration and invasion [39]. This indicated the crucial role of ER in the development, migration, and invasion of breast cancer.

GeneGo was also used to perform the pathway analysis. Ten pathways were found to be significantly enriched with P-value < 0.01 and FDR < 0.05 (Table 7). The result showed that ER binding sites were enriched in breast cancer related pathways. Among the top five maps, development_prolactin receptor signaling and development_glucocorticoid receptor signaling had been reported to associate with ER [40, 41]. development_ligand-independent activation of ESR1 and ESR2 was another enriched map which might have close

Table 5: The comparison of top enriched GO categories between different expressed and other nearest-neighbor genes of ER binding sites (number of genes ≥ 100).

Genes set	Biological process	Molecular function	Cellular component
Differently expressed	Biological regulation, metabolic process, cell communication, organismal process, localization, developmental process	Protein binding, iron binding	Membrane, nucleus
Others	Biological regulation, metabolic process	Protein binding	Membrane

Table 6: KEGG pathways enriched with the nearest-neighbor genes of ER binding sites (P-value < 0.01).

KEGG ID	Pathways name	P-value	Number of genes	Number of different expressed genes
hsa05200	Pathways in cancer	$2.24E-05$	22	16
hsa04510	Focal adhesion	0.0002	15	14
hsa04360	Axon guidance	0.0009	11	8
hsa04810	Regulation of actin cytoskeleton	0.0012	14	11
hsa04010	MAPK signaling pathway	0.0022	15	12
hsa04114	Oocyte meiosis	0.0024	9	8
hsa04144	Endocytosis	0.0024	12	11
hsa04115	p53 signaling pathway	0.0024	7	7
hsa05216	Thyroid cancer	0.0024	5	4
hsa05218	Melanoma	0.0033	7	3
hsa04020	Calcium signaling pathway	0.004	11	4
hsa04062	Chemokine signaling pathway	0.0064	11	9
hsa04914	Progesterone-mediated oocyte maturation	0.0085	7	7
hsa01100	Metabolic pathways	0.0086	35	28
hsa00450	Selenoamino acid metabolism	0.0088	4	3
hsa05414	Dilated cardiomyopathy	0.0096	7	7
hsa03440	Homologous recombination	0.0097	4	3

Table 7: Terms of the enriched GeneGo pathway maps (P-value < 0.01, FDR < 0.05).

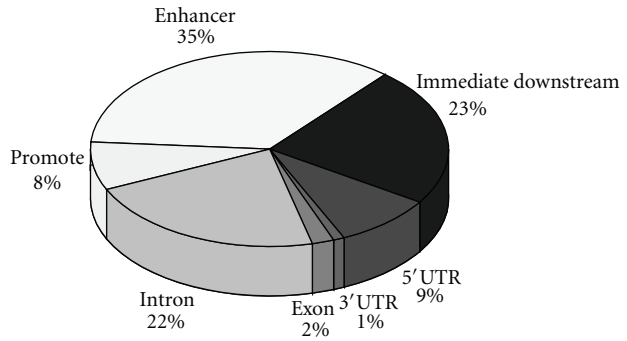
GeneGo pathway terms	P-value
Apoptosis and survival_APRIL and BAFF signaling	$1.29889E-05$
Development_prolactin receptor signaling	$4.95517E-05$
Development_glucocorticoid receptor signaling	$5.81237E-05$
Development_ligand -independent activation of ESR1 and ESR2	0.000295251
Immune response_IL-22 signaling pathway	0.000381484
Development_EPO-induced Jak-STAT pathway	0.000531744
Development_growth hormone signaling via STATs and PLC/IP3	0.000531744
Cytoskeleton remodeling_keratin filaments	0.000622315
Development_GM-CSF signaling	0.000660576
Transcription_transcription regulation of aminoacid metabolism	0.000752764

relationship with ER. APRIL and BAFF were the members of tumor necrosis factor family which related to a plethora of cellular events from proliferation and differentiation to apoptosis and tumor reduction [42]. IL-22 might play a role in the control of tumor growth and progression in breast [43]. However, the relationship between ER and these two pathways need further experimental study.

