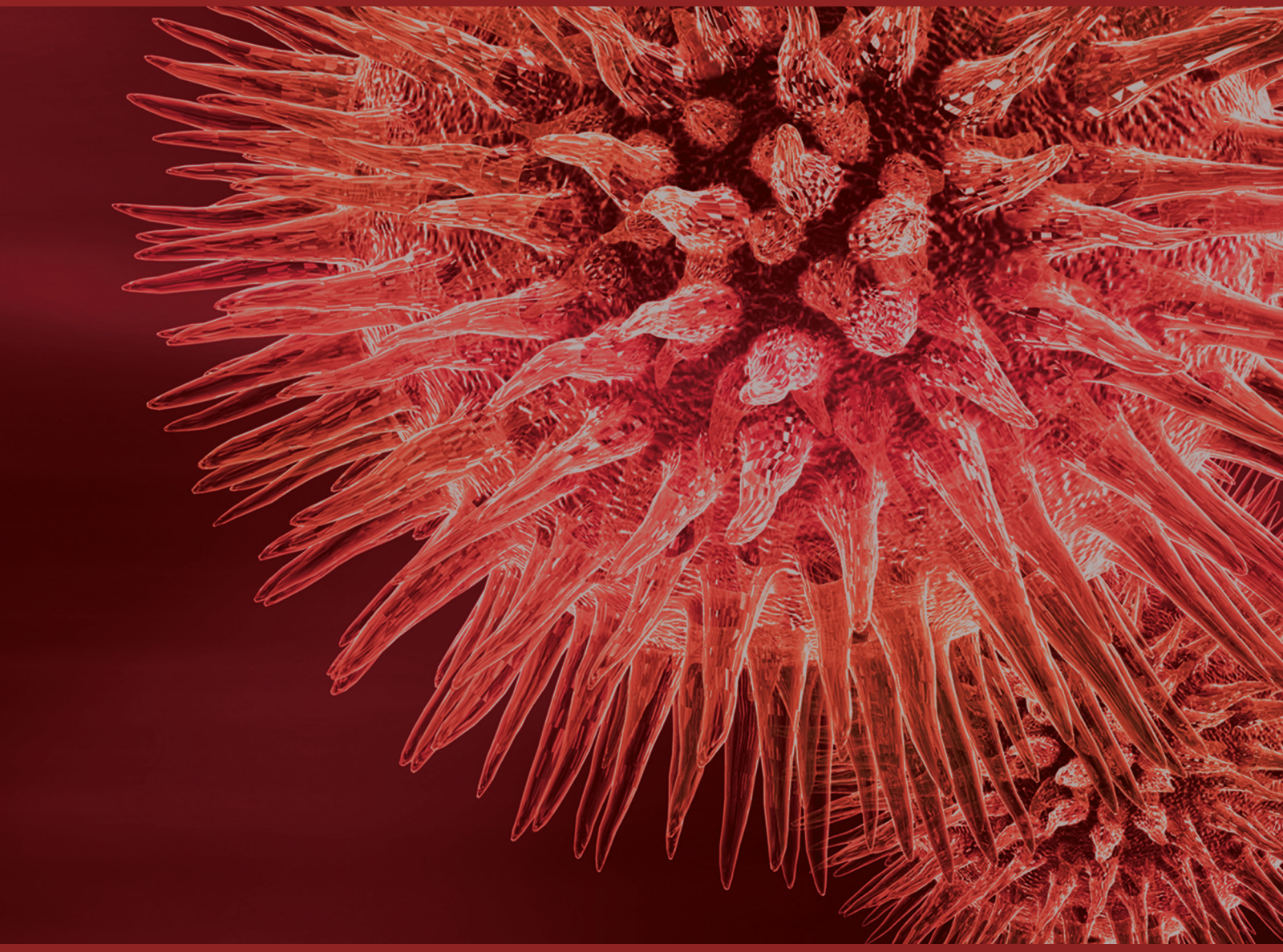
BioMed Research International

Biomedical Data Integration, Modeling, and Simulation in the Era of Big Data and Translational Medicine

Guest Editors: Bairong Shen, Andrew E. Teschendorff, Degui Zhi, and Junfeng Xia



Biomedical Data Integration, Modeling, and Simulation in the Era of Big Data and Translational Medicine

BioMed Research International

Biomedical Data Integration, Modeling, and Simulation in the Era of Big Data and Translational Medicine

Guest Editors: Bairong Shen, Andrew E. Teschendorff, Degui Zhi, and Junfeng Xia

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 731546, 1 page
http://dx.doi.org/10.1155/2014/731546

Editorial

Biomedical Data Integration, Modeling, and Simulation in the Era of Big Data and Translational Medicine

Bairong Shen,[1] Andrew E. Teschendorff,[2] Degui Zhi,[3] and Junfeng Xia[4]

[1] *Center for Systems Biology, Soochow University, P.O. Box 206, Suzhou 215006, China*
[2] *Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT, UK*
[3] *Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA*
[4] *Institute of Health Sciences, Anhui University, Hefei 230601, China*

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

Received 21 July 2014; Accepted 21 July 2014; Published 24 July 2014

Advances in high throughput technologies, specially next-generation sequencing, have generated massive amounts of biological data. To take full advantage of these data and to extract as much information and knowledge from them as possible, we face many challenges. To help address and overcome these challenges and promote the application of informatics to translational research, we launched this special issue.

The biomedical data analyzed in this issue covers molecular, imaging, and clinical data. For instance, J. Shang et al. evaluated and compared multiple aligners for next-generation sequencing data, providing an important guide for biologists to select suitable aligners, and H. Li et al. proposed a method to identify mutated driver pathways in cancer. Y. Li et al. established a tissue type assignment method for glioblastoma multiforme by analyzing the magnetic resonance spectroscopy imaging data and tissue distribution information. F. Liu et al. applied multiple technologies to integrate the clinical and genomic information and to investigate their association for facilitating the diagnosis and treatment of colorectal cancer. X. Zhang et al. analyzed extensive clinical data, summarizing the disease spectrum, in China, and suggesting to pay more attention on disease prevention by promoting lifestyle changes.

In terms of translational research, two aspects, that is, biomarker discovery and drug design for diagnosis or treatment of diseases, were discussed based on computational studies. J. Huang and W. Yan identified micro-RNA biomarkers for sepsis and gastric cancer, based on miRNAs regulatory network analysis and systems biological approach, respectively. Y. Sun et al. successfully implemented big data technologies to a study of drug combinatorial effects. The Hadoop-based model showed higher efficiency and better performance than the traditional methods for the prediction of drug combination effects.

In this issue, we also compiled two technical works for biomedical data analysis. First is the work by H. Li et al., where they developed a hybrid support vector machine (SVM) model for privacy preserving data classification. The second is the work by J. Lei et al., where they made a systematic study on the usability of information technology, especially health information technologies, in China, by the analysis of publications during the past 30 years.

In addition to the analysis of static data, dynamic simulation is also important for biomedical data analysis. L. Jin performed a metadynamics simulation study of conformational transformation of HhaI methyltransferase and proposed that the induced fit model is necessary to understand the function of the studied molecule.

By launching this issue, we wish to give the readers a wider perspective on the future of data modeling and simulation and to leave the readers with the impression that informatics will be the key for successful translational research.

Bairong Shen
Andrew E. Teschendorff
Degui Zhi
Junfeng Xia

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 196858, 5 pages
http://dx.doi.org/10.1155/2014/196858

Research Article

A Hadoop-Based Method to Predict Potential Effective Drug Combination

Yifan Sun, Yi Xiong, Qian Xu, and Dongqing Wei

State Key Laboratory of Microbial Metabolism and College of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

Correspondence should be addressed to Yi Xiong; xiongyi@sjtu.edu.cn and Dongqing Wei; dqwei@sjtu.edu.cn

Received 31 March 2014; Revised 5 July 2014; Accepted 15 July 2014; Published 23 July 2014

Academic Editor: Degui Zhi

Combination drugs that impact multiple targets simultaneously are promising candidates for combating complex diseases due to their improved efficacy and reduced side effects. However, exhaustive screening of all possible drug combinations is extremely time-consuming and impractical. Here, we present a novel Hadoop-based approach to predict drug combinations by taking advantage of the MapReduce programming model, which leads to an improvement of scalability of the prediction algorithm. By integrating the gene expression data of multiple drugs, we constructed data preprocessing and the support vector machines and naïve Bayesian classifiers on Hadoop for prediction of drug combinations. The experimental results suggest that our Hadoop-based model achieves much higher efficiency in the big data processing steps with satisfactory performance. We believed that our proposed approach can help accelerate the prediction of potential effective drugs with the increasing of the combination number at an exponential rate in future. The source code and datasets are available upon request.

1. Introduction

In the past few years, the novel effective drugs come out slowly although there is a substantial investment into the development of drugs. It is common for the pharmaceutical industry to develop novel drugs targeting a certain target. However, the once dominating paradigm of "mono drug mono target" in drug development is now being challenged by the clinical and pharmaceutical people, since the single drug cannot always be effective for the complex diseases (such as cancer and diabetes), which may involve multiple biological pathways and complex pathological process. Therefore, the drug combination, which consists of multiple drugs (the effective chemical molecules), is now becoming a novel strategy to combat complex diseases [1–3].

It is impractical to screen all possible drug combinations experimentally since there will be an exponential explosion when the number of single drugs increases. Therefore, a great number of computational methods have been recently developed for prediction of drug combinations [4–7]. In general, there are three main kinds of computational approaches to identify effective drug combinations: the method of the first kind is to use the stochastic search technique, which is successfully applied in various applications to solve the large-scale combinatorial optimization problems of highly complex systems, and the fast convergence can be achieved with a small number of iterations to find effective drug combinations [5]; the second type is to build a mathematical model based on the median-effect equation in which the "median" is the unified common link of single entity and multiple entities. The disadvantage of this method is that it is hard to interpret the molecular mechanism that underlies the drug combinations [6]; the third type is based on the systems biology principle, which aims to study the possible effect of the various drug combinations on the molecular networks or pathways which they may be involved in. For example, Zhao et al. [4] integrated the molecular and pharmacological features of drugs to predict new potential drug combinations. Wu et al. [7] assumed that the single drug or the drug combinations affected a subnetwork or pathway in the cellular system. They proposed a molecular interaction network-based method to identify effective drug combinations by evaluating the overall effect of one drug or drug combinations.

Although these existing methods can predict novel drug combinations or provide mechanistic insights into existing ones, they are limited by their efficiency when the size of combination space increases at an exponential growth rate (e.g., the number of drugs increases from pairwise combinations to three-wise combinations). Therefore, it is necessary to develop prediction methods that are scalable to data and computation. The Hadoop MapReduce system [8–10] represents a novel program framework with the potential to greatly accelerate data-intensive application. In the present study, we developed the Hadoop-based method to identify the potential effective drug combinations by integrating the gene expression data under the effect of single drugs, the basic information of drug combination, and human disease pathway information. The classification algorithms were then constructed based on the typical perceptron learning algorithm and generative learning algorithm: support vector machine (SVM) and naïve Bayesian for prediction of novel effective drug combinations. The preliminary results indicated that our Hadoop-based implementation of these classification algorithms achieved higher efficiency than the traditional implementation of the algorithms on the dataset with a small number of samples due to insufficient number of effective drug combinations validated. We believe that the proposed Hadoop-based approach will be useful on the larger dataset when the number of drug combinations greatly increases in future.

2. Methods

2.1. Datasets. All the basic information about single drugs and effective drug combinations was extracted from the Drug Combination Database (DCDB) (http://www.cls.zju.edu.cn/dcdb/) [11]. In total, our data set contains 76 pairwise drug combinations involving 103 single drugs, which have well annotated gene expression information (more details explained in the next section). The 76 drug combinations were assigned as the positive samples in the classification models, while the noneffective pairs (called the negative set) were generated by randomly pairing drugs that appeared in the set of the 103 single drugs. The negative set meets the two requirements: (i) the noneffective pairs cannot exist in the set of 76 effective pairs, and (ii) the number of noneffective (negative) pairs is equal to that of effective (positive) pairs.

2.2. Feature Construction. In order to encode the drug combinations, we focus on the possible effect of different drug combinations on the pathways that they may be involved in. The gene expression profiles of the 1309 small-molecule drugs or compounds were downloaded from the Broad Institute Connectivity Map Build02 (http://www.broadinstitute.org/cmap/) [12], and the size of total data is up to 45 GB. We kept the genes which have microarray experiments with at least 3 replicates. The raw expression profiles were processed by using MAS5 algorithm supplied by Affymetrix, which is much faster than RMA (robust multichip average) running on our limited computing capability [13, 14]. The annotated gene set in each human disease pathway was sourced from the Molecular Signatures Database (MSigDB, http://www.broadinstitute.org/gsea/msigdb/) [15]. We finally got 186 gene sets which are related to the human disease pathways.

For the fact that we can only directly obtain the gene expression data of single drugs, we should first represent the feature of pairwise (or multiple) drug combinations. In this study, we applied two different strategies to define the combination feature described as below.

(1) This first kind of representation is a direct way to define the combination feature as a linear function of single drugs. For a drug D_i in the drug combination (D_1, D_2), the expression data of gene G_i is denoted as P_i if it is not affected by drug D_i, and denoted as C_i, if it is affected by drug D_i. Thus, the combination effect of the pairwise drug combination of D_1 and D_2 on G_i is defined as

$$D_{1,2|G} = \left(\frac{P_{1|G}}{C_{1|G}} - 1\right) + \left(\frac{P_{2|G}}{C_{2|G}} - 1\right). \quad (1)$$

Obviously, this is a simple way to get the combination feature of any pairwise drug combination. However, the representation cannot convey the intricacy of drug combinations due to the complexity of human disease mechanism.

(2) Instead, we try another way to find the frequent feature pattern of effective drug combinations and take them as the feature of potential effective drug combinations. Here, we assume that a pathway is affected if there exist genes in this pathway whose expression level is significantly changed under the effect of a single drug. We first performed the Student's t-test for each single drug to get the significantly changed gene set and then mapped them into 186 human disease pathways. This method is finally compared with Zhao et al.'s definition [4], which directly maps the target of the drug into human disease pathway. Finally, we calculated the frequency score of all pairwise drug combinations. The frequency score is defined as below:

$$S_{i,j} = \frac{N_{i,j}(EC)}{N_{i,j}(RC)}, \quad (2)$$

where the denominator shows the number of patterns that emerged in effective pairwise drug combinations and the numerator presents the background frequent patterns in randomly distributed pairwise drug combinations.

2.3. Feature Selection. The feature construction method brings high dimensional feature space on a dataset with small size of samples. To avoid the overfitting, we applied several feature selection methods on our dataset. For the first type of feature construction method mentioned above, we performed the minimum-redundancy-maximum-relevance (mRMR) [16] to select the most important feature for model building, whereas, for the second one, we only need to set a fixed threshold to take the most frequent emerging pattern as the features. In this study, we chose the number of features as one-fourth of the total sample number.

2.4. Model Construction. In the model building step, we employed two popular machine learning algorithms, support vector machine, and naïve Bayesian to train a classifier for predicting effective drug combinations. In the SVM algorithm, the selection of kernel function and related parameters will have a great effect on the performance of the trained classifier. In the training stage, we compared four types of kernel functions: linear kernel, polynomial kernel, Gaussian kernel, and tangent kernel. The SVM classifiers were implemented by using LibSVM package [17]. There are two important parameters when training SVM classifiers, cost factor c for outlier samples and gamma g in kernel functions. There is no smart algorithm to select the best parameters in the training stage, and we searched the optimal parameters using grid search. The search range of the parameters (c and g) is from 0.03125 to 32, with the step as 0.00001. The second type of classification method we used here is the naïve Bayesian algorithm, which can be suitable to be parallelized. In the later section, we will introduce how to implement the MapReduce version of the naïve Bayesian algorithm on the Hadoop platform.

2.5. Scalable Implementation of the Whole Mining Process

2.5.1. Building the Big Data Platform. For scalable implementation of our mining process, we used the machine virtualization to build the Hadoop cluster. The master virtual machines included 4 Intel core i3 processor cores and 4 GB RAM and the two slave virtual machines with 2 Intel core i3 processor cores and 2 GB RAM. The software environment includes Hadoop-1.2.1, Hive-0.11.0, and RHadoop (an integration of R and Hadoop).

After building the scalable Hadoop cluster, we exploited the Hadoop distributed file system to store the raw data and used hive as data ETL tools for relational database and program to process the local files.

2.5.2. Scalable Feature Construction. The feature construction stage can be regarded as a series of independent similar processes on different samples and features. In the Hadoop, we implemented a chain mapper to parallelize the processes, including the gene expression preprocessing and the construction of the proposed drug combination features.

2.5.3. Scalable Model Building. For the SVM algorithm, it is difficult to implement the parallel version. Here, we only parallelized the grid search of the optimal parameters, which are time-consuming in the sequential implementation.

For the naïve Bayesian algorithm, the implementation of the scalable version using MapReduce is mainly composed of three steps (shown in Algorithm 1): the calculation of the prior probability for each class, the conditional probability for each feature under each class, and the conditional probability for each class under each feature.

2.6. Model Validation and Evaluation. A tenfold cross-validation and leave-one-out cross-validation test were used to evaluate the classification performance. To assess the performance of the classification models, we used the accuracy (ACC), sensitivity (SN, also called recall), specificity (SP), and F-measure (F_1). These measures can be calculated by the numbers of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each classifier [18–21]. These performance measures are defined as below:

$$\begin{aligned} \mathrm{ACC} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}, \\ \mathrm{SN} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{SP} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \\ F_1 &= \frac{2\times\mathrm{TP}}{\mathrm{FP}+\mathrm{FN}+2\times\mathrm{TP}}. \end{aligned} \tag{3}$$

Step 1.
map:
  foreach training sample: (C_i, X_j)
    emit $(C_i, 1)$
reduce:
    emit $(C_i, \mathrm{sum}(C_i))$
Step 2.
map:
  foreach training sample: (C_i, X_j)
    foreach feature f_k
      emit $(f_k \mid C_i, 1)$
reduce:
  foreach class
    emit $\left(\frac{f_k \mid C_i}{\mathrm{sum}(C_i)}, \mathrm{sum}(f_k \mid C_i)\right)$
Step 3.
map:
  foreach class C_i
    foreach testing sample: (X_j)
      emit $\left(X_j, P(C_i)\prod P\left(X_j \mid C_i\right)\right)$
reduce:
  emit $\left(X_j, \arg\max\left(P(C_i)\prod P\left(X_j \mid C_i\right)\right)\right)$

Algorithm 1: The workflow of the scalable version of the Naïve Bayesian algorithm implemented by MapReduce.

3. Results

3.1. Optimization of the Prediction Model. The performance of the prediction model using SVM algorithm is determined by the representation of the features, the type of the kernel function, and parameters. Here, the tenfold cross-validation test was conducted to evaluate the model performance. We employed three ways of feature representation, including the linear addition, Zhao's frequent pattern [4], and our frequent pattern. We further used four different types of kernel functions, which are linear, polynomial, Gaussian, and Tanh functions. As shown in Table 1, our proposed frequent pattern performed much better than the other two patterns,

Table 1: Comparison of the accuracy of the prediction models based on SVM using various feature representation and kernel functions.

	Linear	Polynomial	Gaussian	Tanh
Linear addition pattern	47.7%	47.7%	47.7%	53.0%
Zhao's frequent pattern [4]	50.0%	55.1%	57.4%	56.2%
Our frequent pattern	62.2%	64.6%	69.1%	65.4%

Table 2: The performance of the independent test using our definition of frequent pattern and Gaussian kernel.

Run	ACC	SN	SP	F_1
1	67.7%	70.6%	64.3%	0.706
2	65.0%	54.5%	77.8%	0.632
3	60.9%	44.4%	71.4%	0.471
4	64.0%	66.7%	60.0%	0.690
5	68.2%	61.5%	77.8%	0.696
6	65.5%	41.7%	82.4%	0.500
7	77.8%	64.3%	92.3%	0.750
8	72.2%	76.9%	60.0%	0.800
9	72.0%	66.7%	80.0%	0.741
10	70.4%	66.7%	75.0%	0.714
Average	68.4%	61.4%	74.1%	0.670

Table 3: The performance of the one-class SVM classifiers using different kernel functions.

	Linear	Polynomial	Gaussian	Tanh
ACC	46.1%	81.2%	88.2%	80.3%

Table 4: Comparison of the average efficiency between the scalable and sequential version.

Mining steps	Scalable version	Sequential version
Microarray processing	2 h 3 min	6 h 18 m
Feature construction	8 min 34 s	18 min 3 s
Naive Bayesian	15 s	3 s
SVM grid search	27 min 6 s	1 h 11 min

regardless of the types of the kernel functions. The result in Table 1 also suggests that the Gaussian function achieved higher accuracy than the other types of the kernel functions.

3.2. Independent Test. In this section, we evaluated the prediction performance using our proposed frequent pattern and the Gaussian function on the independent test, which is mimicking a true prediction since the model trained on one dataset is used to test on an unseen dataset. We randomly split the whole set of the 76 drug combinations into two datasets (a training set and a testing set). The ratio is about 4 : 1 between the number of the samples of the training set and that of the testing set. The split of the dataset and the independent test is repeated for 10 times. The performance of the 10 runs and their average is presented in Table 2. As shown in Table 2, the model trained by using our proposed frequent pattern performed as well on the independent test, suggesting that our model can predict the unseen data equally well.

3.3. Classification by the One-Class SVM Classifier. In the task of the two-class classification, the assignment of the negative samples (noneffective drug combinations) is not perfect since the unknown pairwise drug combination (we now consider it as noneffective drug combination) may be proved to be an effective drug combination in future. To avoid this problem, we constructed the one-class SVM classifier trained on the dataset with only effective drug combinations. We made use of leave-one-out cross-validation to assess the accuracy of one-class SVM classifiers using different types of kernel functions. As shown in Table 3, without the bias of negative samples, the accuracy of SVM classifiers has significantly increased for polynomial, Gaussian, and hyperbolic tangent kernel, while the linear classifier remains at a lower performance. We have also conducted a test on some nonpositive samples, namely, drug combination that has not yet been approved, and also randomly repeated for 10 times, each testing set containing 76 negative samples. The average result of these 10 repeat experiments suggests that 67.1% of the unknown pairwise samples were predicted as noneffective drug combinations, which is consistent with the fact that there exists a low possibility of the effective drug combinations in the large number of randomly chosen pairs of drugs.

3.4. Extension to a Scalable Mining Process. In this section, we constructed a scalable version of the mining tool for identifying the effective drug combinations and compared its efficiency to that of the sequential implementation by the traditional way. The preprocessing steps (including microarray processing, single drug, and drug combination feature construction) were parallelized by a chain of mappers. The naïve Bayesian algorithm is implemented by a series of MapReduce jobs.

The detailed comparison results of our scalable version and the sequential version in efficiency are listed in Table 4. It is clearly shown in Table 4 that the scalable version achieved higher efficiency in some big data processing steps such as microarray processing, feature construction, and SVM grid search. For naïve Bayesian, the scalable algorithm did not have the advantage against sequential naïve Bayesian, since our final dataset for model construction and evaluation was quite small. However, we believed that the prediction of drug combinations will benefit from our proposed scalable version with the increasing size of the search space of possible drug combinations in future.

4. Conclusions

In this study, we proposed a novel Hadoop-based approach to predict drug combinations by implementing the support vector machine and naïve Bayesian classifiers using the MapReduce programming model, which can advance the improvement of scalability of the prediction algorithm. We believe that our proposed model can be potentially useful when more than two drugs (the increasing availability of the number of the drug combination) are combined for combating the complex diseases in the long run.

Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

Authors' Contribution

Yifan Sun and Yi Xiong contributed equally to this paper.

References

[1] X. Tian and L. Liu, "Drug discovery enters a new era with multi-target intervention strategy," *Chinese Journal of Integrative Medicine*, vol. 18, no. 7, pp. 539–542, 2012.

[2] J. Jia, F. Zhu, X. Ma, Z. W. Cao, Y. X. Li, and Y. Z. Chen, "Mechanisms of drug combinations: interaction and network perspectives," *Nature Reviews Drug Discovery*, vol. 8, no. 2, pp. 111–128, 2009.

[3] G. R. Zimmermann, J. Lehár, and C. T. Keith, "Multi-target therapeutics: when the whole is greater than the sum of the parts," *Drug Discovery Today*, vol. 12, no. 1-2, pp. 34–42, 2007.

[4] X. Zhao, M. Iskar, G. Zeller, M. Kuhn, V. van Noort, and P. Bork, "Prediction of drug combinations by integrating molecular and pharmacological data," *PLoS Computational Biology*, vol. 7, no. 12, Article ID e1002323, 2011.

[5] K. W. Pak, F. Yu, A. Shahangian, G. Cheng, R. Sun, and C. M. Ho, "Closed-loop control of cellular functions using combinatory drugs guided by a stochastic search algorithm," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 13, pp. 5105–5110, 2008.

[6] T. Chou, "Drug combination studies and their synergy quantification using the chou-talalay method," *Cancer Research*, vol. 70, no. 2, pp. 440–446, 2010.

[7] Z. Wu, X. M. Zhao, and L. Chen, "A systems biology approach to identify effective cocktail drugs," *BMC Systems Biology*, vol. 4, article 7, 2010.

[8] Y. Wang, W. Goh, L. Wong, and G. Montana, "Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes," *BMC Bioinformatics*, vol. 14, supplement 16, article S6, 2013.

[9] H. Nordberg, K. Bhatia, K. Wang, and Z. Wang, "BioPig: a Hadoop-based analytic toolkit for large-scale sequence data," *Bioinformatics*, vol. 29, no. 23, pp. 3014–3019, 2013.

[10] R. C. Taylor, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics," *BMC Bioinformatics*, vol. 11, supplement 12, article S1, 2010.

[11] Y. Liu, B. Hu, C. Fu, and X. Chen, "DCDB: drug combination database.," *Bioinformatics*, vol. 26, no. 4, pp. 587–588, 2010.

[12] W. Wang, J. Liu, Y. Xiong, L. Zhu, and X. Zhou, "Analysis and classification of DNA-binding sites in single-stranded and double-stranded DNA-binding proteins using protein information," *IET Systems Biology*, 2014.

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

[14] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucleic acids research*, vol. 31, no. 4, article e15, 2003.

[15] L. Chen, B. Q. Li, M. Y. Zheng, J. Zhang, K. Y. Feng, and Y. D. Cai, "Prediction of effective drug combinations by chemical interaction, protein interaction and target enrichment of KEGG pathways," *BioMed Research International*, vol. 2013, Article ID 723780, 10 pages, 2013.

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

[17] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," LIBSVM software website, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[18] W. Zhang, Y. Xiong, M. Zhao, H. Zou, X. Ye, and J. Liu, "Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature," *BMC Bioinformatics*, vol. 12, article 341, 2011.

[19] Y. Xiong, J. Xia, W. Zhang, and J. Liu, "Exploiting a reduced set of weighted average features to improve prediction of DNA-binding residues from 3D structures," *PLoS ONE*, vol. 6, no. 12, Article ID e28440, 2011.

[20] Y. Xiong, J. Liu, and D. Wei, "An accurate feature-based method for identifying DNA-binding residues on protein surfaces," *Proteins: Structure, Function and Bioinformatics*, vol. 79, no. 2, pp. 509–517, 2011.

[21] J. Xia, X. M. Zhao, J. Song, and D. S. Huang, "APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility," *BMC Bioinformatics*, vol. 11, article no. 174, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 568303, 10 pages
http://dx.doi.org/10.1155/2014/568303

Review Article

The Current Status of Usability Studies of Information Technologies in China: A Systematic Study

Jianbo Lei,[1,2] Lufei Xu,[1] Qun Meng,[3] Jiajie Zhang,[2] and Yang Gong[2]

[1] *Center for Medical Informatics, Peking University, Haidian District, Beijing 100191, China*
[2] *School of Biomedical Informatics, University of Texas Health Sciences Center at Houston, Houston, TX 77030, USA*
[3] *Center for Statistics and Information, National Health and Family Planning Commission, Beijing 100810, China*

Correspondence should be addressed to Qun Meng; mengqun@moh.gov.cn

Received 28 March 2014; Accepted 30 May 2014; Published 19 June 2014

Academic Editor: Bairong Shen

Objectives. To systematically review and analyze the current status and characteristics of usability studies in China in the field of information technology in general and in the field of healthcare in particular. *Methods.* We performed a quantitative literature analysis in three major Chinese academic databases and one English language database using Chinese search terms equivalent to the concept of usability. *Results.* Six hundred forty-seven publications were selected for analysis. We found that in China the literature on usability in the field of information technology began in 1994 and increased thereafter. The usability definitions from ISO 9241-11:1998 and Nielsen (1993) have been widely recognized and cited. Authors who have published several publications are rare. Fourteen journals have a publishing rate over 1%. Only nine publications about HIT were identified. *Discussions.* China's usability research started relatively late. There is a lack of organized research teams and dedicated usability journals. High-impact theoretical studies are scarce. On the application side, no original and systematic research frameworks have been developed. The understanding and definition of usability is not well synchronized with international norms. Besides, usability research in HIT is rare. *Conclusions.* More human and material resources need to be invested in China's usability research, particularly in HIT.

1. Introduction

Usability is essential for the effective, efficient, and safe design, use, and learning of information technology. Research and application of usability have received significant attention by scientists, designers, and industry professionals in Western countries where there are active study populations, comprehensive theories, methods, practices, practical results, and mature industrial and professional organizations. In the field of health information technology (HIT), usability research has been identified as an important cognitive challenge for the adoption and meaningful use of HIT by the Office of National Coordinator for Health Information Technology (ONC), which is part of the U.S. Department of Health and Human Services (DHHS), and has become an active area for research, design, and practice in HIT [1–8]. Methods of usability evaluation have been demonstrated to improve the design and utilization of clinical information systems [9–11]. Usability, under the name of "safety enhanced design," has become a requirement of Stage 2 meaningful use requirement for electronic health records (EHR) in the United States [12]. In China, with 30 years of rapid economic development, China's investment in social development and scientific research has been increasing dramatically. Given this context, one question we would like to answer is as follows. What is the current status of usability research in China? In this paper, through systematically reviewing the literature published in China and applying a combined qualitative and quantitative approach, we present the current status and existing problems of usability research and practice in China's information technology field.

2. Materials and Methods

2.1. Literature Search Strategy

2.1.1. Chinese Publication Search Strategy. Usability is a broad and interdisciplinary field with inconsistent terminologies.

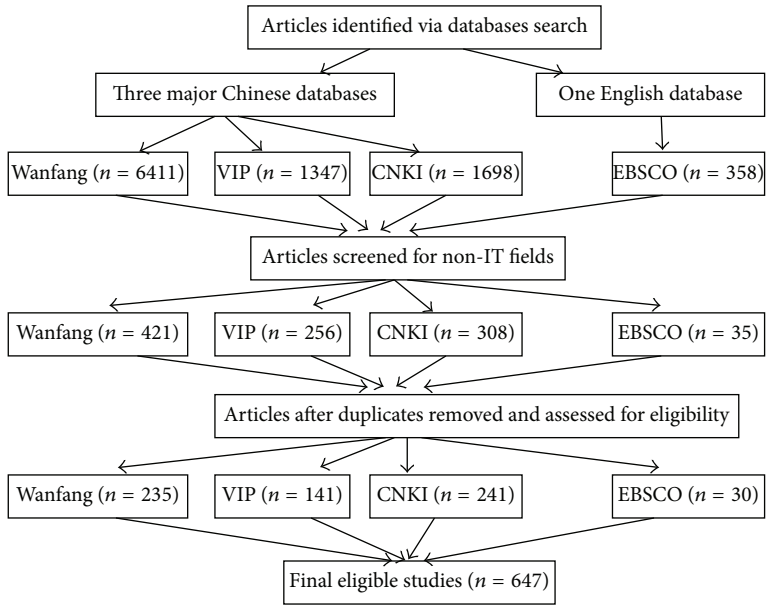


Figure 1: The PRISMA (preferred reporting items for systematic reviews and meta-analyses) flow chart.

This problem is compounded when translating these terms into Chinese because the words often have multiple meanings. After carefully reviewing the literature in August 2012, we decided to conduct a literature review using the Chinese equivalents of the following search terms: "usefulness," "usability," "user experience," "user satisfaction," and "user-centered design." We used the advance search functions of following three mainstream databases in China: China National Knowledge Infrastructure (cnki.net), China Science and Technology Periodical Database (cqvip.com), and Wanfang Electronic Journal Database (wanfangdata.com.cn). We only searched the "title" and "keyword" fields in the literature published from the time period between 1980 and 2012.

2.1.2. English Publication Search Strategy. Although our research was to review the literature originating from China using Chinese academic databases, we used the English language database, EBSCO, because it had the advanced search function that allowed us to select papers that originated from China. We searched using the advanced functionality provided by EBSCO, in the fields of "title" and "abstract" with any of the search terms "usability," "user experience," "user-centered design," "user satisfaction," "customer satisfaction," "user interface," "UCD," and "China" and limited the results to those originating in China.

2.2. Literature Inclusion and Exclusion Criteria

2.2.1. Inclusion Criteria. Inclusion criteria include (1) screening the published literature on usability in the field of HIT in China.

2.2.2. Exclusion Criteria. Exclusion criteria include (1) excluding research literature about usability in other fields; (2) for duplicate entries, excluding those without complete information; and (3) excluding those without full text.

2.3. Data Extraction and Statistical Processing. The information summary sheet was designed to extract data from the selected literature. Information extracted included author, article title, year, journals, subcategory types for theoretical research, subcategory type for empirical studies, definition of usability, evaluation objects in empirical study, subcategory type of evaluation research in network application, and evaluation method used in empirical study. See Figure 1 for the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow chart.

3. Results

3.1. Literature Search Results. Through searching of the literature for thirty-two years in the three databases and EBSCO, we obtained a total of 9814 publications. After applying our inclusion and exclusion criteria, 617 Chinese and 30 English publications were included in our analysis and accounted for 6.6% of the total retrieved literature. Reasons for the high exclusion rate are as follows.

(1) Repeated retrievals: it is common that the same article may be retrieved in all three databases.

(2) Study fields involved in the retrievals are unrelated. For example, topics on architecture, transportation, and so on rather than "information technology" were excluded.

(3) Limitation of search function by certain database: for instance, the Wanfang database returned 6411 items, far more than 1347 and 1698 items from VIP and CNKI databases. The reason is that Wanfang is

Table 1: Inclusion and exclusion process.

Databases	Wanfang database	VIP	CNKI	EBSCO	Total
Retrievals from database	6411	1347	1698	358	9814
After exclusion of repeated and irrelevant	235	141	241	30	647
screening rate	3.7%	10.5%	14.2%	8.8%	6.6%

Table 2: General author information.

Author	Institutional affiliation	Number of publications	Percentage
Liu, Zhenjie	EU Usability Chinese Center, Dalian Maritime University	12	1.9%
Qiu, Minghui	Consulting and Management Department, Sun Yat-sen University	8	1.2%
Ge, Liezhong	Psychological Department, Zhejiang University of Technology	8	1.2%
Zhang, Kan	Psychological Institute, Chinese Academy of Sciences	7	1.1%
Rob Law	Hong Kong Institute of Technologies	7	1.1%
Zhang, Liping	EU Usability Chinese Center, Dalian Maritime University	6	0.9%
He, Guihe	School of Economics and Management, Jingchu University of Technology	6	0.9%
Sun, Qingzhen	Zhengzhou Institute of Aeronautical Industry Management	6	0.9%
Huang, Xiaobin	Consulting and Management Department, Sun Yat-sen University	6	0.9%
Ren, Zhongbin	Institute of Surveying and Mapping, Information Engineering University	6	0.9%

Note: quantities of published literature of other authors were all less than 0.5% and are not listed here.

unique in the way that it does not support whole word search; rather, it will return all results containing each composite word of a complete word. In Chinese usability is composed of three "words."

(4) Excluded nontechnical items: for example, some items retrieved were advertisements released by companies for new products or news published on nontechnical magazines. The inclusion and exclusion process and the results are listed in Table 1.

3.2. General Characteristics of the Literature

3.2.1. Year Distribution of the Literature. Publication dates of the 647 publications span from 1994 to 2012. The distribution across the years is shown in Figure 2.

3.2.2. Type Distribution of the Literature. In the 647 items incorporated into our study, 564 are periodical publications and account for 87.2% of the total items, 52 are conference proceedings and account for 8.0%, 30 are theses/dissertations and account for 4.6%, and one was a book chapter accounting for 0.2%.

3.2.3. Author Distribution of the Literature. Among the 1190 authors (including the second and the third coauthors) that contributed to the selected research publications, the following 10 authors listed in Table 2 published the most publications.

3.2.4. Journal Distribution of Published Literature. Journals or conference proceedings in the 647 publications (top 14 with the highest quantity) are listed in Table 3.

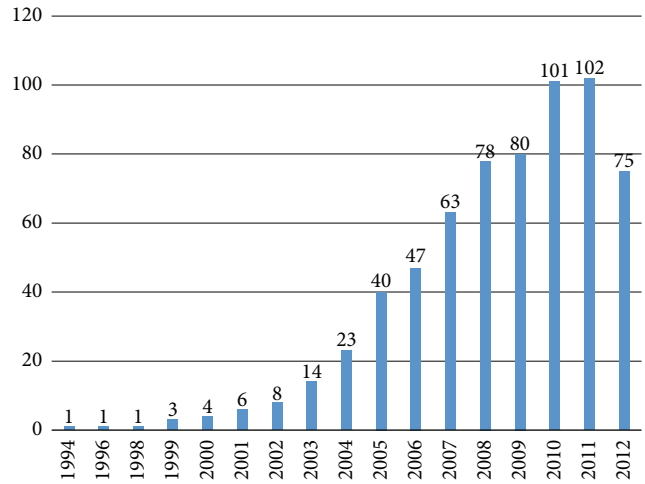


Figure 2: Distribution of the literature publications by year. Note: the abscissa represents the publication year and the ordinate represents the number of publications.

3.2.5. Distribution of Keywords Involved in the Literature. There are a total of 2229 keywords in the 647 publications. The keywords with the top 8 occurrence frequencies are listed in Table 4.

3.3. Results of Usability Research in China's Information Technology Field

3.3.1. Types of Domestic Usability Research in Information Technology Field. Usability research in China may be roughly divided into two categories. The first category is the study about usability theories; there are a total of 395 publications

Table 3: List of journals and conference proceedings.

Journal or conference name	Number of publications	Percentage	Core journals*
Ergonomics	23	3.6%	Peking University core and Technology core
Information science	13	2.0%	Peking University core
Library and information service	13	2.0%	Peking University core
Packaging engineering	12	1.9%	Peking University core
Modern library and information technology	12	1.9%	Peking University core
Art and design	12	1.9%	
Library studies	10	1.5%	Peking University core
Programmer	9	1.4%	
Intelligence theory and practice	9	1.4%	Peking University core
Computer engineering and applications	8	1.2%	Peking University core and Technology core
Market modernization	8	1.2%	Peking University core
Computer engineering and design	7	1.1%	Peking University core and Technology core
Computer science	7	1.1%	Peking University core and Technology core
Modern information	7	1.1%	Peking University core

Note: the publication quantity of other journals or conference proceedings is less than 1% and is not listed here. *aka "Peking University core journal" refers to the classification by Peking University Library on Chinese academic journals, published every 3-4 years and currently widely recognized by the Chinese academia. Publications in core journals are viewed with relative high academic levels and this is an important part of the academic evaluation system in China.

Table 4: Information about keywords in literatures.

Keywords	Publication usage frequency	Percentage
Usefulness	140	6.3%
User experience	84	3.8%
User satisfaction	28	1.3%
High usability	26	1.2%
Usability assessment	23	1.0%
Usability test	23	1.0%
Usability	21	0.9%
Website	21	0.9%

Note: occurring percentage of other keywords less than 0.9% is not listed here.

accounting for 61% of the total retrieved publications. The other category includes empirical studies of usability, of which qualitative research methods were primarily utilized. The theoretical study mainly includes the following three aspects in contents: (a) usability history of development, influencing factors and problems encountered in usability (in different branch fields); (b) usability evaluation methods (in different branch fields); and (c) usability design principles and design concepts (in different branch fields).

Empirical studies related to usability accounted for 39% (i.e., 252) of the retrieved publications. Those publications related to the integration of usability during software or technology development account for 17% (i.e., 42) of the total retrieved publications. Most studies (83% or 210 total) focused on usability evaluations in specific target areas through the selected evaluation methods. The evaluation methods mostly applied are the combination of qualitative and quantitative studies.