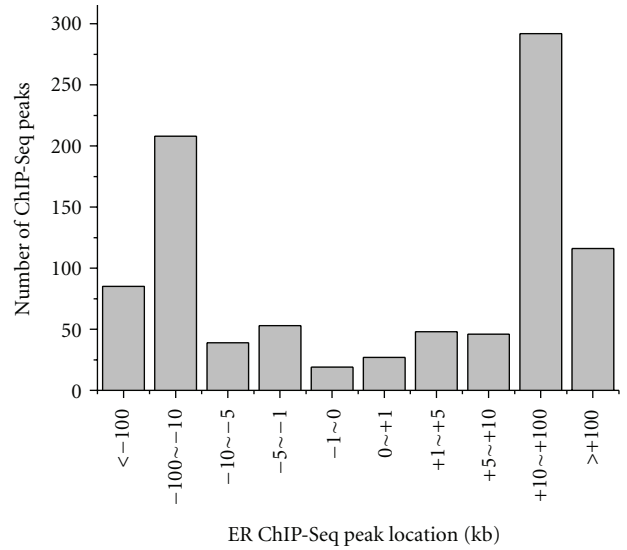
4. Conclusions

ER is an important molecular symbol of breast cancer. A full understanding of the molecular mechanisms of ER will be useful for the research in the prediction and treatment of breast cancer. The ChIP-Seq technology is useful to study the interaction of protein and DNA on a genome-wide scale. ChIP-Seq data can effectively analyze the regulatory mechanism of transcription factor in genome-wide scale. In this study, we used ChIP-Seq data to identify the global sites regulated by ER in MCF-7 breast cancer cell line. In order to get more reliable result, three different tools were used to analyze two datasets. And 933 binding sites were identified, and the ERE motif was refined here.

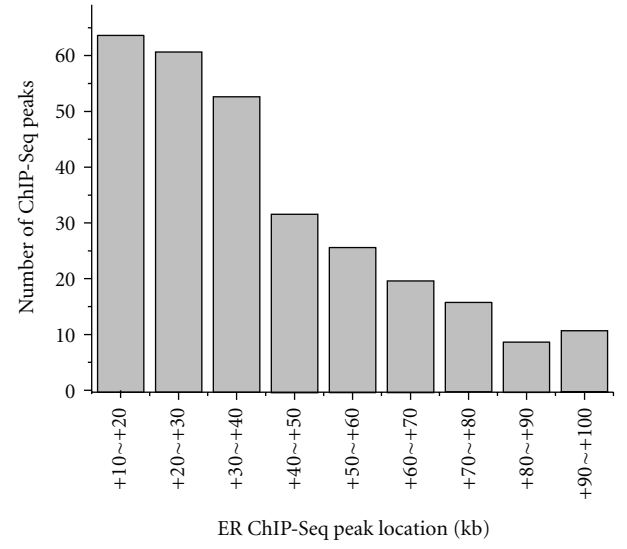
The analysis of the global genomic occupancy of ER-regulated genes revealed that 92% of the total 933 ER-binding



(a)



(b)



(c)

Figure 4: Location analysis of ER binding sites. (a) locations relative to nearest-neighbor genes. (b) Genomic Locations of ER ChIP-Seq peaks. (c) Genomic locations of ER ChIP-Seq peaks within +10~+100 kb.

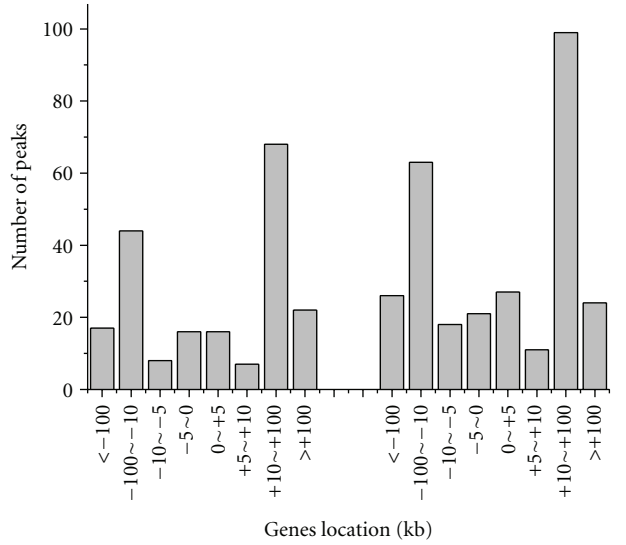


Figure 5: Genomic Locations of differentially expressed genes in the vicinity of ER binding sites.

sites were located far away from promoters. This suggested that the canonical mode of ER factor function involved long-range control. Previous research had reported that ER-α includes looping [44]. Using ChIP-PET, Lin et al. [27] had analyzed the genome-wide ER-α chromatin occupancy and revealed abundant nonpromoter sites. Our findings provided further support for this mode of ER factor function.

We compared the ER binding sites found in this study with published differentially expressed genes between ER+ and ER− breast tumors. A set of 487 genes was found significant in discriminating ER status in breast tumors. This indicated that these genes appeared to affect ER response. Only 9% (44) of the genes have been identified by Lin et al. [27], while the remaining need further validations. We found that binding sites were preferentially associated with ER up-regulated genes, indicating that ER was more frequently involved in the regulation of upregulated genes. The location of 487 genes verified the long-range control mode of ER factor.

In this study, we found 2694 single nucleotide polymorphisms loci located in or near the ER binding sites. Among these SNPs, the 157 genes of 836 SNPs were also differentially expressed between ER+ and ER− breast tumors. It indicated that this set of SNPs might have close relationship with ER in breast.

The functional annotation provided a deeper understanding of ER and ER-associated genes. Enrichment analysis of GO gave an overview of gene function. As shown in Figure 6, significantly enriched terms belonged to three classes, biological regulation, cellular processes, and developmental processes. The result of KEGG enrichment analysis was similar. Five pathways were involved in cellular processes, including focal adhesion, regulation of actin cytoskeleton, oocyte meiosis, endocytosis, and p53 signaling pathway. These pathways were associated with cell communication,