The categorization of theoretical studies is shown in Figure 3. Classification of empirical studies is shown in Figure 4.

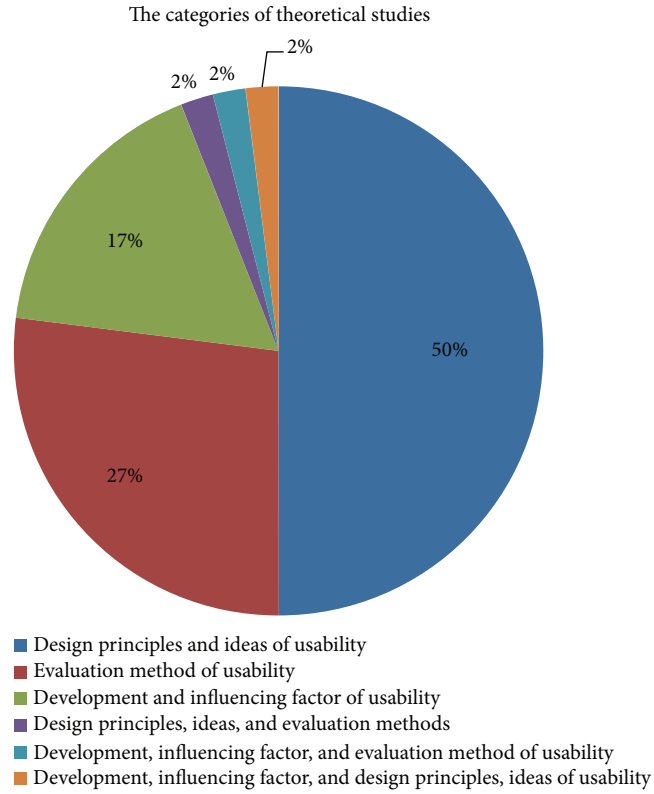


Figure 3: Classification of theoretical studies.

3.3.2. Understanding of Chinese Researchers on the Concept of Usability. Based on the retrieved publications, the usability concepts in Table 5 are more commonly recognized by Chineseresearchers.

Table 5: List of usability concepts.

Usability definition	Mentioning rate of the definitions	Percentage
Definition given in ISO9241-11 "Ergonomic requirements for office work with visual display terminals (VDTs)"	151	36.8%
Definition given by Nielsen in 1993	120	29.3%
Other definitions (definitions in various branch fields by combining with specific study contents in the field)	120	29.3%
Definition given in ISO9126-1:2000 "Software Product Evaluation: Quality Characteristics and Guidelines for their Use-standard"	14	3.4%
Definition given in China national standard GB/T 162602006 "Software engineering products quality"	5	1.2%
Total	410	100%

Note: among the 647 articles, 282 did not provide definite usability definition. Among the articles providing usability definitions, an article may provide more than one definition.

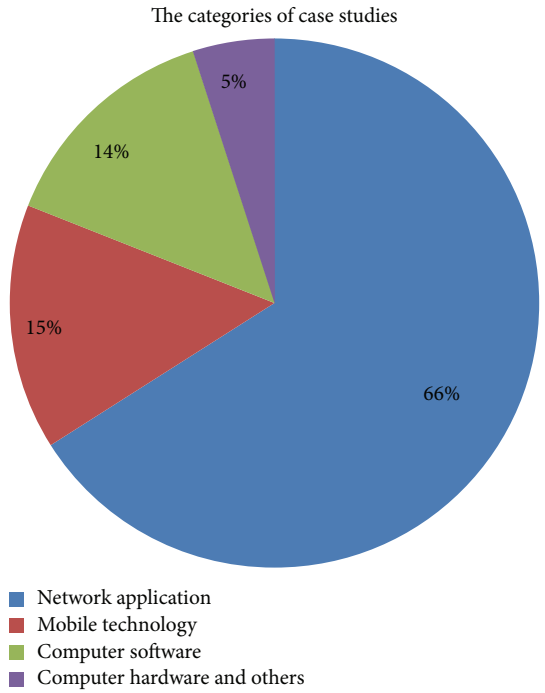


Figure 4: Classification of empirical case studies. Note: study targets in network application mainly include websites and network services; mobile study targets mainly include the interface design and applications of mobile phones and other mobile terminals.

Specific definitions are as follows.

(i) In the China national standard GB/T16260-2006 "Software engineering products quality," usability is defined as "the ability of a software product to be understood, studied, used and as well as [sic] the ability to attract users in a particular use environment."

(ii) In the usability definition given by Nielsen in 1993 [13], usability includes 5 aspects, which are learnability, efficiency, memorability, errors, and satisfaction, respectively.

(iii) In the ISO 9241-11:1998 [14] "Ergonomic requirements for office work with visual display terminals (VDTs)," usability is defined as "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use."

(iv) In the ISO/IEC 9126-1:2001 [15] "Software products evaluation-quality properties and operation directions," usability is defined as "the capability of the software product to be understood, learned and liked by the user, when used under specified conditions."

(v) In recent years, with the advancement of usability research, new terminology emerges constantly. In the 647 publications, the "user experience, user satisfaction, and user-centered design" are also mentioned; among them, "user experience" is mentioned 118 times, "user satisfaction" 25 times, and "user-centered design" 12 times.

3.3.3. Specific Types of Network Study Targets in Empirical Studies for Usability Evaluation. Through analyzing the retrieved publications, we found that the Chinese empirical studies on usability focused mainly on website application (138 publications; see Figure 5 for the breakdown of website types).

Regarding the trend of evaluation objects, we can see from Table 6 that usability evaluation studies in web application still have a dominant position in recent years. However, usability evaluation studies on mobile Internet (including mobile phones) and applications did not increase as would be expected with the current rapid growth of mobile technology.

3.3.4. Types of Study Methods in Empirical Studies for Usability Evaluation. Through analyzing the retrieved publications, we found that the usability evaluation methods used mainly include questionnaires, usability tests, heuristic evaluation, usability guidelines (such as MUG: Microsoft Usability Guideline), statistical analyses through system logs, cognitive walkthrough, behavior analyses, observation and interviews, eye movement analyses, distance of information-state transition (DIT), and other methods. The specific applications of these methods in the empirical studies are listed in Table 7.

Table 6: Distribution of evaluation objects by years.

Year	Classification of evaluation targets in usability evaluation study				Total (literature quantity)
	Web application	Mobile technology	Computer applications	Computer hardware	
1996	0	0	1	0	1
2002	0	0	2	0	2
2003	2	0	1	0	3
2004	3	0	2	2	7
2005	4	3	1	2	10
2006	15	3	0	0	18
2007	19	3	4	3	29
2008	23	6	2	1	32
2009	14	4	4	0	22
2010	23	3	6	1	33
2011	21	5	3	1	30
2012	14	4	4	1	23
Total (literature quantity)	138	31	30	11	210

Table 7: List of usability evaluation methods.

Usability evaluation method	Application times in study	Percentage
Questionnaires	90	30.8%
User/researcher usability test	80	27.4%
Following existing guidelines (such as MUG)	61	20.9%
Eye movement analysis and DIT theory and other methods	33	11.3%
Observation and interviews	12	4.1%
Heuristic evaluation	11	3.8%
Statistical analysis through system log files	4	1.4%
Cognitive walkthrough	1	0.3%
Total	292	100%

Note: multiple evaluation methods may be used in the same study.

In usability evaluation research, some studies may use two or more evaluation methods to obtain a greater understanding of the system's usability. Table 8 provides a breakdown of the number of research methods used in each article.

3.3.5. Studies in the Field of Healthcare Information Technologies. Of special note is that among the 647 publications only nine are about usability as it relates to HIT (specific information about the literatures is listed in Table 10).

Table 8: Combination uses of evaluation methods.

Use of evaluation methods	Quantity of literatures involved	Percentage
Single evaluation method	106	58.2%
Two evaluation methods	66	36.2%
Three or more evaluation methods	10	5.6%
Total	182	100%

4. Discussion

(1) From our analyses of the publication dates and publication quantities, it is evident that China's literature on usability research in the information technology field began in 1994. The quantity of publications has been increasing year by year, starting with one publication identified in 1994 and culminating in a total of 102 publications in 2011. Publications released from 2010 to August 2012 account for 43% of the total publications. Our data show that usability research in China's information technology field started relatively late and the history is not long, but it is attracting more attention.

(2) From the data on the authors and the publishing journals, we conclude that the targets of usability research in China's information technology field are relatively scattered and no coordinated usability efforts, such as research institutes or centers, have been established.

The most prolific author with the largest number of publications has only published 12 publications, and the top 10 authors with the most publications account for 11% (including the coauthors) of the total publications printed. The fields of the main publications

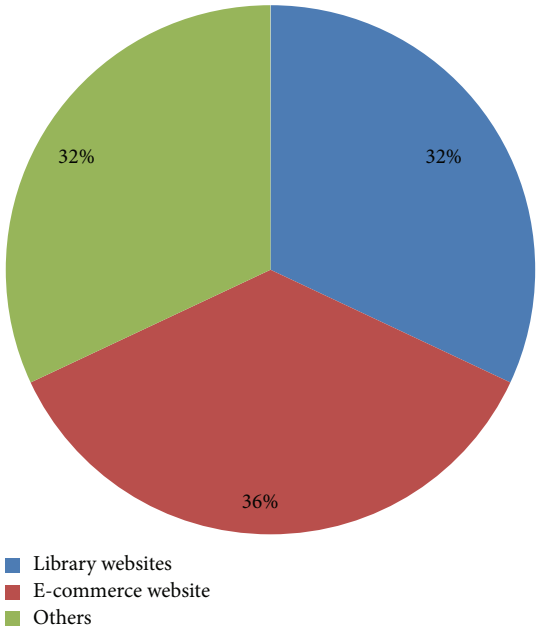


Figure 5: Distribution of study targets in web application study. Note: other targets mainly include other types of websites (government websites, various major portals, etc.), search engines, and network services.

in which the publications are produced are relatively simple and centralized. The top 11 journals that publish the most literature are mostly library information journals. The publication quantity distribution shows that the top 14 journals published 150 publications about usability research, accounting for 23.3% of the total publications. This shows that, in the information technology usability research field, articles are published across a diverse spectrum of journals and there is no journal with a major focus on usability. Of course, these journals are respectively included in the "China Journal Citation Report" by the China Science and Technology Information Institute and the "Chinese core journals" of the Peking University, which are believed to have a good reputation. In the United States, there are many pioneers and authorities in the field of usability, such as Jakob Nielsen, who have produced many publications, but China has yet to have such established and prolific experts on these topics. What is more, besides specific peer reviewed journals for usability, such as "Journal of Usability Studies", there are more than 50 types of journals/magazines which are directly associated with usability.

(3) The data on study types show that theoretical studies on usability account for 61% of the total publications, including 27% about usability evaluation methods. Through analyzing the retrieved literature, we found that theoretical studies typically combine usability with specific branch fields (such as study of usability in web application) in terms of usability influencing factors, usability problems, usability design principles, and usability evaluation methods. However, the theoretical studies typically do not have significant original proposals or innovative methods and frameworks. At present, major advancements in theoretical and methodological studies are lacking in China.

(4) From the data on the understanding on usability concepts, we found that the Nielsen [13] and ISO 9241-11:1998 [14] definitions account for 66.1% of the publications, showing that these two definitions are widely recognized by Chinese researchers. However, the above two definitions were proposed in earlier years. Based on the newest usability definition provided in ISO 25010:2011, "System and software quality model", usability is not only a property about product quality but also a property about quality in use of the project (comprising effectiveness, efficiency, and satisfaction) [16]. Yet, this newest definition is not mentioned in any current Chinese usability literature, so, to a certain extent, this may demonstrate that Chinese researchers are falling behind on international usability research. A new definition of usability was just proposed last year by Zhang and Walji [5], with the intent to unify all the variations of usability definitions, concepts, and applications under a single theoretical framework.

(5) From the data on study targets, we can see that most Chinese usability studies are combined with evaluation studies and the evaluation objects are mainly focused on Internet applications. In evaluation studies, 66% are usability studies in Internet applications; this is partially because usability studies on Internet started relatively early in western countries. Many usability experts in web applications have proposed various website usability evaluation methods and practice guidelines. For example, in the US, Story argues that a website developer should follow 10 usability principles when the site is designed [17]. The Northwest Alliance for Computational Science and Engineering (NACES) formulated common website usability guidelines for website design, webpage design, and navigation help [18]. Borges et al. from University of Puerto Rico also proposed 16 usability principles for web design and proved the effectiveness of these principles by experiments [19]. Nielsen, a pioneer in usability research, conducted many important studies on usability of websites, addressing theories, methods, practice, and other aspects of usability [13, 20, 21].

Meanwhile, with the development of digital multimedia technology and wireless network technology, evaluation objects in usability evaluation studies are also changing gradually.

(6) From the data on usability methods we can see that evaluation methods are divided into two categories: usability testing and questionnaires. They account for 58.2% of all evaluation methods. In most cases, a single method is applied in usability evaluation and this

Table 9: Specific combinations between evaluation methods.

Combination use of evaluation methods	Number of publications involved	Percentage in literature about evaluation study
Questionnaires and following existing guidelines (such as MUG)	29	13.8%
Questionnaires and user/researcher usability testing	15	7.1%
User/researcher usability testing and observation/interviews	5	2.4%
User/researcher usability testing and following existing guidelines (such as MUG)	5	2.4%
User/researcher usability testing and heuristic evaluation	4	1.9%
Questionnaires and user/researcher usability testing and observation/interviews	4	1.9%
User/researcher usability testing and eye movement analysis and DIT theory	3	1.4%
Questionnaires and heuristic evaluation	2	1.0%
Heuristic evaluation and following existing guidelines (such as MUG)	1	0.5%
User/researcher usability testing and heuristic evaluation and cognitive walkthrough	1	0.5%
Questionnaires and user/researcher usability testing and guidelines (such as MUG)	1	0.5%
Questionnaires and user/researcher usability testing and heuristic evaluation	1	0.5%
Questionnaires and user/researcher usability testing and following existing guidelines (such as MUG) and statistical analysis through system log files.	1	0.5%
Questionnaires and heuristic evaluation and observation/interviews	1	0.5%
Following existing guidelines (such as MUG) and eye movement analysis and DIT theory and other methods	1	0.5%
Following existing guidelines (such as MUG) and statistical analysis through system log files.	1	0.5%
User/research usability testing and following existing guidelines (such as MUG) and eye movement analysis and DIT theory	1	0.5%
Total	76	36.2%

accounts for 63.8% of the evaluation methods. The data on combination of different evaluation methods are listed in Table 9. Obviously, usability testing and questionnaires, as two prominent methods, appear to play an important role in usability research. Although each evaluation method has its own use conditions, combining multiple methods may evaluate usability from a more comprehensive perspective. Combination of multiple methods will likely become a development trend for use of evaluation methods in the future.

(7) Further analysis of the data on distribution of study targets shows that Chinese usability studies are mostly concentrated on web information technology and usability studies in the healthcare information technology field are quite limited. Among the 647 publications, only nine are about this field (specific information about the literatures is listed in Table 10). Comparatively, as of 25 November 2012, the biomedical database PubMed returned about 942 publications (search from titles) and 4861 publications (search from abstracts) preliminary search results using the similar combination of query terms: "usability or user experience/s or user centered design or user satisfaction." Compared with the rapid development of the HIT industry in China, usability research in HIT in China is very underdeveloped. Many studies have shown that usability improvement of HIT could effectively reduce medical errors [20], thus improving patient quality. Obviously, it is both important and urgent to carry out usability research in China's HIT field.

5. Study Limitations

This paper studies the current status and characteristics of usability research in China's information technology field by using the systematic review and a quantitative literature analysis. Limitations of the study are listed as follows. (1) Usability research is an interdisciplinary field and researchers in different disciplines often use different terminologies. To minimize any effects of overrepresentation, we used many different keywords such as "usability," "user satisfaction," "user experience," and "user-centered design" to query the related literature. (2) Three major databases are used in our study and we searched via the field of keywords. However, keywords in Chinese publications do not have corresponding vocabulary similar to Mesh, and all keywords in Chinese publications are manually added by the author. We had to assume that if the authors had considered the subject of the article to be mainly about usability, they would have used one of the above related usability terms in the keywords, especially given the frequent references to Nielsen [13] and ISO 9241-11:1998 [14]. (3) This paper is limited to the information technology field, but typical publications may not

Table 10: Summary about literatures involving usability study in the health information technology.

Literature title	Year	Type	Contents studied
Telehealth for older patients: the Hong Kong experience	2002	Empirical study	Evaluation study about usability
Maintaining high usability of database, ensuring stable operation of hospital information systems	2005	Theoretical study	Usability design principles
Usability design study on human-machine interface of medical equipment	2007	Empirical study	Usability-oriented system software or technology development
Design on high usability of hospital information systems	2008	Empirical study	Usability-oriented system software or technology development
Study on user experience testing of China Disease Prevention and Control Center website in 2009	2010	Empirical study	Evaluation study about usability
Practice and improvement of clinic HIS high usability programs	2011	Theoretical study	Usability development and influence factors
Achieving high reliability and high usability of regional health information system database through ESX4	2012	Empirical study	Usability-oriented system software or technology development
Using recommendation to support adaptive clinical pathways	2012	Empirical study	Evaluation study about usability
A mobile nursing information system based on human-computer interaction design for improving quality of nursing	2012	Empirical study	Usability-oriented system software or technology development

explicitly use the keywords such as information technology in the titles or keywords; thus, we cannot enter "information technology" in the search query. We could only manually screen publications about information technology after all usability related publications are retrieved; this process needs more manual efforts. (4) Finally, this study focuses on the academic literature only; thus, the results obtained here are not inclusive. Furthermore, as usability is also an application intensive discipline, it is possible that usability related efforts are more active in industrial society than in academic domain reflected from this research.

Conflict of Interests

The authors declare that they have no competing interests.

Authors' Contribution

Yang Gong and Jianbo Lei developed the conceptual framework and templates for the literature review and guided Lufei Xu in the management and quantitative analysis of the literature review. Jianbo Lei drafted the paper and Yang Gong made revisions. Jiajie Zhang and Qun Meng supervised the study and provided comments. All authors read and approved the final paper.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (NSFC) Grant no. 81171426. The authors would like to thank Dr. Timothy McEwen for his comments and revisions on the paper.

References

[1] Y. Y. Han, J. A. Carcillo, S. T. Venkataraman et al., "Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system," *Pediatrics*, vol. 116, no. 6, pp. 1506–1512, 2005.

[2] R. Koppel, J. P. Metlay, A. Cohen et al., "Role of computerized physician order entry systems in facilitating medication errors," *Journal of the American Medical Association*, vol. 293, no. 10, pp. 1197–1203, 2005.

[3] S. Z. Lowry, M. Ramaiah, D. Brick et al., *A Human Factors Guide to Enhance EHR Usability of Critical User Interactions When Supporting Pediatric Patient Care*, National Institute of Standards and Technology, 2012.

[4] G. Southon, C. Sauer, and K. Dampney, "Lessons from a failed information systems initiative: issues for complex organisations," *International Journal of Medical Informatics*, vol. 55, no. 1, pp. 33–46, 1999.

[5] J. Zhang and M. F. Walji, "TURF: toward a unified framework of EHR usability," *Journal of Biomedical Informatics*, vol. 44, no. 6, pp. 1056–1067, 2011.

[6] D. Armijo, C. McDonne, and K. Werner, Eds., *Electronic Health Record Usability: Evaluation and Use Case Framework*, Agency for Healthcare Research and Quality, 2009.

[7] D. Armijo, C. McDonne, and K. Werner, Eds., *Electronic Health Record Usability: Interface Design Considerations*, Agency for Healthcare Research and Quality, 2009.

[8] W. Stead and H. Lin, Eds., *Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions*, Committee on Engaging the Computer Science Research Community in Health Care Informatics, National Research Council, Washington, DC, USA, 2009.

[9] D. W. Bates and A. A. Gawande, "Improving safety with information technology," *The New England Journal of Medicine*, vol. 348, no. 25, pp. 2526–2534, 2003.

[10] D. W. Bates, L. L. Leape, D. J. Cullen et al., "Effect of computerized physician order entry and a team intervention on prevention of serious medication errors," *Journal of the American Medical Association*, vol. 280, no. 15, pp. 1311–1316, 1998.

[11] A. C. Li, J. L. Kannry, A. Kushniruk et al., "Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support," *International Journal of Medical Informatics*, vol. 81, no. 11, pp. 761–772, 2012.