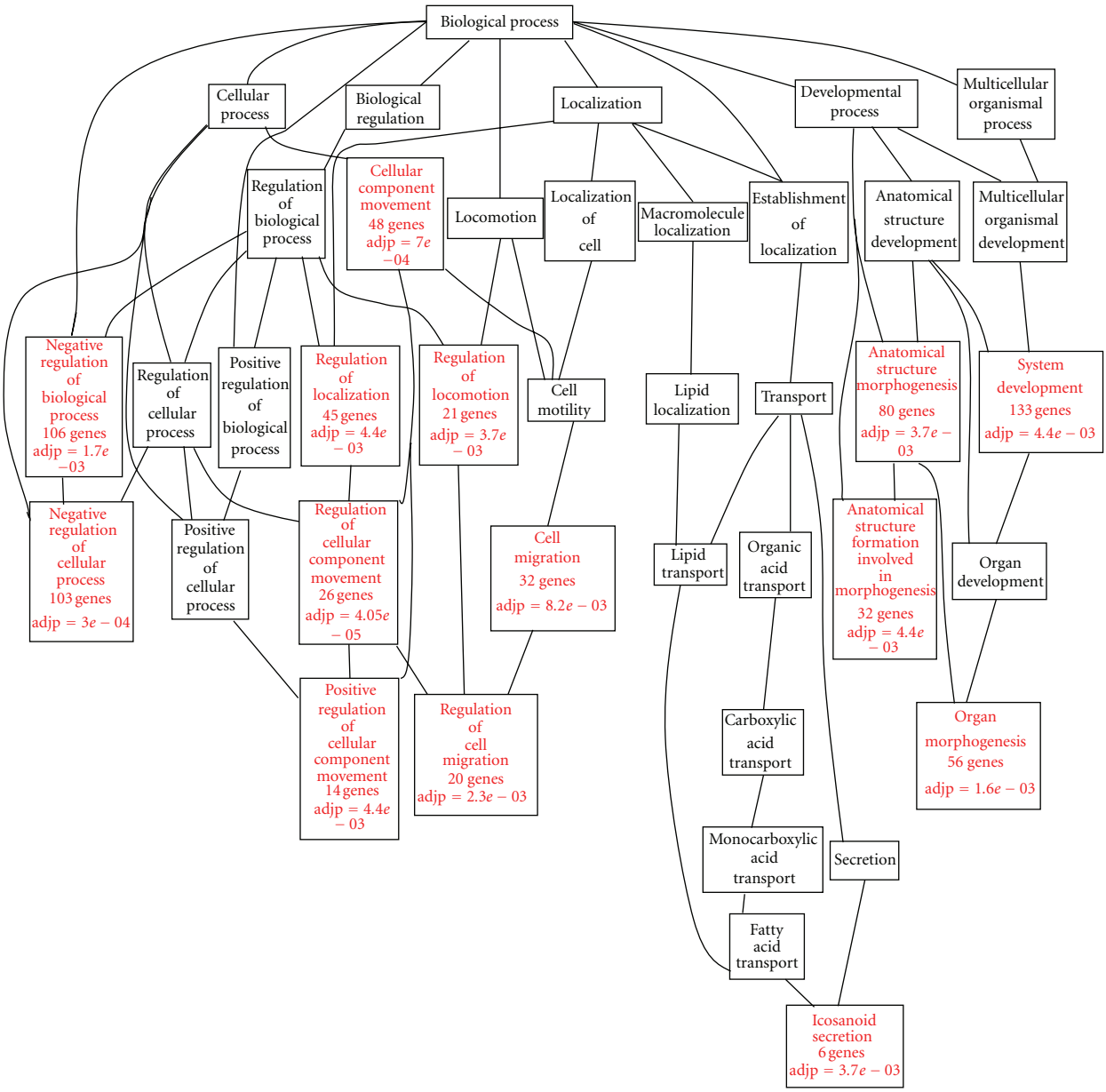


Figure 6: Directed Acyclic Graphs (DAGs) of significantly enriched GO (Gene Ontology) categories ($P < 0.01$).

movement, growth, and death. Most enriched terms determined by GeneGO were development pathways. It was suggested that ER-regulated genes participated in various development processes. Moreover, KEGG pathway analysis suggested that ER-regulated genes were enriched in some diseases related pathways. Both KEGG and GeneGO pathway analysis revealed that some immune-related pathways were enriched, such as chemokine signaling pathway and immune response_IL-22 signaling pathway. These results indicated that ER-regulated genes related to the development, progression, and metastasis of breast. ER affected every developed stage of breast. However, the regulatory mechanisms of ER in different stages and different pathways still need further studies.

Conflict of Interests

The authors declare that they have no conflict of interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China Grants (no. 91230117 and 31170795), the Specialized Research Fund for the Doctoral Program of Higher Education of China (20113201110015), International S&T Cooperation Program of Suzhou (SH201120), and the National High Technology Research and Development Program of China (863 program, Grant no. 2012AA02A601).

References

[1] M. J. van de Vijver, Y. D. He, L. J. van'T Veer et al., "A gene-expression signature as a predictor of survival in breast cancer," *The New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[2] M. A. Lopez-Garcia, F. C. Geyer, M. Lacroix-Triki, C. Marchió, and J. S. Reis-Filho, "Breast cancer precursors revisited: molecular features and progression pathways," *Histopathology*, vol. 57, no. 2, pp. 171–192, 2010.

[3] W. F. Anderson, N. Chatterjee, W. B. Ershler, and O. W. Brawley, "Estrogen receptor breast cancer phenotypes in the surveillance, epidemiology, and end results database," *Breast Cancer Research and Treatment*, vol. 76, no. 1, pp. 27–36, 2002.

[4] S. Mandal and J. R. Davie, "An integrated analysis of genes and pathways exhibiting metabolic differences between estrogen receptor positive breast cancer cells," *BMC Cancer*, vol. 7, article 181, 2007.

[5] M. C. Abba, Y. Hu, H. Sun et al., "Gene expression signature of estrogen receptor α status in breast cancer," *BMC Genomics*, vol. 6, no. 1, article 37, 2005.

[6] M. Hu, J. Yu, J. M. G. Taylor, A. M. Chinnaiyan, and Z. S. Qin, "On the detection and refinement of transcription factor binding sites using ChIP-Seq data," *Nucleic Acids Research*, vol. 38, no. 7, Article ID gkp1180, pp. 2154–2167, 2010.

[7] W. J. Welboren, M. A. van Driel, E. M. Janssen-Megens et al., "ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands," *EMBO Journal*, vol. 28, no. 10, pp. 1418–1428, 2009.

[8] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, article R25, 2009.

[9] A. P. Fejes, G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, and S. J. M. Jones, "FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology," *Bioinformatics*, vol. 24, no. 15, pp. 1729–1730, 2008.

[10] A. P. Boyle, J. Guinney, G. E. Crawford, and T. S. Furey, "F-Seq: a feature density estimator for high-throughput sequence tags," *Bioinformatics*, vol. 24, no. 21, pp. 2537–2538, 2008.

[11] H. Ji, H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong, "An integrated software system for analyzing ChIP-chip and ChIP-seq data," *Nature Biotechnology*, vol. 26, no. 11, pp. 1293–1300, 2008.

[12] Y. Zhang, T. Liu, C. A. Meyer et al., "Model-based analysis of ChIP-Seq (MACS)," *Genome Biology*, vol. 9, no. 9, article R137, 2008.

[13] R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao, "Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data," *Nucleic Acids Research*, vol. 36, no. 16, pp. 5221–5231, 2008.

[14] A. Valouev, D. S. Johnson, A. Sundquist et al., "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data," *Nature Methods*, vol. 5, no. 9, pp. 829–834, 2008.

[15] R. Redon, S. Ishikawa, K. R. Fitch et al., "Global variation in copy number in the human genome," *Nature*, vol. 444, no. 7118, pp. 444–454, 2006.

[16] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, vol. 2, pp. 28–36, 1994.

[17] J. Li and R. Tibshirani, "Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data," *Statistical Methods in Medical Research*. In press.

[18] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[19] S. T. Sherry, M. H. Ward, M. Kholodov et al., "DbSNP: the NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, 2001.

[20] 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.

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

[22] B. Zhang, D. Schmoyer, S. Kirov, and J. Snoddy, "GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies," *BMC Bioinformatics*, vol. 5, article 16, 2004.

[23] B. Zhang, S. Kirov, and J. Snoddy, "WebGestalt: an integrated system for exploring gene sets in various biological contexts," *Nucleic Acids Research*, vol. 33, no. 2, pp. W741–W748, 2005.

[24] K. Graham, X. Ge, A. de Las Morenas, A. Tripathi, and C. L. Rosenberg, "Gene expression profiles of estrogen receptor-positive and estrogen receptor-negative breast cancers are detectable in histologically normal breast epithelium," *Clinical Cancer Research*, vol. 17, no. 2, pp. 236–246, 2011.

[25] Y. Wang, J. G. M. Klijn, Y. Zhang et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[26] B. Lu, X. Liang, G. K. Scott et al., "Polyamine inhibition of estrogen receptor (ER) DNA-binding and ligand-binding functions," *Breast Cancer Research and Treatment*, vol. 48, no. 3, pp. 243–257, 1998.

[27] C. Y. Lin, V. B. Vega, J. S. Thomsen et al., "Whole-genome cartography of estrogen receptor α binding sites," *PLoS Genetics*, vol. 3, no. 6, article e87, 2007.

[28] A. Beeghly-Fadiel, W. Zheng, W. Lu et al., "Replication study for reported SNP associations with breast cancer survival," *Journal of Cancer Research and Clinical Oncology*, vol. 138, no. 6, pp. 1019–1026, 2012.

[29] W. Han, K. Y. Kim, S. J. Yang, D. Y. Noh, D. Kang, and K. Kwack, "SNP-SNP interactions between DNA repair genes were associated with breast cancer risk in a Korean population," *Cancer*, vol. 118, no. 3, pp. 594–602, 2012.