[12] ONC, 2014 Test procedure for, "Safety-enhanced design", September, 2012, http://www.healthit.gov/sites/default/files/standards-certification/2014-edition-draft-test-procedures/170-314-g-3-safety-enhanced-design-2014-test-procedures-draft-v-1.0.pdf.

[13] J. Nielsen, *Usability Engineering*, Academic Press, Boston, Mass, USA, 1993.

[14] ISO 9241-11:1998, "Ergonomic requirements for office work with visual display terminals (VDTs)—part 11: guidance on usability," December 2012, http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=16883.

[15] ISO/IEC 9126-1:2001, "Software engineering—product quality—part 1: quality model," October 2012, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=22749.

[16] ISO/IEC 25010:2011, "Systems and software engineering—systems and software quality requirements and evaluation (SQuaRE)—system and software quality models," October 2012, http://www.iso.org/iso/catalogue_detail.htm?csnumber=35733.

[17] D. Story, "Usability Checklist for Site Developer," October 2012, http://www.drdobbs.com/usability-checklist-for-site-developers/184412660.

[18] Northwest Alliance for Computational Science and Engineering (NACES), "Web usability guide," October 2012, http://www.nacse.org/home/usability/usability_guide/index.html.

[19] J. A. Borges, I. Morales, and N. J. Rodriguez, "Guidelines for designing usable World Wide Web pages," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '96)*, pp. 277–278, April 1996.

[20] J. Nielsen, "A mathematical model of the finding of usability problems," in *Proceedings of the ACM (INTERCHI '93) Conference*, pp. 206–213, Amsterdam, The Netherlands, April 1993.

[21] J. Nielsen, "User interface directions for the web," *Communications of the ACM*, vol. 42, no. 1, pp. 65–72, 1999.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 304563, 13 pages
http://dx.doi.org/10.1155/2014/304563

Research Article

Metadynamics Simulation Study on the Conformational Transformation of HhaI Methyltransferase: An Induced-Fit Base-Flipping Hypothesis

Lu Jin,[1,2] Fei Ye,[1] Dan Zhao,[3] Shijie Chen,[3] Kongkai Zhu,[3] Mingyue Zheng,[3] Ren-Wang Jiang,[2] Hualiang Jiang,[3] and Cheng Luo[3]

[1] *College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China*
[2] *Institute of Traditional Chinese Medicine and Natural Products, Jinan University, Guangzhou 510632, China*
[3] *Drug Design and Discovery Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China*

Correspondence should be addressed to Cheng Luo; cluo@mail.shcnc.ac.cn

Received 31 March 2014; Accepted 12 May 2014; Published 19 June 2014

Academic Editor: Junfeng Xia

DNA methyltransferases play crucial roles in establishing and maintenance of DNA methylation, which is an important epigenetic mark. Flipping the target cytosine out of the DNA helical stack and into the active site of protein provides DNA methyltransferases with an opportunity to access and modify the genetic information hidden in DNA. To investigate the conversion process of base flipping in the HhaI methyltransferase (M.HhaI), we performed different molecular simulation approaches on M.HhaI-DNA-S-adenosylhomocysteine ternary complex. The results demonstrate that the nonspecific binding of DNA to M.HhaI is initially induced by electrostatic interactions. Differences in chemical environment between the major and minor grooves determine the orientation of DNA. Gln237 at the target recognition loop recognizes the GCGC base pair from the major groove side by hydrogen bonds. In addition, catalytic loop motion is a key factor during this process. Our study indicates that base flipping is likely to be an "induced-fit" process. This study provides a solid foundation for future studies on the discovery and development of mechanism-based DNA methyltransferases regulators.

1. Introduction

DNA methylation at the position 5 of cytosine, which is closely related to development and differentiation, genome stability, genomic imprinting, X-chromosome inactivation, and silencing of retrotransposons [1–4], is commonly found in bacteria, plants, and mammalians. Hypermethylation of specific genes is found to be closely related to many malignant diseases [5]. DNA methylation is catalyzed by DNA methyltransferases (DNMTs), which have been identified in at least 16 kinds of bacterial DNMTs [6] and 3 kinds of mammalian ones. Crystal structures of different DNMTs [7, 8] show that the catalytic domains of these methyltransferases are relatively conserved. Recent studies demonstrate that these enzymes also share a similar catalytic mechanism.

HhaI methyltransferase (M.HhaI) belongs to restriction-modification systems of bacterial DNMTs [9] and methylates certain CpG sequences specifically. To access the target base and modify the genetic information, M.HhaI flips the target base out of the DNA double helix during the catalytic process. Base flipping was first discovered by Cheng et al. in the cocrystal structure of cytosine-5 DNA methyltransferase binding to DNA [8]. Structures of M.HhaI can be divided into three parts: a large domain (residues 1–193 and 304–327), a small domain (residues 194–275), and a hinge region (residues 276–303) [10–12] (Figure 1). The target recognition domain (TRD) is located in the small domain and plays an important role in recognizing cognate GCGC base pairs. The catalytic loop (residues 81–100), a very flexible motif in the large domain, is located opposite the TRD. Based on

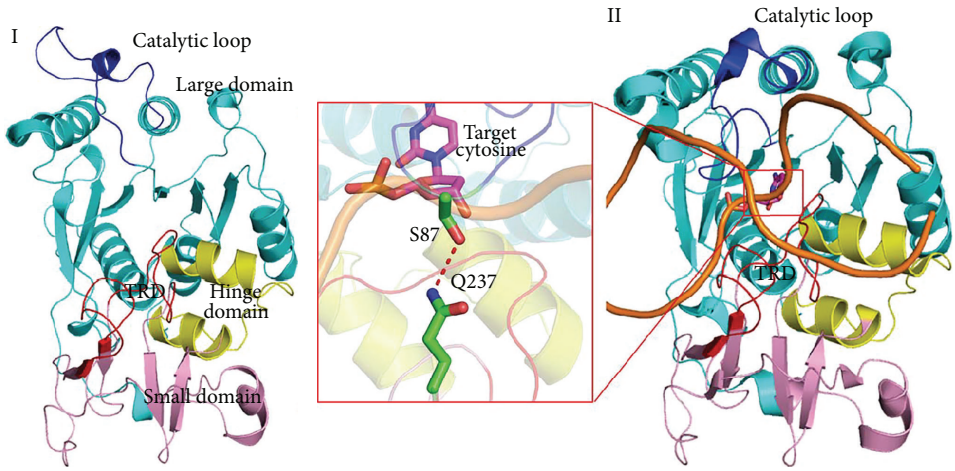


Figure 1: Two conformations of M.HhaI. (I) The structure of the M.HhaI-SAM binary complex (PDBID:2HMY) shows the inactive status of M.HhaI. The large, hinge, and small domains are colored marine cyan, yellow, and pink. (II) Structure of the M.HhaI-SAH-DNA ternary complex (PDBID:2HR1) and the active status of this enzyme. Carbon atoms of target cytosine are colored magenta. The target recognition domain (TRD) and catalytic loop in both structures are colored red and blue, respectively. Flipped cytosine, Gln237, and Ser87 are shown using sticks.

the conformations of the catalytic loop, these structures can be classified into two distinct classes: the catalytic inactive (Figure 1(I)) and catalytic active (Figure 1(II)) forms. Various assays, including fluorescence [13], NMR [14], and molecular dynamics simulation, have been used to investigate M.HhaI, and valuable information on different aspects of the base flipping process has been obtained. (1) M.HhaI binds to the nonspecific DNA or cognate DNA (DNA containing GCGC nucleotides) with different affinities [15, 16], and the DNA sequence plays an important role in substrate recognition and conformational transition of the catalytic loop. (2) Base flipping is involved in extensive protein conformational changes, including closure of the catalytic loop (residues 81–100 of M.HhaI) [17], target base flipping, and correct assembly of the active site [18]. (3) The target base preferably rotates out of the double helix through a major groove path [19] because of interactions between the HhaI methyltransferase and the backbone of the cognate DNA [20]. However, these studies mainly focused on flipped base and surrounding residues in active state; these states may not adequately describe nonspecific binding pattern and following sequence recognition process. Thus, understanding the detailed process of structural rearrangement of catalytic loop and the relationships between target sequence recognition and catalytic loop reorganization remains challenging.

To gain new insights into the conformational transition of M.HhaI, we performed a mechanistic investigation on the dynamic transition of this enzyme using a combination of molecular docking, conventional molecular dynamics (MD) simulation, and metadynamics simulation. The DNA-M.HhaI- (open form-) S-adenosylhomocysteine (SAH) ternary complex built by protein-DNA docking model is used as the starting structure and then optimized by conventional MD simulation. Subsequently, metadynamics simulation is employed to monitor the motion of catalytic loop. The results show that DNA binds to a shallow pocket close to the catalytic loop before it falls into a cleft between the TRD and the catalytic loop. DNA binds to this nonspecific binding site and evokes the conformational change of residues at the tip of this motif. Target recognition loops I and II detect the target DNA and facilitate target base flipping by destabilizing hydrogen bonds between base pairs. Our study shows the nonspecific binding patterns of DNA, sequence recognition process of M.HhaI, and conformational reorganization of the catalytic loop. We propose that DNA evokes the conformational change of M.HhaI, which then selects the target cytosine to fit into its catalytic pocket actively. Understanding the mechanism of DNA recognition process in base flipping at the atomic level is of great help to future researchers. This study explains DNA recognition process in atomic detail and will aid the future discovery and development of mechanism-based DNMT regulators.

2. Materials and Methods

2.1. Starting Structure for Simulation. Crystal structures of the binary complex (PDB ID: 2HMY [7]) were used to construct the protein model, in which all water molecules were removed. For ligands in the crystal, SAM was converted into SAH by simply removing the methyl group attached to the sulfur atom, whereas solvents and other molecules were deleted. DNA used in this simulation was generated by the 3D-DART server [21], and the sequence employed was identical to that in the M.HhaI-DNA-SAH complex (PDBID:2HR1). To place a piece of cognate DNA in its "nonspecific" site, protein-DNA docking was employed. Docking was performed on the PatchDock web server [22]: the prepared protein was used as the receptor and the B-form DNA generated by 3D-DART was used as the ligand.

Little is known about nonspecific binding sites and the binding poses of M.HhaI. To find an appropriate starting

point, biomolecular docking was employed. This method is commonly used to gain structural insights into macromolecule structures that X-ray crystallography or NMR spectroscopy cannot elucidate [23]. Patchdock [24, 25] is a geometry-based molecular docking algorithm that can be used for protein-protein, protein-ligand, or protein-DNA docking. We docked B-form DNA into proteins through the Patchdock web server. Fifty different poses were downloaded from the server. We then separated these poses into two categories: (1) DNA approaching the TRD and (2) DNA approaching the catalytic loop. Combining the structures of M.HhaI at different states and the NMR experiment results (see Supporting Information, available online at http://dx.doi.org/10.1155/2014/304563), we chose the top scored poses in category two as our initial structures.

2.2. Conventional MD Simulation. This initial structural model was prepared using Charmm27 all-atom force field by pdfgen. Then, the ternary complex was embedded into an explicit TIP3P water molecule box with 10 Å widths. 61 Na^+ and 39 Cl^- ions were added to this box to ensure charge neutrality. Finally, the concentration of NaCl was adjusted to 0.11 mM by the Autoionization plug-in (version 1.3).

The system described above of ~62,000 total atoms underwent 5,000 steps of water molecule minimization keeping all heavy atoms of protein, DNA, and SAH fixed, 2,000 steps of minimization with only the protein backbone fixed to allow protein side chains to relax, and another 5,000 steps of minimization without any constraint on the system. The energy-minimized system was gradually heated to 300 K in 50,000 steps at a rate of 5 K per 1,000 steps at constant volume using a Berendsen thermostat [26]. The L-J potential cutoff of molecular dynamics simulation was set to 14 Å. Then, the whole system was equilibrated with unbiased MD simulations for 5 ns under NPT conditions.

2.3. Targeted MD Simulation (TMD). Targeted molecular dynamics (TMD) simulation was employed to guide a set of atoms moving from its initial to a given target structure by means of the steering forces. In this experiment, the transitions of M.HhaI from inactive to activate state were driven by applying RMSD restraints with a force constant of about 1 kcal/mol/Å^2 to each heavy atom of the catalytic loop (residues Cys81–Leu100). The offset parameter of RMSD decreased by about 0.027 Å per ps until it reached zero deviation. The total TMD simulation lasted for 2 ns.

2.4. Path CV Based Well-Tempered Metadynamics. Metadynamics [27] in its new well-tempered variant [28] was used for free energy calculation. The free energy at time (t) was defined using the following formula:

$$F(s,t) = -\frac{T+\Delta T}{\Delta T} V(s,t), \tag{1}$$

where $F(s,t)$ stands for the free energy at time t, $V(s,t)$ is the bias potential added to the system, and T is the temperature used for this simulation. ΔT is the difference between the fictitious temperature of the CV and the temperature of the simulation. The bias potential is made up by the sum of the Gaussians deposited along the trajectories of the CVs.

To trace the path, two variables $s(R)$ and $z(R)$ were introduced as [29]

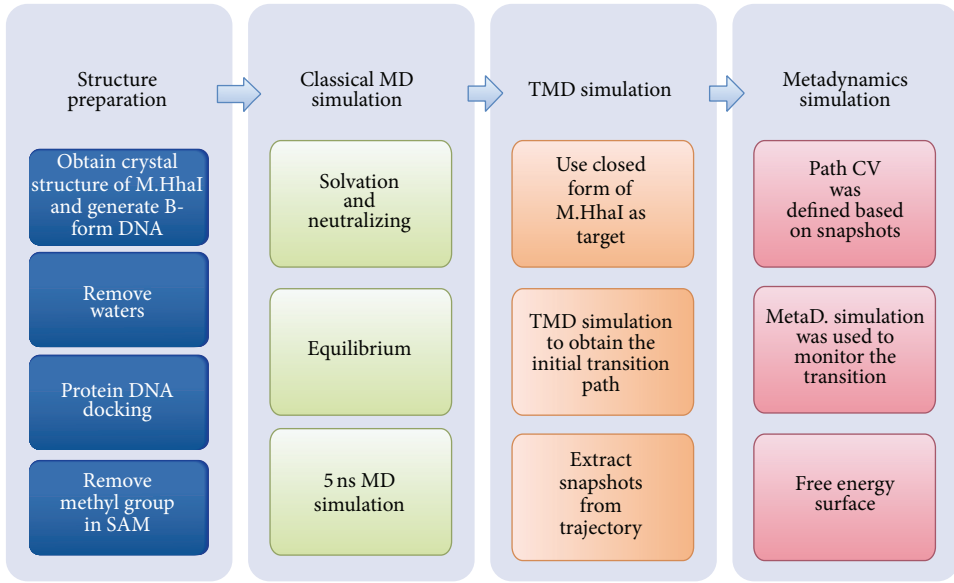
$$s(r) = \frac{\sum_{l=1}^{P} l e^{-\lambda \|S(r)-S(l)\|^2}}{\sum_{l=1}^{P} e^{-\lambda \|S(r)-S(l)\|^2}},$$
$$z(r) = -\frac{1}{\lambda} \ln\left(\sum_{l=1}^{P} e^{-\lambda \|S(r)-S(l)\|^2}\right), \tag{2}$$

where the distance between the current position $S(R)$ and the reference frames of the path $S(l)$ is calculated using a DMSD metric after alignment to the reference using a rototranslation matrix [30]. To define $s(R)$, 31 reference frames were selected from the TMD trajectory. Heavy atoms of the catalytic loop (residues Cys81–Leu100) were selected for DMSD calculation. In addition, nonhydrogen atoms of an alpha helix immediately after the catalytic loop (residues 102–120) were used for alignment. The mean interframe RMSD of these frames is 0.46 Å^2. According to the relationship between mean RMSD and λ, the λ value was set to five. We performed metadynamics only in the space of $s(R)$ whereas $z(R)$ was constrained to 3 Å^2. The hill height was set to 1.5, and the bias factor was set to 8. NAMD 2.8 [31] with the Plumed 1.3 plug-in [32] was employed for all simulations. A summary of simulation protocol is surmised in Scheme 1. Detailed parameters, preliminary simulations, and postprocessing protocols are listed in Supporting Information.

3. Results and Discussion

3.1. Biased MD Simulations. In order to avoid clash between modeled structure and surrounding solvents, a short conventional MD simulation was performed. According to the RMSD profile relevant to the starting structure along MD trajectory, the complex structure appeared to have reached a stable state after 4 ns equilibration, where the RMSD value converged to a value around 4.0 Å (shown in Supplementary Information). Given the limitation of standard molecular dynamics (MD), enhanced sampling was employed to overcome the energy barrier. Among the techniques currently available, metadynamics has shown to be useful in studies of conformational changes of proteins [33], peptide folding [34], or chemical reactions [35]. In this study, we performed metadynamics in its new variant, named well-tempered metadynamics, which allows reconstruction of the free-energy profile of the process of interest by adding an adaptive bias on a selected number of collective variables (CVs) [28]. Thus, choosing an appropriate CV is vital to successful metadynamics simulations.

Considering both protein and DNA participated in DNA recognition process, we employed two CVs to describe the transition path of catalytic loop and cognate GCGC sequence, respectively. Path CV is very useful tools which transform the high-dimensional phase space to a one-dimensional description [36]. As a result, we employ RMSD of heavy atoms to demonstrate the motion of catalytic loop. On the



Scheme 1: Molecular simulation protocol.

other hand, we choose the distance between center of mass (COM) of GCGC and COM of TRD as another CV defines the transition path. Under the acceleration of metadynamics, this loop is able to move from one free-energy minimum state to another, thereby overcoming the large free-energy barriers that are encountered during the transition process.

3.2. DNA Migrates into a Binding Cleft between TRD and Catalytic Loop as a Result of Electrostatic Attraction. During the simulation, DNA initially enters into the binding cleft between the TRD and the catalytic loop. This process can be divided into two phases. In the first phase (0 ns to 10 ns), DNA induces a conformational change of M.HhaI, leading to formation of a binding cleft between the TRD and the catalytic loop. As shown in Figure 2(III), the distance between the cognate GCGC base pair and the two target recognition loops (residues 233–240 and 252–258) [37] decreases by about 7 Å over 10 ns. When the DNA moves towards the TRD, the direction of the catalytic loop changes simultaneously as shown in Figure 2(II). Residues at the tip of catalytic loop, such as Ser85, Ser87, and Lys89, move towards the TRD and evoke rearrangement of the entire catalytic loop. The RMSD profile confirms that the catalytic loop undergoes a distinct conformational change (Figure 2(III)). When the DNA and catalytic loop move towards the TRD, a cleft between the TRD and the catalytic loop formed. DNA enters this cleft and interacts with TRD through the phosphodiester backbone. Cleft formation and DNA binding may be largely attributed to electrostatic attraction, because the TRD of this enzyme is a positively charged motif (as shown in Figure 2(I)), whereas the phosphor group at the DNA scaffold is negatively charged.

In the second phase (10–50 ns), DNA is accommodated into the binding cleft by adjusting its groove width and orientation. As shown in Figure 2(VII), both the major and minor groove widths of DNA fluctuate as the simulation proceeds. The groove width profile and snapshots derived from the trajectory demonstrate that the groove width affects the location of DNA and the catalytic loop conformation. The average minor groove width increased as the major groove narrowed at 19 ns (average groove width of the GCGC motif is 3.2 Å) and the catalytic loop approached the major groove of the GCGC sequence. The distance between DNA and the TRD increased to accommodate the TRD (Figure 2(IV)). At about 24 ns, the minor groove narrowed but the volume of the major groove increased (shown in Figure 2(V), average groove width, 9.8 Å). As the groove width changed, the catalytic loop gradually penetrated into the minor groove of DNA (RMSD value increased) and the major groove was accommodated into the TRD of M.HhaI. During this period, DNA is rotated about 45° along with the groove width fluctuation. At the end of this period (about 50 ns), DNA adopts a relatively stable orientation with the major groove facing the TRD and the minor groove facing the catalytic loop (Figure 2(VI)).

3.3. The Target Recognition Domain Recognizes Cognate DNA by Hydrogen Bonded to the GCGC Base Pair in Both the Target and Complementary Strands. Binding and recognition of the target GCGC site in DNA is a key event that occurs before base flipping [38]. Formation of hydrogen bonds may play an important role in this recognition process. Here, we monitored the hydrogen bonds number and existence map along the trajectory. As shown in Figure 3(I), the hydrogen bonds number between TRD and GCGC increased along the trajectory. The existence map, which presents the hydrogen bond formation process, indicates that Gln237 detects the target cytosine (DC2) and GC bases in the complementary strand, whereas target recognition loop II identifies DG3 and DG4 in the target strand and DG5 in the complementary strand (Figure 3(III–X)). Similar results were

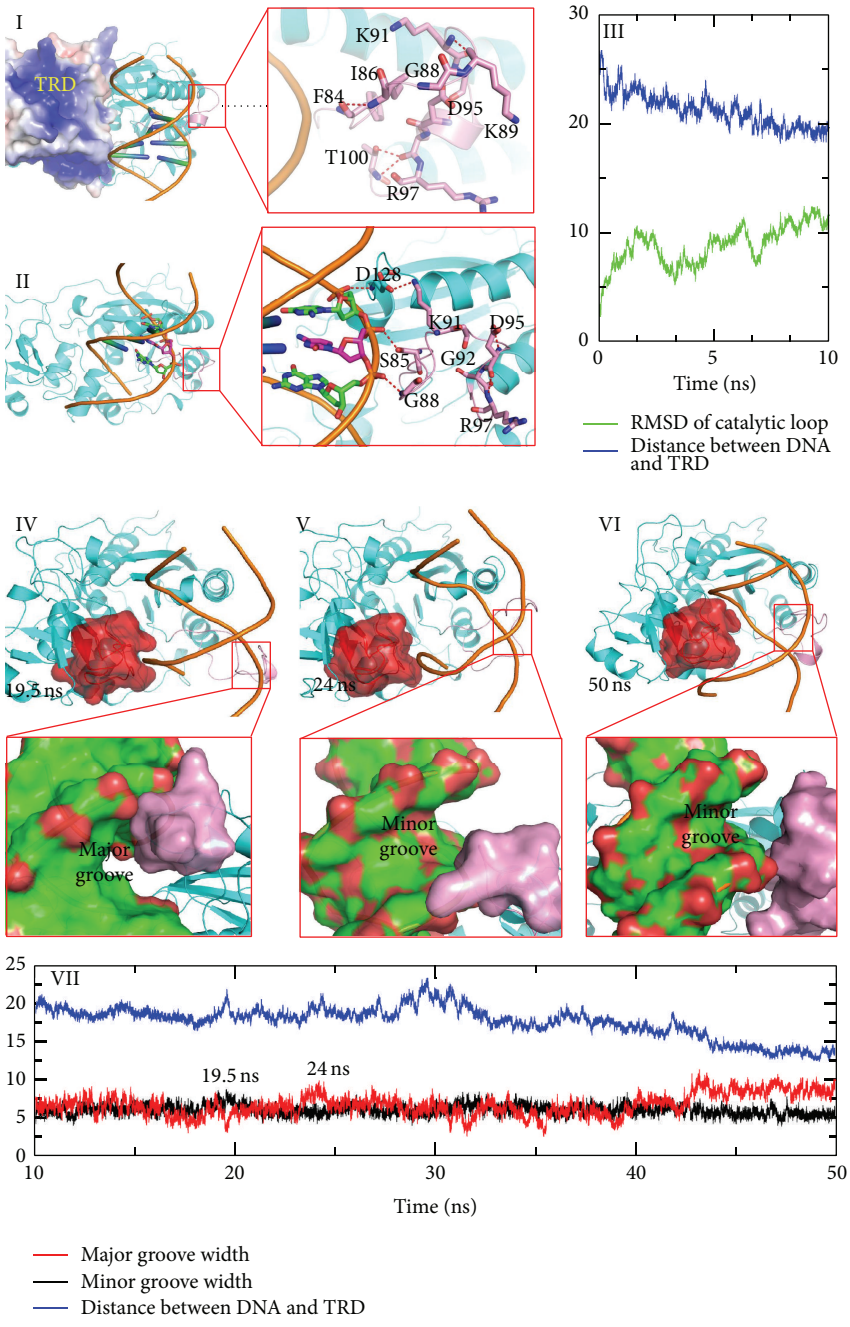


Figure 2: DNA migrates into a cleft between TRD and catalytic loop. Nonspecific DNA binding initiating the conformational transition of the catalytic loop is shown in the top. (I) Structure of the initial model. The open state of the catalytic loop is stabilized by hydrogen bonds between the main chain atoms. The electrostatic surface is generated by APBS and the positively charged area is colored blue. (II) Snapshots extracted from the MD simulation. Residues at the tip of the catalytic loop move toward the DNA backbone because hydrogen bonds between the side chain atoms and the DNA backbone replace the original hydrogen bond network. (III) RMSD and distance plot along trajectory. The distance COM of the target recognition loop (residues 230–260) and COM of GCGC are plotted in blue lines, whereas the RMSD values of the catalytic loop are plotted in light green lines. A series of snapshots that describe DNA rotation in a cleft between the TRD and the catalytic loop is shown at the bottom. (IV)–(VI) Different snapshots extracted from the trajectory. TRDs are represented with a red extended surface; major and minor grooves are also highlighted using the extended surface. (VII) Groove width and distance plot determined from the MD simulation. The major width plot is colored red, the minor groove is colored black, and the distance is colored blue.

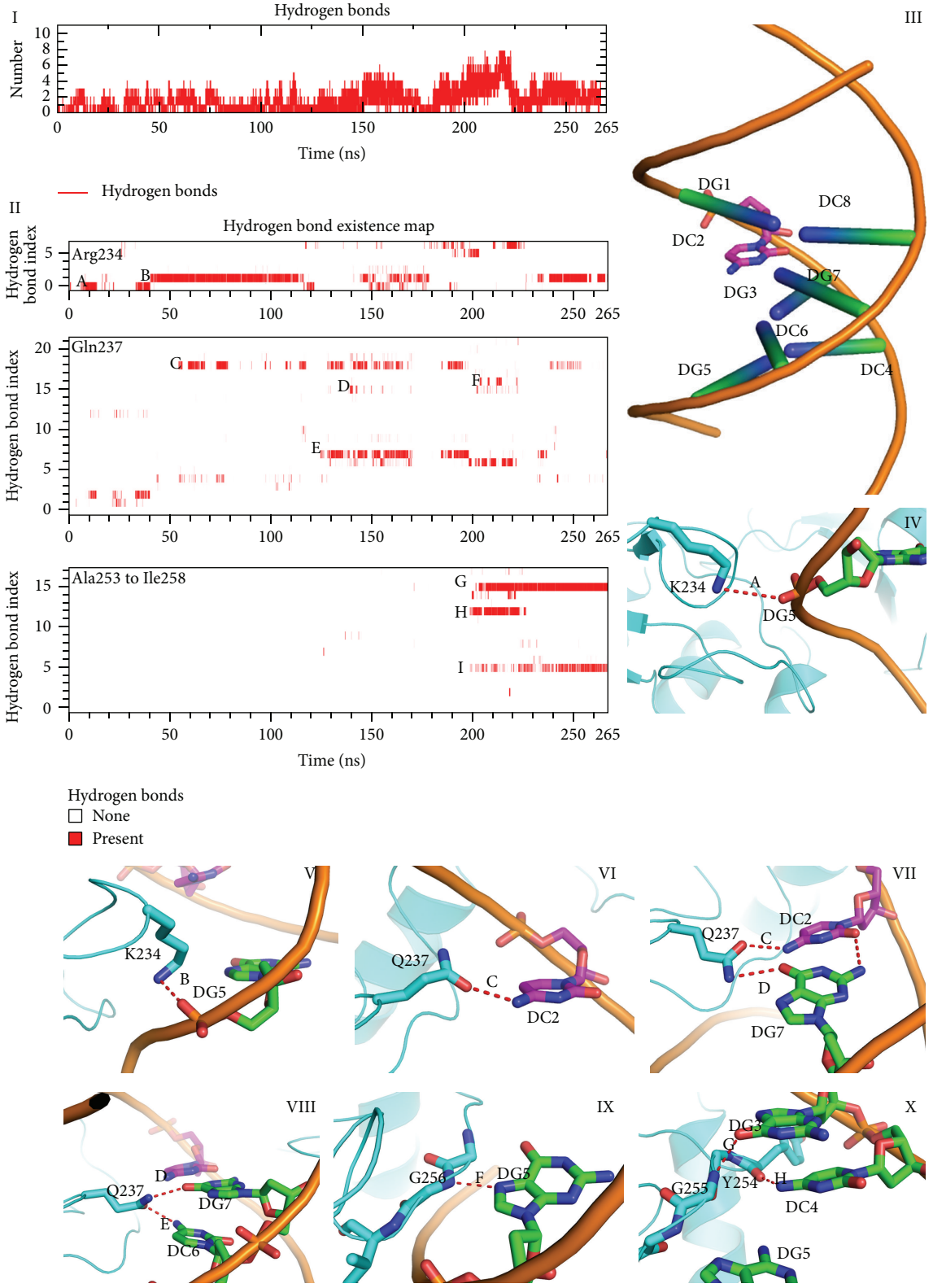


Figure 3: Hydrogen bonds between GCGC motif and target recognition loop. (I) Hydrogen bond number plot along the trajectory. (II) Hydrogen bond existence map of Arg234, Gln237, and target recognition loop (II). (III) Illustration of the notions used in Figures 4, 5, and 6. (IV)–(X) Highlighted hydrogen bonds in the existence map.

found in the NMR and fluorescence experiments [15, 39]. The sequence specificity of C5-MTases is largely attributed to two recognition loops located in the target recognition domain [39]. Furthermore, target recognition loops I and II recognize different parts of cognate GCGC sequences. Target recognition loop II recognizes DG3 and DG4, which are located at the $3'$ end of the target cytosine, by hydrogen bonding to guanine and cytosine bases [15, 39]. By contrast, target recognition loop I recognizes the $5'$ end of the target site. Gln237 plays an important role in this process; thus Q237A mutations show significantly decreased base flipping rates [40, 41].

3.4. Target Recognition Loops and the Catalytic Loop Facilitate Base Flipping by Evoking and Stabilizing the Preflipping Statue. After the GCGC sequence is recognized by M.HhaI, the distance between DNA and the TRD achieves its minimum at about 10 Å. Residues at the tip of the catalytic loop sense the translocation of the DNA backbone and rearrange its conformation to move along with the DNA scaffold. Ile86 and Ser87 are inserted into the minor groove of DNA because of reorganization of the catalytic loop. DNA backbone twisting results in increased distances between the target cytosine and the complementary guanine. The hydrogen bonds between the G:C pair are impaired, and the original hydrogen bonds loss force guanine or cytosine hydrogen bond to surrounding Ser87, Gln237, and Ser252 (Figure 4(I–IV)). After the target base is stabilized by surrounding residues, cytosine rotates about 15° out of the DNA double helix. This observation is coincident with previous molecular dynamics simulations [42], fluorescence tracking [43], and NMR experiments [44].

Preflipping leads to the loss of hydrogen bonds π-π staking; as a result, this state is not very stable and the cytosine was quickly flipped out of the double helix from the major groove (Figure 4(V–VIII)). After base flipping, residues surrounding the flipped cytosine, such as Phe79 and Gln304, stabilized the flipped status by bonding hydrogen to the cytosine base ring. This "major groove pathway" is also observed in the crystal structures, molecular dynamic simulations [19, 45], and NMR experiments [46]. On the other hand, base flipping is observed before catalytic loop is fully close, and the dynamic properties are a very important factor that affect base flipping process. This observation is coincident with results of NMR, molecular dynamic studies [16, 20, 47, 48], and mutation experiments [49] and previous research [9, 50, 51].

3.5. Conformations Transition of Catalytic Loop. The catalytic loop is a very flexible motif and has an important function in base flipping, catalytic pocket formation, and methyltransfer reactions [48]. Molecular dynamics simulations [45], crystallography studies [19, 52, 53], and mutation experiments [51, 54, 55] show that the dynamics of conformational rearrangement occurring in the catalytic loop are closely related to the base flipping process. We monitored the transformation of the secondary structure between Pro70 and Leu110 (Figure 5(I)) to observe conformational changes in the catalytic loop. M.HhaI adopts an open conformation in the solution (PDBID:2HMY) [15]: the catalytic loop stays away from the TRD and the heteroatom of the polar side chain in the catalytic loop points opposite to target recognition loops. This conformation is stabilized by the hydrogen bonds between the main chain atoms. When M.HhaI binds DNA in a nonspecific manner, residues at the tip of the catalytic loop, such as Lys89, Lys91, and Gln90, flip their side chain and approach the DNA backbone gradually (about 6–50 ns) (Figures 2(II) and 2(VI)), and the number of hydrogen bonds between the main chains decreases. Then, catalytic loop undergoes an extensive conformational change: (1) unfolding of short helix from Lys91 to Asp95 to a coiled structure (snapshot at 6 ns in Figure 5(II)), (2) gradual formation of antiparallel beta-pleated sheet between Ser85 and Gln90 (snapshot at 20 ns in (Figure 5(II))), (3) rotating of the β-sheet of Ser87 to Asp95 around the β-sheet axis by about 90° (snapshots at 50 and 66 ns in Figure 5(II)), (4) refolding of a helix between Gly92 and Ser96 (snapshot at 180 ns in Figure 5(II)), and (5) main chain atoms between Gln82 to Ser85 changing their orientation (snapshot at 240 ns in Figure 5(II)). However, Gly98 preserves its conformation, and the phi and psi angles between Gly98 and its adjacent Thr99 remain unchanged. As Matje et al. mentioned, this "hinge" may aid the refolding process because the mutation of Gly98 is believed to affect the base flipping process [51]. Besides, the orientation of guanidyl of Arg97 changes along with the unfolding of short helix from Gly92 to Ser96 and refolding process of the short helix. Helix unfolding forced the guanidyl of Arg97 to leave the DNA backbone (snapshot at 100 ns in Figure 5(II)). Nevertheless, electrostatic attraction induced Arg97 to move toward a phosphor group of the DNA (snapshot at 121 ns in Figure 5(II) and following refolding of the Phe93 to Ser96 segments.

Combining the conformational reorganization of the catalytic loop, five basins were acquired from the free energy surface and the snapshots extracted from the MD trajectory. We propose that enzymes undergo "open," "semiopen," "semiclosed," and "closed" states to accomplish the entire transition and facilitate the base flipping process (Figure 6), and the enzyme uses these different conformations to sense DNA binding and screen DNA sequence, find cognate GCGC, and flip the base, respectively.

3.6. The Mechanism of M.HhaI Screens Different DNA Sequence. When M.HhaI binds to the DNA loosely in the "semiopen" state, the DNA twists and translocates. This binding pattern provides a platform through which the enzyme can search for its target sequence [15]. Conformations (IIa) and (IIb) share similar probabilities, as shown in the free energy surface (Figures 6(IIa) and 6(IIb)); basin IIb is approximately 0.5 kcal/mol deeper than basin IIa. Thus, both the major and minor grooves have the opportunity to face the TRD or the catalytic loop of M.HhaI. The DNA rotates around the screw axis of the double helix at approximately 45° when the complex transforms from a IIa-like pattern to a IIb-like one. If a cognate sequence is detected by the target recognition loops, the GCGC will create a hydrogen bond with these loops and decrease the dynamics of the

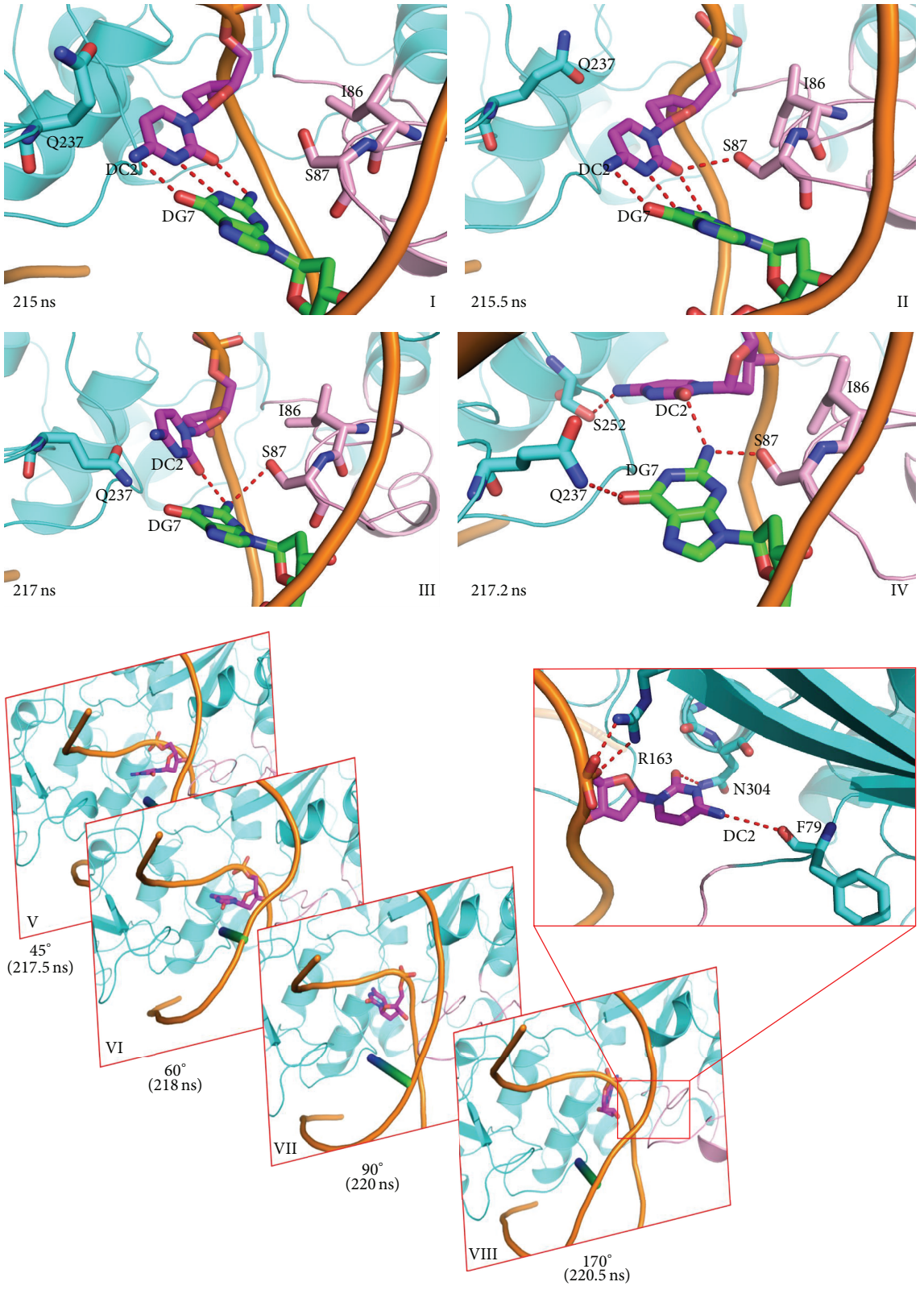


Figure 4: Base preflipping and base flipping process. (I)–(IV) Preflipping states evoked by the pushing of the DNA backbone by the catalytic loop using Ile86 and Ser87. (V)–(VII) The target cytosine flips out of the DNA double helix from the major groove side. Time and flip angles are noted at the bottom of the snapshots.