[30] C. H. Yang, L. Y. Chuang, Y. J. Chen, H. F. Tseng, and H. W. Chang, "Computational analysis of simulated SNP interactions between 26 growth factor-related genes in a breast cancer association study," *OMICS A Journal of Integrative Biology*, vol. 15, no. 6, pp. 399–407, 2011.

[31] K. D. Graves, B. N. Peshkin, G. Luta, W. Tuong, and M. D. Schwartz, "Interest in genetic testing for modest changes in breast cancer risk: implicationsfor SNP testing," *Public Health Genomics*, vol. 14, no. 3, pp. 178–189, 2011.

[32] R. J. Hartmaier, S. Tchatchou, A. S. Richter et al., "Nuclear receptor coregulator SNP discovery and impact on breast cancer risk," *BMC Cancer*, vol. 9, article 438, 2009.

[33] A. L. Lark, C. A. Livasy, L. Dressler et al., "High focal adhesion kinase expression in invasive breast carcinomas is associated with an aggressive phenotype," *Modern Pathology*, vol. 18, no. 10, pp. 1289–1294, 2005.

[34] A. S. Oh, L. A. Lorant, J. N. Holloway, D. L. Miller, F. G. Kern, and D. El-Ashry, "Hyperactivation of MAPK induces

loss of ERα expression in breast cancer cells," *Molecular Endocrinology*, vol. 15, no. 8, pp. 1344–1359, 2001.

[35] H. Cam, H. Griesmann, M. Beitzinger et al., "p53 family members in myogenic differentiation and rhabdomyosarcoma development," *Cancer Cell*, vol. 10, no. 4, pp. 281–293, 2006.

[36] M. P. DeYoung, C. M. Johannessen, C. O. Leong, W. Faquin, J. W. Rocco, and L. W. Ellisen, "Tumor-specific p73 up-regulation mediates p63 dependence in squamous cell carcinoma," *Cancer Research*, vol. 66, no. 19, pp. 9362–9368, 2006.

[37] G. Dominguez, J. M. Silva, J. Silva et al., "Wild type p73 overexpression and high-grade malignancy in breast cancer," *Breast Cancer Research and Treatment*, vol. 66, no. 3, pp. 183–190, 2001.

[38] A. Chédotal, "Chemotropic axon guidance molecules in tumorigenesis," *Progress in Experimental Tumor Research*, vol. 39, pp. 78–90, 2007.

[39] H. Yamaguchi and J. Condeelis, "Regulation of the actin cytoskeleton in cancer cell migration and invasion," *Biochimica et Biophysica Acta*, vol. 1773, no. 5, pp. 642–652, 2007.

[40] K. McHale, J. E. Tomaszewski, R. Puthiyaveettil, V. A. Livolsi, and C. V. Clevenger, "Altered expression of prolactin receptor-associated signaling proteins in human breast carcinoma," *Modern Pathology*, vol. 21, no. 5, pp. 565–571, 2008.

[41] P. Moutsatsou and A. G. Papavassiliou, "The glucocorticoid receptor signalling in breast cancer: breast carcinoma," *Journal of Cellular and Molecular Medicine*, vol. 12, no. 1, pp. 145–163, 2008.

[42] V. Pelekanou, M. Kampa, M. Kafousi et al., "Expression of TNF-superfamily members BAFF and APRIL in breast cancer: immunohistochemical study in 52 invasive ductal breast carcinomas," *BMC Cancer*, vol. 8, article 76, 2008.

[43] G. F. Weber, F. C. Gaertner, W. Erl et al., "IL-22-mediated tumor growth reduction correlates with inhibition of ERK1/2 and AKT phosphorylation and induction of cell cycle arrest in the G 2-M phase," *Journal of Immunology*, vol. 177, no. 11, pp. 8266–8272, 2006.

[44] J. S. Carroll, X. S. Liu, A. S. Brodsky et al., "Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1," *Cell*, vol. 122, no. 1, pp. 33–43, 2005.