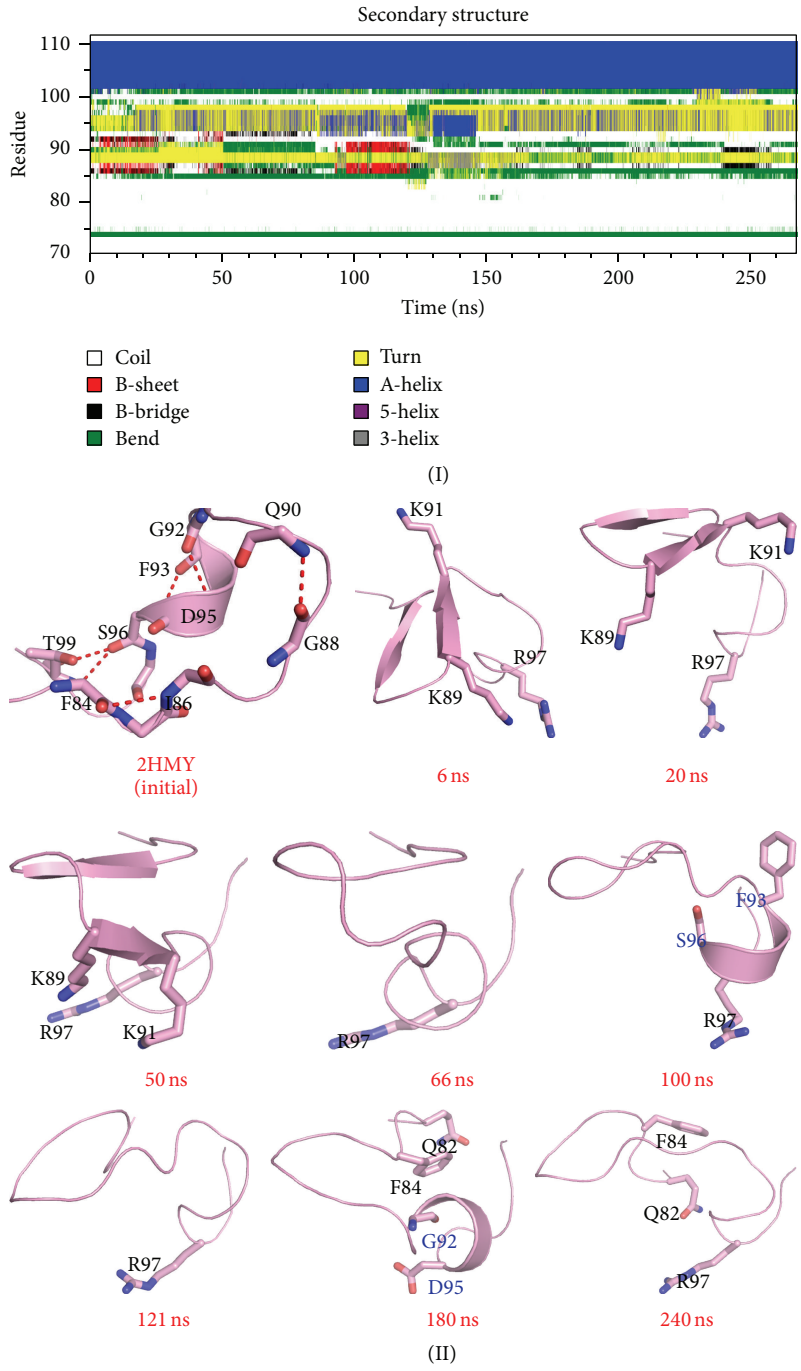


Figure 5: Conformational transition of catalytic loop during simulation. (I) Secondary structure profile of the catalytic loop. (II) Snapshots extracted from the MD simulation. In the 100 and 180 ns snapshots, the beginning and end of the short helix are colored blue. Carbon atoms are colored pink, and other atoms are colored using default settings in PyMol.

DNA backbone. Finally, DNA rotation is hindered and the recognition process begins. However, if the bases located in the major groove cannot be recognized by the TRD, the major groove leaves the TRD through "rotation-coupled sliding along the DNA helix," which is a general phenomenon found in DNA glycosylases and other similar enzymes [56]. Therefore, we speculate that catalytic loop is responsible for evoking DNA rotation and searching for the appropriate DNA sequence simultaneously when a segment of a double stranded DNA molecule or a plasmid binds to M.HhaI because the energy barrier between (IIa) and (IIb) is about 7.2 Kcal/Mol.

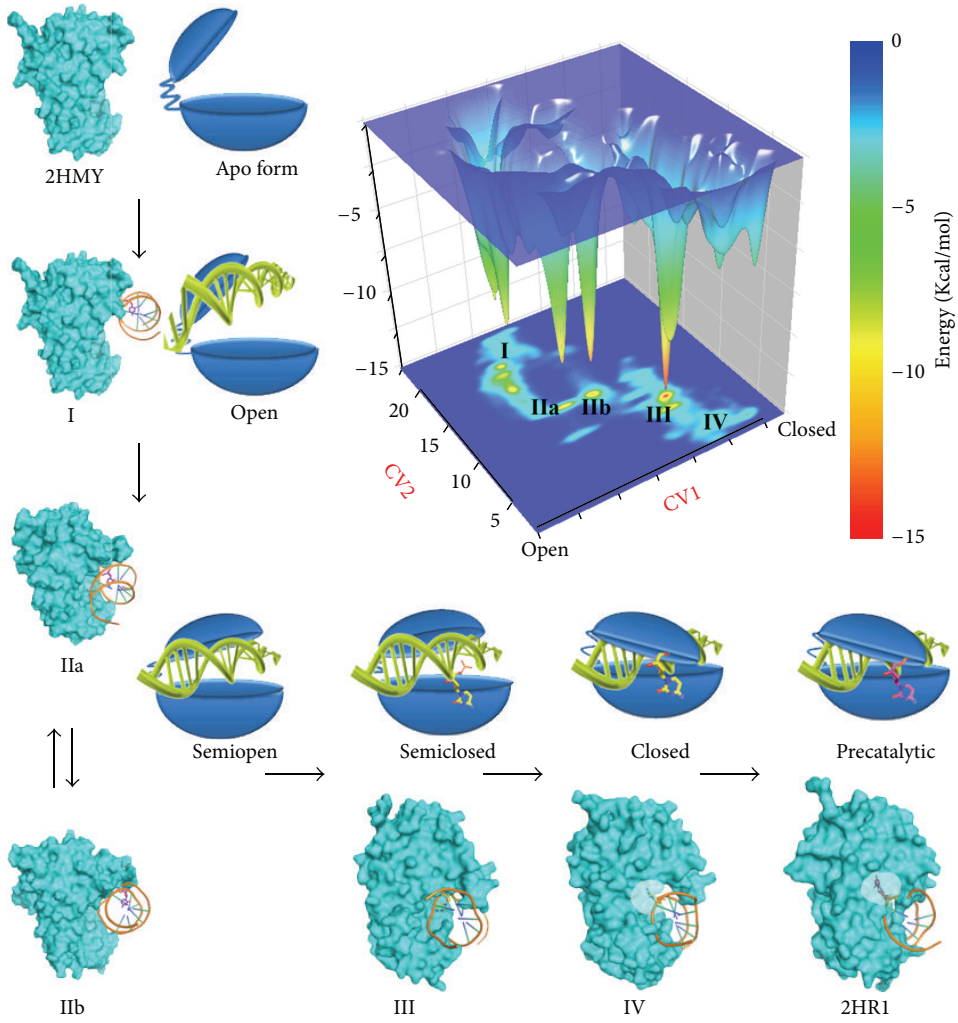


Figure 6: Inactive to active form transition. Free energy profile acquired from metadynamics simulation is shown in upper right corner. To describe the open to closed conformational transition of catalytic loop, path CV (CV1) [29] is used. In addition, we used a distance CV (CV2) that measures the distance between the COM of the GCGC base pair and the COM of the two target recognition loops. No bias is added to the distance CV, and the statistics collected during the metadynamics simulations are used to generate the final FES using the reweighting protocol [60] according to Limongelli's approach [61]. Proposed transition path from inactive form (2HMY) to active form (2HR1) is represented in an "L" shaped manner. (I)–(IV) show different basins obtained from the metadynamics simulation. Basins IIa and IIb are noted because their protein conformations are similar. Proteins are represented by an extended surface. In snapshots (IV) and (2HR1), residues around the flipped base are semitransparent to aid visualization of the flipped base. Schematic diagrams of these different states are placed beside the corresponding snapshots. Proteins are represented in a Pac-Man-like form [62] with the upper cap and lower base connected by a hinge. DNA is represented using the double helix.

Recognition of the cognate sequence is an important process prior to base flipping and methyltransfer [15]. After the target sequence is detected by M.HhaI, the catalytic loop moves closer to a TRD, and the system enters basin (III) (Figure 6(III)), which is deeper than any other minima. Experimental data indicate that the side chains of Gln237 and Arg240 are vital to evoke DNA methylation and base flipping; however, according to the metadynamics simulation, their role in the recognition phase may be different. Experiments suggest that whether or not guanine is replaced by other purine base or purine-like substrates, the base flipping rate is similar as long as hydrogen bonds between Arg240 and the base are preserved [57]. While both Q237 mutating and GCGC sequence missing abolished the catalytic activity of M.HhaI, in our simulation, the hydrogen bonds of Gln237 and GCGC are distributed within 150–200 ns (Figure 3). These hydrogen bonds include the target cytosine, the orphan guanine in the complementary strand, and cytosine 5′ between the side chains of Gln237. During this period, the distance of GCGC and the target recognition loops decreases. The driving force of GCGC motif approaching TRD is speculated to be polar interactions. Thus, besides stabilizing the flipped cytosine, Gln237 may function as a probe to detect CG binucleosides in the target and complementary strands.

3.7. Base Flipping: An Induced-Fit Process. Different hypotheses have attempted to demonstrate the base flipping process. Research shows that closure of the catalytic loop occurs after base flipping [9, 58] and that the enzyme utilizes the hydrogen bonds between Gln237 and Ser87 to lock the flipped base. Other research studies indicate that target base flipping and closure of the mobile catalytic loop occur simultaneously [17]. Our results are in agreement with the second theory, which is also known as the "induced-fit" hypothesis. This model was first presented in 2004, in which tight DNA binding is thought to be coupled with base flipping and protein loop rearrangement [11]. Subsequent fluorescence experiments [18] and molecular dynamic simulations [59] also support this theory, which demonstrates that loop rearrangements are directly coupled with base flipping. Our simulations show that the induced-fit process of DNA-protein recognition begins immediately after DNA binding to a nonspecific binding site. The negatively charged DNA backbone triggers conformational rearrangement of the catalytic loop. Then, a DNA binding cleft emerges after the catalytic loop changes its conformation. Formation of this cleft provides the enzyme with the ability to bind to DNA loosely and search for its target sequence and target base. Selected base or sequences make contact with the TRD, whereas other sequences are rejected by DNA movement. As target bases fit into the TRD, the phosphate backbone of the DNA initiates another conformational rearrangement of the catalytic loop. While the target is flipped out of the DNA double helix, the DNA approaches the TRD even as the catalytic loop is not fully closed. Subsequently, another extensive conformational transition of the catalytic loop elicited by base flipping contributes to the folding of the catalytic pocket and stabilization of the flipped cytosine.

4. Conclusion

Base flipping appears in a number of systems and enzymes, and debates regarding the detailed process and mechanism of this phenomenon persist. Here, we performed metadynamics simulation on the M.HhaI-SAH-DNA ternary complex to provide a better understanding of this interesting phenomenon. Consistent with previous experimental findings, we found that both protein and DNA play important roles in nonspecific binding, DNA sequence recognition, and the flipping process. Moreover, during the open to closed transition process, we captured a series of intermediates, the transition process into four phases according to the free energy landscape constructed based on MD simulation, and the transition process can be divided into four phases. In each phase, key residues found in the simulation coincided with data from previous experiments. Combining these findings, we proposed an "induced-fit" model to illustrate the base flipping process in M.HhaI. The results of our simulations demonstrate base flipping at the atomic level and help elucidate the mechanism underlying the base flipping process.

Conflict of Interests

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

Authors' Contribution

Lu Jin and Fei Ye contributed equally to this paper.

Acknowledgments

The authors gratefully acknowledge financial support from the Hi-Tech Research and Development Program of China (2012AA020302), the National Natural Science Foundation of China (91229204, 81230076, and 21210003), the National Science and Technology Major Project "Key New Drug Creation and Manufacturing Program" (2013ZX09507-004, 2013ZX09507001, 2014ZX09507002-005-012), and the Zhejiang Province Natural Science Foundation (LQ14H300003) and Zhejiang Provincial Top Key Discipline of Biology.

References

[1] A. Bird, "DNA methylation patterns and epigenetic memory," *Genes and Development*, vol. 16, no. 1, pp. 6–21, 2002.

[2] E. Li, "Chromatin modification and epigenetic reprogramming in mammalian development," *Nature Reviews Genetics*, vol. 3, no. 9, pp. 662–673, 2002.

[3] W. Reik and A. Lewis, "Co-evolution of X-chromosome inactivation and imprinting in mammals," *Nature Reviews Genetics*, vol. 6, no. 5, pp. 403–410, 2005.

[4] M. G. Goll and T. H. Bestor, "Eukaryotic cytosine methyltransferases," *Annual Review of Biochemistry*, vol. 74, pp. 481–514, 2005.

[5] K. D. Robertson, "DNA methylation, methyltransferases, and cancer," *Oncogene*, vol. 20, no. 24, pp. 3139–3155, 2001.

[6] E. Merkiene and S. Klimašauskas, "Probing a rate-limiting step by mutational perturbation of AdoMet binding in the Hhal methyltransferase," *Nucleic Acids Research*, vol. 33, no. 1, pp. 307–315, 2005.

[7] M. O'Gara, X. Zhang, R. J. Roberts, and X. Cheng, "Structure of a binary complex of HhaI methyltransferase with S-adenosyl-L-methionine formed in the presence of a short non-specific DNA oligonucleotide," *Journal of Molecular Biology*, vol. 287, no. 2, pp. 201–209, 1999.

[8] X. Cheng, S. Kumar, J. Posfai, J. W. Pflugrath, and R. J. Roberts, "Crystal structure of the Hhal DNA methyltransferase complexed with S-adenosyl-L-methionine," *Cell*, vol. 74, no. 2, pp. 299–307, 1993.

[9] A. Reisenauer, L. S. Kahng, S. Mccollum, and L. Shapiro, "Bacterial DNA methylation: a cell cycle regulator?" *Journal of Bacteriology*, vol. 181, no. 17, pp. 5135–5139, 1999.

[10] X. Cheng, S. Kumar, S. Klimasauskas, and R. J. Roberts, "Crystal structure of the HhaI DNA methyltransferase," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 58, pp. 331–338, 1993.

[11] Y.-F. Lee, D. S. Tawfik, and A. D. Griffiths, "Investigating the target recognition of DNA cytosine-5 methyltransferase Hhal by library selection using in vitro compartmentalisation," *Nucleic Acids Research*, vol. 30, no. 22, pp. 4937–4944, 2002.

[12] W. Choe, S. Chandrasegaran, and M. Ostermeier, "Protein fragment complementation in M.HhaI DNA methyltransferase," *Biochemical and Biophysical Research Communications*, vol. 334, no. 4, pp. 1233–1240, 2005.

[13] B. Holz, N. Dank, J. E. Eickhoff, G. Lipps, G. Krauss, and E. Weinholdt, "Identification of the binding site for the extrahelical target base in N6-adenine DNA methyltransferases by photocross-linking with duplex oligodeoxyribonucleotides containing 5-iodouracil at the target position," *Journal of Biological Chemistry*, vol. 274, no. 21, pp. 15066–15072, 1999.

[14] X. Cheng and R. J. Roberts, "AdoMet-dependent methylation, DNA methyltransferases and base flipping," *Nucleic Acids Research*, vol. 29, no. 18, pp. 3784–3795, 2001.

[15] H. Zhou, M. M. Purdy, F. W. Dahlquist, and N. O. Reich, "The recognition pathway for the DNA cytosine methyltransferase M.HhaI," *Biochemistry*, vol. 48, no. 33, pp. 7807–7816, 2009.

[16] K. Pederson, G. A. Meints, Z. Shajani, P. A. Miller, and G. P. Drobny, "Backbone dynamics in the DNA HhaI protein binding site," *Journal of the American Chemical Society*, vol. 130, no. 28, pp. 9072–9079, 2008.

[17] R. Gerasimait, E. Merkiene, and S. Klimašauskas, "Direct observation of cytosine flipping and covalent catalysis in a DNA methyltransferase," *Nucleic Acids Research*, vol. 39, no. 9, pp. 3771–3780, 2011.

[18] R. A. Estabrook and N. Reich, "Observing an induced-fit mechanism during sequence-specific DNA methylation," *Journal of Biological Chemistry*, vol. 281, no. 48, pp. 37205–37214, 2006.

[19] J. R. Horton, G. Ratner, N. K. Banavali et al., "Caught in the act: visualization of an intermediate in the DNA base-flipping pathway induced by Hhal methyltransferase," *Nucleic Acids Research*, vol. 32, no. 13, pp. 3877–3886, 2004.

[20] X. Cheng and R. J. Roberts, *Base Flipping, ELS*, John Wiley and Sons, 2010.

[21] M. van Dijk and A. M. J. J. Bonvin, "3D-DART: a DNA structure modelling server," *Nucleic Acids Research*, vol. 37, no. 2, pp. W235–W239, 2009.

[22] I. Banitt and H. J. Wolfson, "ParaDock: a flexible non-specific DNA: rigid protein docking algorithm," *Nucleic Acids Research*, vol. 39, no. 20, article e135, 2011.

[23] A. D. J. van Dijk, R. Boelens, and A. M. J. J. Bonvin, "Data-driven docking for the study of biomolecular complexes," *FEBS Journal*, vol. 272, no. 2, pp. 293–312, 2005.

[24] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfson, "PatchDock and SymmDock: servers for rigid and symmetric docking," *Nucleic Acids Research*, vol. 33, no. 2, pp. W363–W367, 2005.

[25] D. Duhovny, R. Nussinov, and H. Wolfson, "Efficient unbound docking of rigid molecules," in *Algorithms in Bioinformatics*, R. Guigó and D. Gusfield, Eds., pp. 185–200, Springer, Berlin, Germany, 2002.

[26] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of Chemical Physics*, vol. 81, no. 8, pp. 3684–3690, 1984.

[27] A. Laio and F. L. Gervasio, "Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science," *Reports on Progress in Physics*, vol. 71, no. 12, Article ID 126601, 2008.

[28] A. Barducci, G. Bussi, and M. Parrinello, "Well-tempered metadynamics: a smoothly converging and tunable free-energy method," *Physical Review Letters*, vol. 100, no. 2, Article ID 020603, 2008.

[29] D. Branduardi, F. L. Gervasio, and M. Parrinello, "From A to B in free energy space," *Journal of Chemical Physics*, vol. 126, no. 5, Article ID 054103, 2007.

[30] G. Saladino, L. Gauthier, M. Bianciotto, and F. L. Gervasio, "Assessing the performance of metadynamics and path variables in predicting the binding free energies of p38 inhibitors," *Journal of Chemical Theory and Computation*, vol. 8, no. 4, pp. 1165–1170, 2012.

[31] J. C. Phillips, R. Braun, W. Wang et al., "Scalable molecular dynamics with NAMD," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1781–1802, 2005.

[32] M. Bonomi, D. Branduardi, G. Bussi et al., "PLUMED: a portable plugin for free-energy calculations with molecular dynamics," *Computer Physics Communications*, vol. 180, no. 10, pp. 1961–1972, 2009.

[33] M. A. Brolich, L. Wang, and M. A. O'Neill, "Folding kinetics of recognition loop peptides from a photolyase and cryptochrome-DASH," *Biochemical and Biophysical Research Communications*, vol. 391, no. 1, pp. 874–878, 2010.

[34] V. Volkov and M. Bonn, "Structural properties of gp41 fusion peptide at a model membrane interface," *The Journal of Physical Chemistry B*, vol. 117, no. 49, pp. 15527–15535, 2013.

[35] X. Qian, "Free energy surface for Brønsted acid-catalyzed glucose ring-opening in aqueous solution," *The Journal of Physical Chemistry B*, vol. 117, pp. 11460–11465, 2013.

[36] L. Sutto, S. Marsili, and F. L. Gervasio, "New advances in metadynamics," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 2, no. 5, pp. 771–779, 2012.

[37] W. M. Lindstrom Jr., J. Flynn, and N. O. Reich, "Reconciling structure and function in HhaI DNA cytosine-C-5 methyltransferase," *Journal of Biological Chemistry*, vol. 275, no. 7, pp. 4912–4919, 2000.

[38] G. Vilkaitis, E. Merkiene, S. Serva, E. Weinhold, and S. Klimašauskas, "The mechanism of DNA cytosine-5 methylation. Kinetic and mutational dissection of HhaI methyltransferase," *Journal of Biological Chemistry*, vol. 276, no. 24, pp. 20924–20934, 2001.

[39] R. Gerasimaite, G. Vilkaitis, and S. Klimašauskas, "A directed evolution design of a GCG-specific DNA hemimethylase," *Nucleic Acids Research*, vol. 37, no. 21, Article ID gkp772, pp. 7332–7341, 2009.

[40] X. D. Cheng, "Structure and function of DNA methyltransferases," *Annual Review of Biophysics and Biomolecular Structure*, vol. 24, pp. 293–318, 1995.

[41] D. Daujotyte, S. Serva, G. Vilkaitis, E. Merkiene, Č. Venclovas, and S. Klimašauskas, "HhaI DNA methyltransferase uses the protruding Gln237 for active flipping of its target cytosine," *Structure*, vol. 12, no. 6, pp. 1047–1055, 2004.

[42] N. Huang and A. D. MacKerell Jr., "Atomistic view of base flipping in DNA," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 362, no. 1820, pp. 1439–1460, 2004.

[43] D. M. Matje, H. Zhou, D. A. Smith et al., "Enzyme-promoted base flipping controls DNA methylation fidelity," *Biochemistry*, vol. 52, no. 10, pp. 1677–1685, 2013.

[44] S. Klimašauskas, T. Szyperski, S. Serva, and K. Wüthrich, "Dynamic modes of the flipped-out cytosine during Hhal methyltransferase-DNA interactions in solution," *The EMBO Journal*, vol. 17, no. 1, pp. 317–324, 1998.

[45] N. Huang, N. K. Banavali, and A. D. MacKerell Jr., "Protein-facilitated base flipping in DNA by cytosine-5-methyltransferase," *Proceedings of the National Academy of*

Sciences of the United States of America, vol. 100, no. 1, pp. 68–73, 2003.

[46] G. A. Meints and G. P. Drobny, "Dynamic impact of methylation at the M. HhaI target site: a solid-state deuterium NMR study," *Biochemistry*, vol. 40, no. 41, pp. 12436–12443, 2001.

[47] M. Fuxreiter, N. Luo, P. Jedlovszky, I. Simon, and R. Osman, "Role of base flipping in specific recognition of damaged DNA by repair enzymes," *Journal of Molecular Biology*, vol. 323, no. 5, pp. 823–834, 2002.

[48] X. Zhang and T. C. Bruice, "The mechanism of M.Hhal DNA C5 cytosine methyltransferase enzyme: a quantum mechanics/molecular mechanics approach," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 16, pp. 6148–6153, 2006.

[49] F.-K. Shieh, B. Youngblood, and N. O. Reich, "The role of Arg165 towards base flipping, base stabilization and catalysis in M.HhaI," *Journal of Molecular Biology*, vol. 362, no. 3, pp. 516–527, 2006.

[50] A. R. Fersht, "The hydrogen bond in molecular recognition," *Trends in Biochemical Sciences*, vol. 12, pp. 301–304, 1987.

[51] D. M. Matje, C. T. Krivacic, F. W. Dahlquist, and N. O. Reich, "Distal structural elements coordinate a conserved base flipping network," *Biochemistry*, vol. 52, no. 10, pp. 1669–1676, 2013.

[52] G. Vilkaitis, A. Dong, E. Weinhold, X. Cheng, and S. Klimašauskas, "Functional roles of the conserved threonine 250 in the target recognition domain of HhaI DNA methyltransferase," *Journal of Biological Chemistry*, vol. 275, no. 49, pp. 38722–38730, 2000.

[53] M. O'Gara, J. R. Horton, R. J. Roberts, and X. Cheng, "Structures of hhai methyltransferase complexed with substrates containing mismatches at the target base," *Nature Structural Biology*, vol. 5, no. 10, pp. 872–877, 1998.

[54] G. Lukinavičius, A. Lapinaitė, G. Urbanavičiūtė et al., "Engineering the DNA cytosine-5 methyltransferase reaction for sequence-specific labeling of DNA," *Nucleic Acids Research*, vol. 40, pp. 11594–11602, 2012.

[55] B. Youngblood, F.-K. Shieh, S. De Los Rios, J. J. Perona, and N. O. Reich, "Engineered extrahelical base destabilization enhances sequence discrimination of DNA methyltransferase M.HhaI," *Journal of Molecular Biology*, vol. 362, no. 2, pp. 334–346, 2006.

[56] P. C. Blainey, G. Luo, S. C. Kou et al., "Nonspecifically bound proteins spin while diffusing along DNA," *Nature Structural and Molecular Biology*, vol. 16, no. 12, pp. 1224–1229, 2009.

[57] R. A. Estabrook, T. T. Nguyen, N. Fera, and N. O. Reich, "Coupling sequence-specific recognition to DNA modification," *Journal of Biological Chemistry*, vol. 284, no. 34, pp. 22690–22696, 2009.

[58] D. M. Matje and N. O. Reich, "Molecular drivers of base flipping during sequence-specific DNA methylation," *ChemBioChem*, vol. 13, no. 11, pp. 1574–1577, 2012.

[59] K.-I. Okazaki and S. Takada, "Dynamic energy landscape view of coupled binding and protein conformational change: induced-fit versus population-shift mechanisms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 32, pp. 11182–11187, 2008.

[60] M. Bonomi, A. Barducci, and M. Parrinello, "Reconstructing the equilibrium boltzmann distribution from well-tempered metadynamics," *Journal of Computational Chemistry*, vol. 30, no. 11, pp. 1615–1621, 2009.

[61] V. Limongelli, L. Marinelli, S. Cosconati et al., "Sampling protein motion and solvent effect during ligand binding," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 5, pp. 1467–1472, 2012.

[62] J. Wereszczynski and I. Andricioaei, "Free energy calculations reveal rotating-ratchet mechanism for DNA supercoil relaxation by topoisomerase IB and its inhibition," *Biophysical Journal*, vol. 99, no. 3, pp. 869–878, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 827371, 10 pages
http://dx.doi.org/10.1155/2014/827371

Research Article

Privacy Preserving RBF Kernel Support Vector Machine

Haoran Li,[1] Li Xiong,[1] Lucila Ohno-Machado,[2] and Xiaoqian Jiang[2]

[1] *Department of Mathematics & Computer Science, Emory University, Atlanta, GA 30322, USA*
[2] *Division of Biomedical Informatics, UC San Diego, La Jolla, CA 92093, USA*

Correspondence should be addressed to Haoran Li; hli57@emory.edu

Received 16 February 2014; Accepted 8 April 2014; Published 12 June 2014

Academic Editor: Bairong Shen

Data sharing is challenging but important for healthcare research. Methods for privacy-preserving data dissemination based on the rigorous differential privacy standard have been developed but they did not consider the characteristics of biomedical data and make full use of the available information. This often results in too much noise in the final outputs. We hypothesized that this situation can be alleviated by leveraging a small portion of open-consented data to improve utility without sacrificing privacy. We developed a hybrid privacy-preserving differentially private support vector machine (SVM) model that uses public data and private data together. Our model leverages the RBF kernel and can handle nonlinearly separable cases. Experiments showed that this approach outperforms two baselines: (1) SVMs that only use public data, and (2) differentially private SVMs that are built from private data. Our method demonstrated very close performance metrics compared to nonprivate SVMs trained on the private data.

1. Introduction

Data sharing is important for accelerating scientific discoveries, especially when there are not enough local samples to test a hypothesis [1, 2]. However, medical data are sensitive as they essentially contain personal information and can reveal much about ethnicity, disease risk [3], and even family surnames [4]. To promote data sharing, it is important to develop privacy-preserving algorithms that respect data confidentiality and present data utility [5], especially when one wants to leverage cloud computing [6].

Privacy preserving data analysis and publishing [7, 8] have received considerable attention in recent years as a promising approach for sharing information while preserving data privacy. Differential privacy [9–11] has recently emerged as one of the strongest privacy guarantees for statistical data release [12–17]. A statistical aggregation or computation is DP (we shorten differentially private to DP) if the outcome is formally indistinguishable when run with and without any particular record in the dataset. The level of indistinguishability is quantified as a privacy parameter ϵ. A common mechanism to achieve differential privacy is the Laplace mechanism [18] which injects calibrated noise to a statistical measure determined by the privacy parameter ϵ and the sensitivity of the statistical measure influenced by the inclusion and exclusion of a record in the dataset. A lower privacy parameter requires larger noise to be added and provides a higher level of privacy.

General purpose algorithms for privacy protection (e.g., [19, 20]) often introduce too much perturbation error, which renders the resulting information useless for healthcare research. Our contribution is to leverage a small portion of open-consented data to maximally explore information that resides in the private data through a hybrid framework. Figure 1 shows an example of an environment in this case. We recently published differentially private distributed logistic regression using public and private biomedical datasets [21], which demonstrated advantages over pure private or public models. However, logistic regression is a generalized linear model, which has limited flexibility in classifying complex patterns. In this paper, we sought to extend our previous effort to the more powerful, RBF-kernel based support vector machines.

The remainder of the paper is organized as follows. Section 2 reviews background knowledge of differential privacy and SVM and RBF kernel. Section 3 describes the framework and details for our hybrid SVM mechanism. Then, Section 4 contains an extensive set of experimental

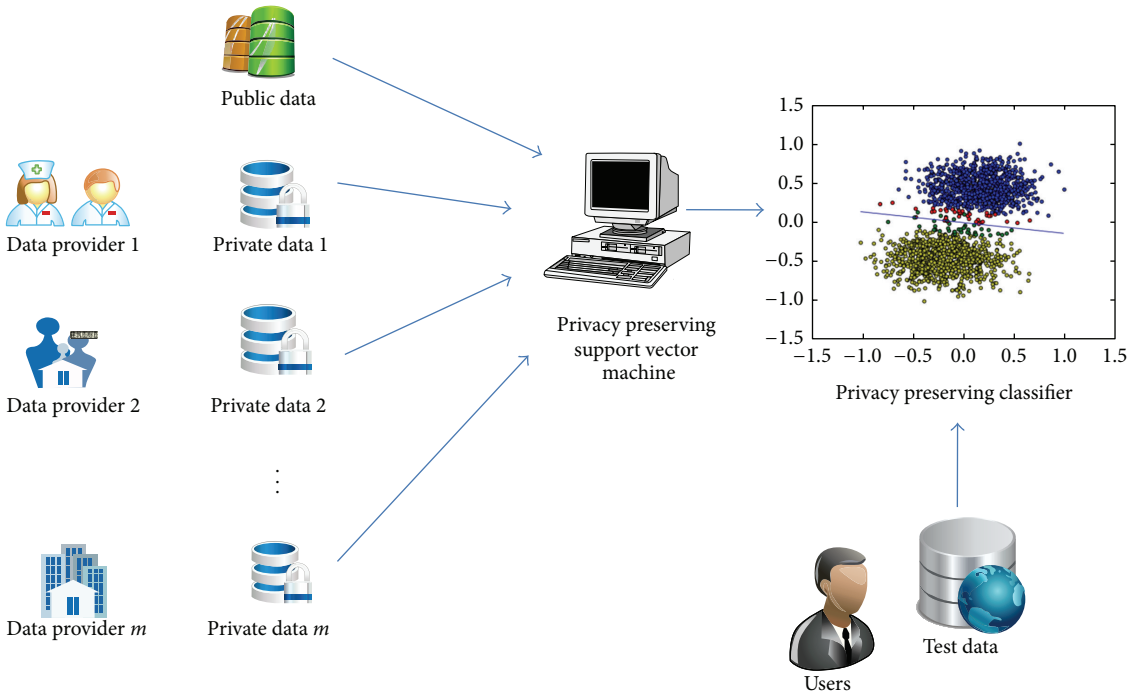


Figure 1: Biomedicine data sharing system. A small amount of public data and a large amount of private data are available for different data providers. A privacy preserving support vector machine can leverage both public and private data to maximize the classification accuracy under differential privacy. Then users can classify their test data via the released privacy preserving classifier.

evaluations. Finally, Section 5 concludes the paper with conclusions, limitations, and directions for future work.

2. Related Work

Rubinstein et al. [22] propose a private kernel SVM algorithm (shortened as PrivateSVM) which only works for a translation-invariant kernel $g(\Delta)$. The method approximates the original infinite feature space Ω of $g(\Delta)$ with a finite feature space $\widetilde{\Omega}$ using the Fourier transform $p(\omega)$ of $g(\Delta)$. Then add the noise to the weight parameters in the primal form based on the new space $\widetilde{\Omega}$. One weakness is that the parameters used to construct $\widetilde{\Omega}$ are randomly generated from $p(\omega)$ which degrades the approximation accuracy of $\widetilde{\Omega}$ to Ω. Another problem is that the utility bounds use the same regularization parameter value to compare the private and nonprivate classifiers. They take no consideration into the change of regularization parameter incurred by privacy constraints. Chaudhuri et al. [23] investigated a general mechanism, namely, DPERM, to produce private approximations of classifiers by regularized empirical risk minimization (ERM) with good perturbation error. Akin to PrivateSVM, DPERM requires that the underlying kernel is translation invariant. In this paper, we will compare our method to the PrivateSVM algorithm, since DPERM has comparable performance with PrivateSVM.

3. Preliminary

Consider an original dataset $D = \{(\mathbf{x}_i, y_i) \mid i \in Z^+, 1 \leq i \leq n\}$ that contains a small portion of public data D_{public} and a large part of private data D_{private}. Our goal is to release a differentially private support vector machine using both public and private data. In this section, we first introduce the definition of differential privacy; then, we give a brief overview of SVM and RBF kernel.

3.1. Differential Privacy. Differential privacy has emerged as one of the strongest privacy definitions for statistical data release. It guarantees that if an adversary knows complete information of all the tuples in D except one, the output of a differentially private randomized algorithm should not give the adversary too much additional information about the remaining tuples. We say that datasets D and D' differ in only one tuple if we can obtain D' by removing or adding only one tuple from D. A formal definition of differential privacy is given as follows.

Definition 1 (ϵ-differential privacy [18]). Let $\mathscr{A}$ be a randomized algorithm over two datasets D and D' differing in only one tuple, and let $\mathcal{O}$ be any arbitrary set of possible outputs f $\mathscr{A}$. Algorithm $\mathscr{A}$ satisfies ϵ-differential privacy if and only if the following holds:

$$\Pr\left[\mathscr{A}(D) \in \mathcal{O}\right] \leq e^{\epsilon} \Pr\left[\mathscr{A}\left(D'\right) \in \mathcal{O}\right]. \quad (1)$$

Intuitively, differential privacy ensures that the released output distribution of $\mathscr{A}$ remains nearly the same whether or not an individual tuple is in the dataset.

A common mechanism to achieve differential privacy is the Laplace mechanism [18] that adds a small amount of independent noise to the output of a numeric function f to

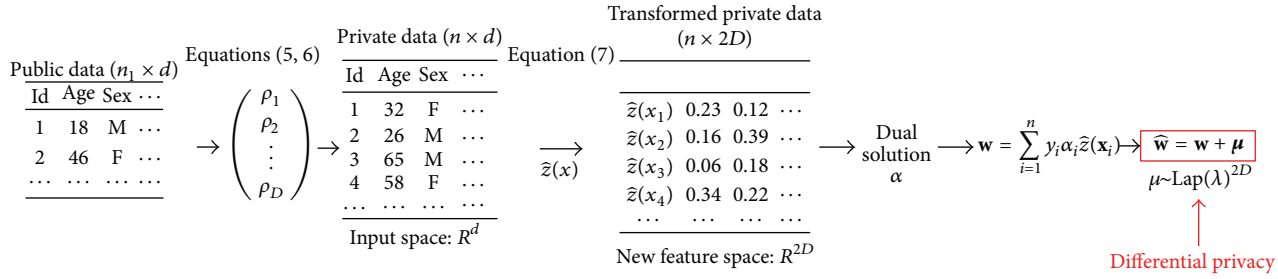


Figure 2: Detailed framework of our hybrid SVM.

fulfill ϵ-differential privacy of releasing f, where the noise is drawn from *Laplace distribution* with a probability density function $\Pr[\eta = x] = (1/2b)e^{-|x|/b}$. A Laplace noise has a variance $2b^2$ with a magnitude of b. The magnitude b of the noise depends on the concept of *sensitivity* which is defined as follows.

Definition 2 (sensitivity [18]). Let f denote a numeric function, and the sensitivity of f is defined as the maximal L_1-norm distance between the outputs of f over the two datasets D and D' which differ in only one tuple. Formally,

$$\Delta_f = \max_{D,D'} \left\| f(D) - f\left(D'\right) \right\|_1. \tag{2}$$

With the concept of sensitivity, the noise follows a zero-mean Laplace distribution with the magnitude $b = \Delta_f/\epsilon$. To fulfill ϵ-differential privacy for a numeric function f over D, it is sufficient to publish $f(D) + X$, where X is drawn from $\text{Lap}(\Delta_f/\epsilon)$.

3.2. Review of SVM and RBF Kernel. SVM is one of the most popular supervised binary classification methods that takes a sample and a predetermined kernel function as input, and outputs a predicted class label for this sample. Consider training data $D = \{(\mathbf{x}_i, y_i) \mid i \in Z^+, 1 \le i \le n\}$, where $\mathbf{x}_i \in R^d$ denotes the training input points, $y_i \in \{1, -1\}$ are the training class labels, and n is the size of training data. Here, d is the dimension of input data and "+1" and "−1" are class labels. A SVM maximizes the geometric margin between two classes of data and minimizes the error from misclassified data points. The primal form of a soft-margin SVM can be written as

$$\min_{\mathbf{w} \in R^F} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} l\left(y_i, f_{\mathbf{w}}\left(\mathbf{x}_i\right)\right), \tag{3}$$

where $\mathbf{w}$ is the normal vector to the hyperplane separating two classes of data, $l(y, \widehat{y})$ is a loss function convex in $\widehat{y}$, C is a regularization parameter that weighs smoothness and errors (i.e., large for fewer errors, smaller for increased smoothness), and $f_{\mathbf{w}}(\mathbf{x}_i) = \langle \phi(x_i), \mathbf{w} \rangle$, where $\phi(\mathbf{x}) : R^d \to R^F$ is a function mapping training data point from their input space R^d to a new F-dimensional feature space R^F (F may be infinite). Sometimes we map the training data from their input space to another high-dimensional feature space in order to classify nonlinearly separable data. When F is large or infinite, the innerproducts in feature space R^F may be computed efficiently by an explicit representation of the kernel function $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$. For example, $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$ is a linear kernel function for a linear SVM, and $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|_2^2/\sigma^2)$ is a RBF kernel function, which is translation invariant.

In this paper, we use a RBF kernel function. Our method can be applied to any translation invariant kernel SVM. With the hinge loss $l(y_i, f_{\mathbf{w}}(\mathbf{x}_i)) = \max(0, 1 - y_i f_{\mathbf{w}}(\mathbf{x}_i))$, we can obtain a dual form SVM written as

$$\begin{aligned} &\max_{\boldsymbol{\alpha} \in R^n} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k\left(\mathbf{x}_i, \mathbf{x}_j\right) \\ &\text{s.t.} \quad 0 \le \alpha_i \le C, \quad \forall i \in 1, \ldots, n, \end{aligned} \tag{4}$$

where $\alpha_i \in \boldsymbol{\alpha}$, $i \in (1, n)$ is a persample parameter and $w_j \in \mathbf{w}$, $j \in (1, d)$ is a perfeature weight parameter. The weight vector $\mathbf{w}$ can be converted from sample weight vector $\boldsymbol{\alpha}$ via $\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i$ in the linear SVM.

4. Privacy Preserving Hybrid SVM

In this section, we first introduce a framework overview and then the technical details of our hybrid SVM method. We assume that all data samples follow the same distribution. Here, we assume that all original data from different data sets follow some unknown joint multivariate distribution and all data tuples are samples from this distribution.

4.1. The General Framework. Figure 2 illustrates the general framework of hybrid SVM. Algorithm 1 presents the hybrid SVM algorithm. First, we use the small amount of public data and (5) and (6) to compute the parameter $\boldsymbol{\rho} = (\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D)^T$, $\boldsymbol{\rho}_i \in R^d$ in the mapping function of the approximation form to the RBF kernel. Second, with $\boldsymbol{\rho}$, we transform the private data from the original sample space to the new $2D$-dimensional feature space via the mapping function $\widehat{z}(x)$ in (7). Then we can compute the parameter $\boldsymbol{\alpha}$ in the dual space with the transformed private data and $\mathbf{w}$ in the primal space via the linear relationship between $\boldsymbol{\alpha}$ and $\mathbf{w}$ in the linear SVM. Finally, draw $\boldsymbol{\mu}$ from $\text{Lap}(\lambda)^{2D}$ where $\lambda = 2^{2.5} C \sqrt{D}/n\epsilon$ and return $\widehat{\mathbf{w}} = \mathbf{w} + \boldsymbol{\mu}$ and $\boldsymbol{\rho}$. Then users can transform their test data to the new $2D$-dimensional feature space with $\boldsymbol{\rho}$ and classify the transformed data with $\widehat{\mathbf{w}}$. Here the computation

Input: Public data D_{public}, private data D_{private}, the dimensionality D of $\boldsymbol{\rho}$, a regularization parameter C, and privacy budget ϵ;
Output: Differentially private SVM;
(1) Use the public data to compute $\boldsymbol{\rho} = (\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D)^T$ via (5), (6);
(2) Transform each record of the private data to new $2D$-dimensional data via the mapping function $\widehat{z}(x)$ defined by (7);
(3) Compute the parameter $\boldsymbol{\alpha}$ in the dual space with the transformed private data, and $\mathbf{w}$ in the primal space via $\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \widehat{z}(\mathbf{x}_i)$;
(4) Draw $\boldsymbol{\mu}$ from $\text{Lap}(\lambda)^{2D}$, $\lambda = 2^{2.5} C\sqrt{D}/n\epsilon$, then return $\widehat{\mathbf{w}} = \mathbf{w} + \boldsymbol{\mu}$ and $\boldsymbol{\rho}$.

Algorithm 1: Hybrid SVM algorithm.

of parameter $\boldsymbol{\rho}$ has no privacy risk because it is retrieved directly from public data. More details about hybrid SVM will be given in the successive subsections.

Privacy Properties. We present the following theorem showing the privacy property of Algorithm 1.

Theorem 3. *Algorithm 1 guarantees ϵ-differential privacy.*

Proof. For step 1, no private data is used, and hence step 1 does not impact the privacy guarantee. Due to Corollary 15 in [22] and the fact that the hinge-loss is convex and 1-Lipschitz in $\widehat{y}$, the sensitivity of $\mathbf{w}$ over a pair of neighbouring datasets is $\Delta_{\mathbf{w}} = 2^{2.5} C\sqrt{D}/n$. Then the scale parameter λ in step 4 is set to $\lambda = \Delta_{\mathbf{w}}/\epsilon = 2^{2.5} C\sqrt{D}/n\epsilon$ due to the Laplace mechanism introduced in Section 3.1. Therefore, Algorithm 1 preserves ϵ-differential privacy which completes the proof. □

4.2. The Computation of $\boldsymbol{\rho}$. Rahimi and Recht [24] approximate a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ induced by an infinite dimensional feature mapping with a random RKHS $\widehat{\mathcal{H}}$ induced by a random finite-dimensional mapping z. The random finite-dimensional RKHS $\widehat{\mathcal{H}}$ can be constructed by drawing D i.i.d. vectors $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D$ from the Fourier transform of a positive-definite translation-invariant kernel function $k(x, y)$, such as the RBF kernel function. Then we can obtain an approximation form $z(x)^T z(y)$ of $k(x, y)$ using the real-valued mapping function $z(x) : R^d \rightarrow R^D$ defined by the following equation:

$$z(x) = \sqrt{\frac{2}{D}} \left[\cos\left(\boldsymbol{\rho}_1^T \mathbf{x} + b_1\right) \cdots \cos\left(\boldsymbol{\rho}_D^T \mathbf{x} + b_D\right)\right]^T, \quad (5)$$

where $b_1, \ldots, b_D$ are i.i.d. samples drawn from a uniform distribution $U[0, 2\pi]$. $z(x) : R^d \rightarrow R^D$ maps the data from its original d-dimensional input space to the new D-dimensional feature space. Their approach is based on the fact that the kernel function of a continuous positive-definite translation-invariant kernel is the Fourier transform of a nonnegative measure. The uniform convergence property of the approximation form $z(x)^T z(y)$ to the kernel function $k(x, y)$ has also been proved in [24]. In our context, the kernel function $k(x, y)$ refers to the RBF kernel function.

In our problem setting, since a small amount of public data can be considered as x in $z(x)$ and only the vectors $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D$ are needed to construct the random finite-dimensional RKHS $\widehat{\mathcal{H}}$, we can compute the vectors $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D$ with an optimization function defined as follows:

$$\min_{\boldsymbol{\rho} \in R^{D \times d}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{2}{D} z(x_i)^T z\left(x_j\right) - k\left(x_i, x_j\right) \right|. \quad (6)$$

Since (6) is an unconstrained nonlinear optimization function, we solve it using L-BFGS (the full name is Limited-memory Broyden Fletcher Goldfarb Shanno) algorithm.

Thus, we can obtain a more accurate approximation form $z(x)^T z(y)$ of the kernel function $k(x, y)$ by deploying the public data to compute the $\boldsymbol{\rho}$, than randomly sampling $\boldsymbol{\rho}$ from the fourier transform of the kernel function $k(x, y)$ as shown in [25]. To guarantee differential privacy, we need only consider the data-dependent weight parameter $\mathbf{w}$. Fortunately we can employ the differentially private linear SVM approach in [25] to compute $\mathbf{w}$ after transforming all private data to a new $2D$-dimensional feature space using the mapping $\widehat{z}(x) : R^d \rightarrow R^{2D}$ defined in (7) with the vectors $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D$ as follows:

$$\widehat{z}(x) = \frac{1}{\sqrt{D}} \left[\cos\left(\boldsymbol{\rho}_1^T x\right), \sin\left(\boldsymbol{\rho}_1^T x\right), \ldots, \cos\left(\boldsymbol{\rho}_D^T x\right), \sin\left(\boldsymbol{\rho}_D^T x\right)\right]^T. \quad (7)$$

4.3. The Computation of $\widehat{\mathbf{w}}$. With the vectors $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_D$ to approximate the RBF kernel function, we can convert RBF kernel SVM in the d-dimensional input space into the linear SVM in a new $2D$-dimensional feature space with (7), then use the privacy preserving linear SVM algorithm in [25]. The general idea of this algorithm is that with the transformed $2D$-dimensional private data, we first compute the parameter $\boldsymbol{\alpha}$ in the dual space and then $\mathbf{w}$ in the primal space using $\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \widehat{z}(\mathbf{x}_i)$; then we draw $\boldsymbol{\mu}$ from $\text{Lap}(\lambda)^{2D}$, where $\lambda = 2^{2.5} C\sqrt{D}/n\epsilon$ and compute noisy $\widehat{\mathbf{w}}$ with $\widehat{\mathbf{w}} = \mathbf{w} + \boldsymbol{\mu}$.

5. Experiments

In this section, we experimentally evaluate our hybrid SVM and compare it with one state-of-the-art method, called

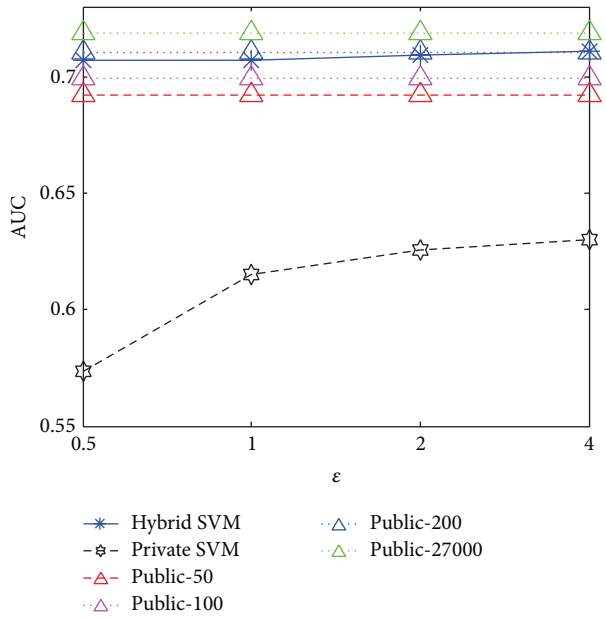


Figure 3: AUC versus privacy budget for US.

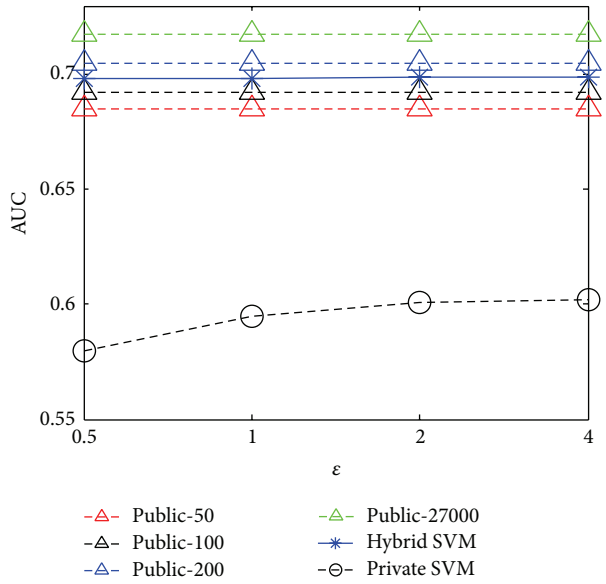


Figure 4: AUC versus privacy budget for Brazil.

private SVM and on baseline method. We evaluate the utility of the trained SVM classifier using the AUC metric. Hybrid SVM and private SVM are implemented in MATLAB R2010b, and all experiments were performed on a PC with 3.2 GHz CPU and 8 G RAM.

5.1. Experiment Setup

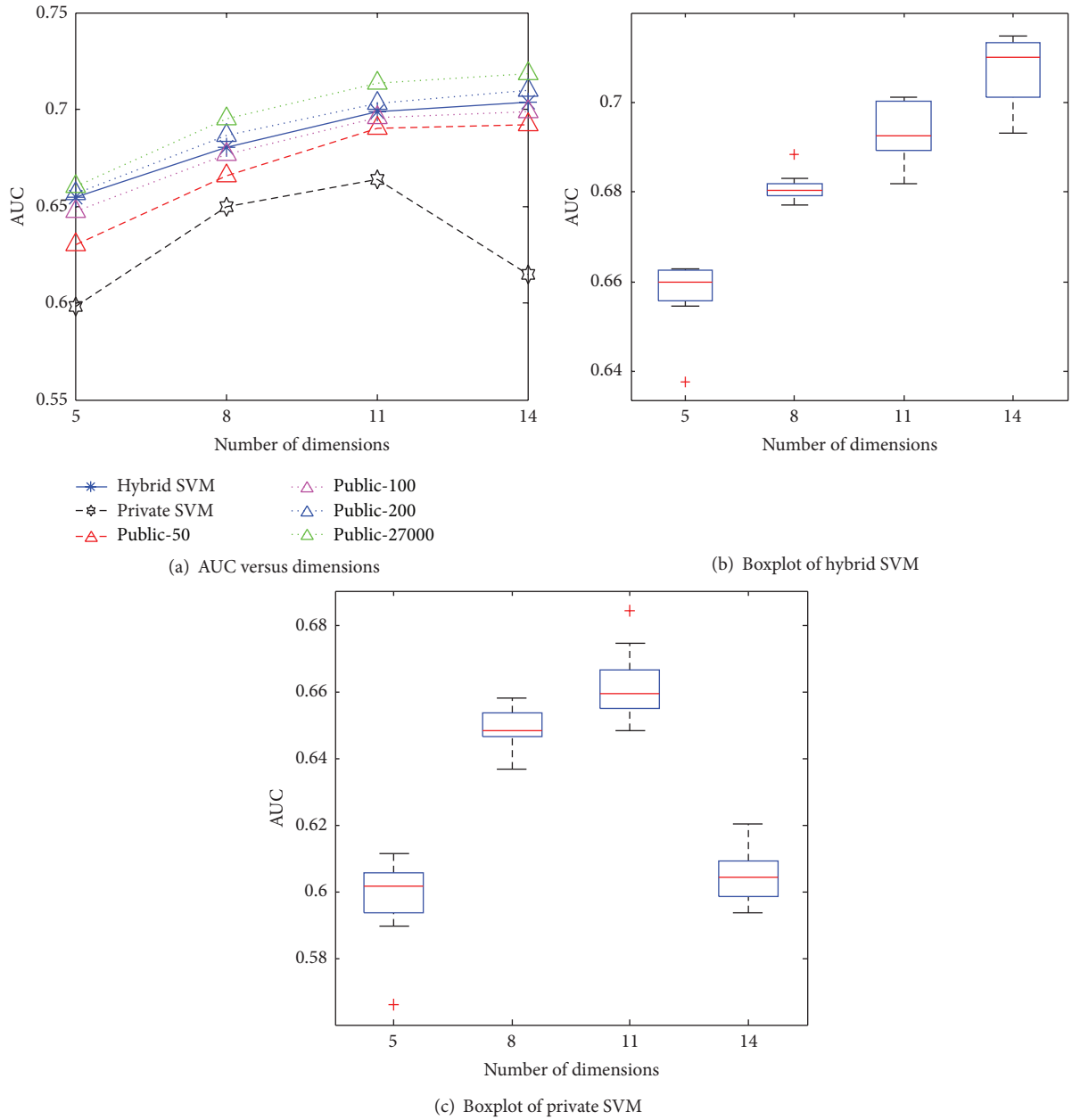
Datasets. We used two open source datasets from the Integrated Public Use Microdata Series (Minnesota Population Center, Integrated public use microdata series—international: Version 5.0., 2009, https://international.ipums.org), the US and Brazil census datasets with 370,000 and 190,000 records collected in the US and Brazil, respectively. One motivation for using these public datasets is that it bears similar attributes (e.g., demographic features) as some medical records, but it is publicly available for testing and comparisons. From each dataset, we selected 40,000 records, with 10,000 records serving as the public data pool. There were 13 attributes in both datasets, namely, *age, gender, marital status, education, disability, nationality, working hours per week, number of years residing in the current location, ownership of dwelling, family size, number of children, number of automobiles*, and *annual income*. Among these attributes, *marital status* is the only categorical attribute containing more than 2 values, that is, *single*, *married*, and *divorced/widowed*. Because SVMs do not handle categorical features by default, we transformed *marital status* into two binary attributes, *is single* and *is married* (an individual divorced or widowed would have false on both of these attributes). With this transformation, our two datasets had 14 dimensions. For each dataset, we randomly extract a subset of original data as a public data pool, from which public data is sampled uniformly, and use the remaining 30000 tuples as the private data.

Table 1: Experiment parameters.

Parameter	Default value
Number of records in the public data used by hybrid SVM	20
Number of records in the private training dataset	27000
Number of records in the test dataset	3000
Number of dimensions	14
Privacy budget ϵ	1.0

Comparison. We experimentally compared the performance of our hybrid SVM against two approaches, namely, public data baseline and private SVM [25]. The public data baseline is a RBF kernel SVM that uses only public data. In our experiment figures, we use "Public—#" to denote the public data baseline method with # as the size of public data. The private SVM is a state-of-the-art differentially private RBF kernel SVM that uses private data only. The parameters in all methods are set to optimal values.

Metrics. We used the other attributes to predict the value of *annual income* by converting *annual income* into a binary attribute: values higher than a predefined threshold were mapped to 1, and otherwise to −1. Here, we set the predefined threshold as the median value of *annual income*. The classification accuracy was measured by the AUC (the area under an ROC curve) [26]. The boxplot was used to measure the stability of our method and private SVM. The boxplots of "Public—50," "Public—100," and "Public—200," are qualitatively similar to our hybrid SVM; hence, we do not report boxplots of these baseline methods. We performed 10-fold cross-validation 10 times for each algorithm and reported the average results. We varied three different parameters: the privacy budget ϵ, the dataset dimensionality, and the



(a) AUC versus dimensions

(b) Boxplot of hybrid SVM

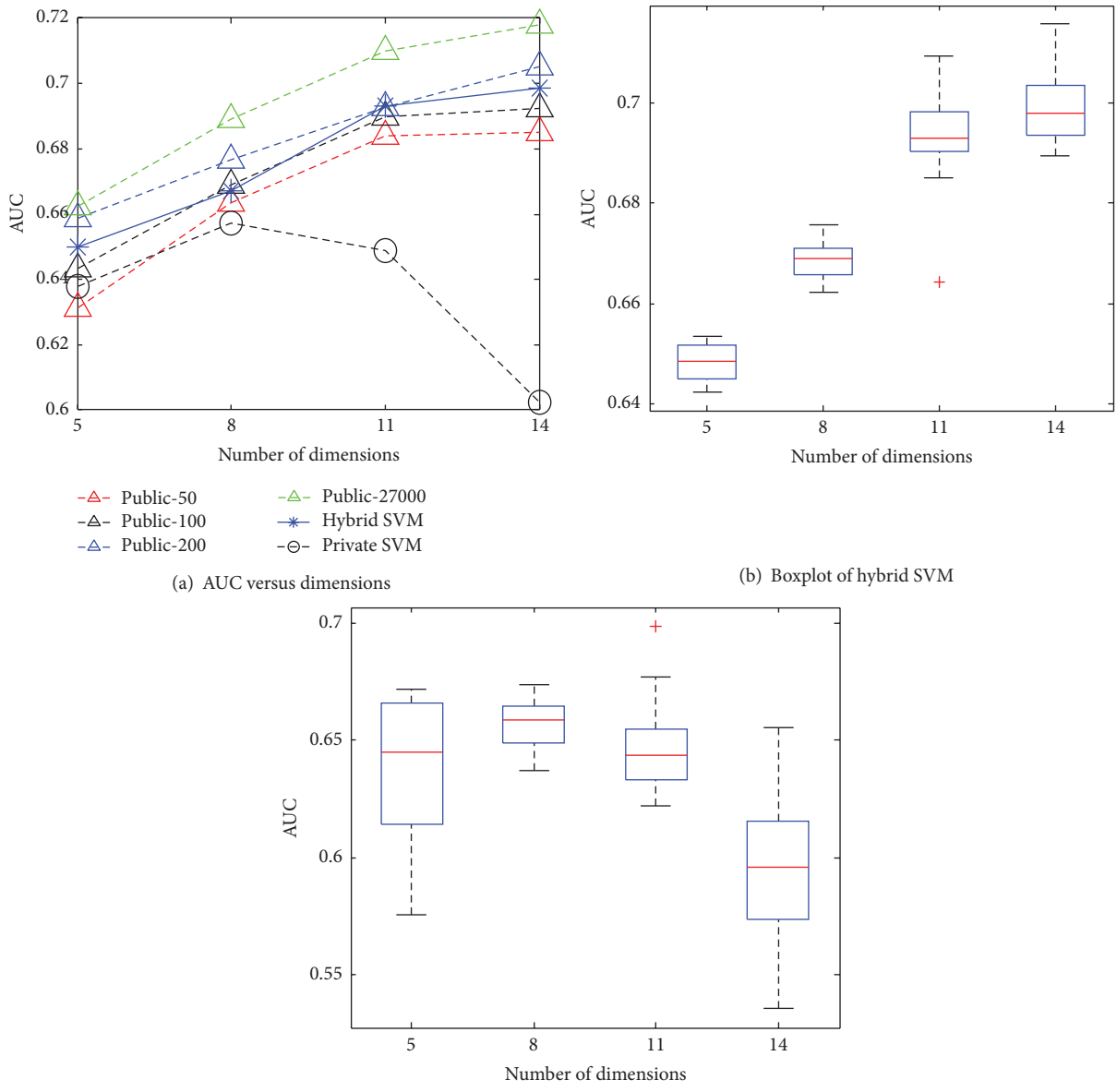
(c) Boxplot of private SVM

FIGURE 5: AUC versus dimensions for US.

data cardinality (i.e., the size of training data). To vary the data cardinality parameter, we randomly generate subsets of records in the training records set, with the sampling rate varying from 0.1 to 1. For various data dimensionalities with the range being 5, 8, 11, and 14, we select three attribute subsets in the US and Brazil datasets for classification. The first five dimensions include: *age*, *gender*, *education*, *family size*, and *annual income*. The second eight dimensions contain the previous five attributes, and additionally *nativity*, *owner of dwelling*, and *number of automobiles*. The third eleven dimensions consist of all the attributes in the second 8 dimensions and *is single*, *is married*, and *number of children*. Table 1 summarizes the parameters and their default values in the experiments.

5.2. AUC versus Privacy Budget. Figures 3 and 4 illustrate the AUCs of each method under various privacy budgets from 0.5 to 4, where "Public—#" means the public data baseline methods with various sizes of public data. Observe that our hybrid SVM outperforms the private SVM and performs better than the public data baseline defined by the public data. The AUC of our method remains stable under all privacy budgets and is significantly close to the public data baseline that uses the complete private data set as public data.

5.3. AUC versus Dataset Dimensionality. Figures 5 and 6 present the AUCs of each algorithm as a function of the dataset dimensionality for the US and Brazil datasets. With



(a) AUC versus dimensions

(b) Boxplot of hybrid SVM

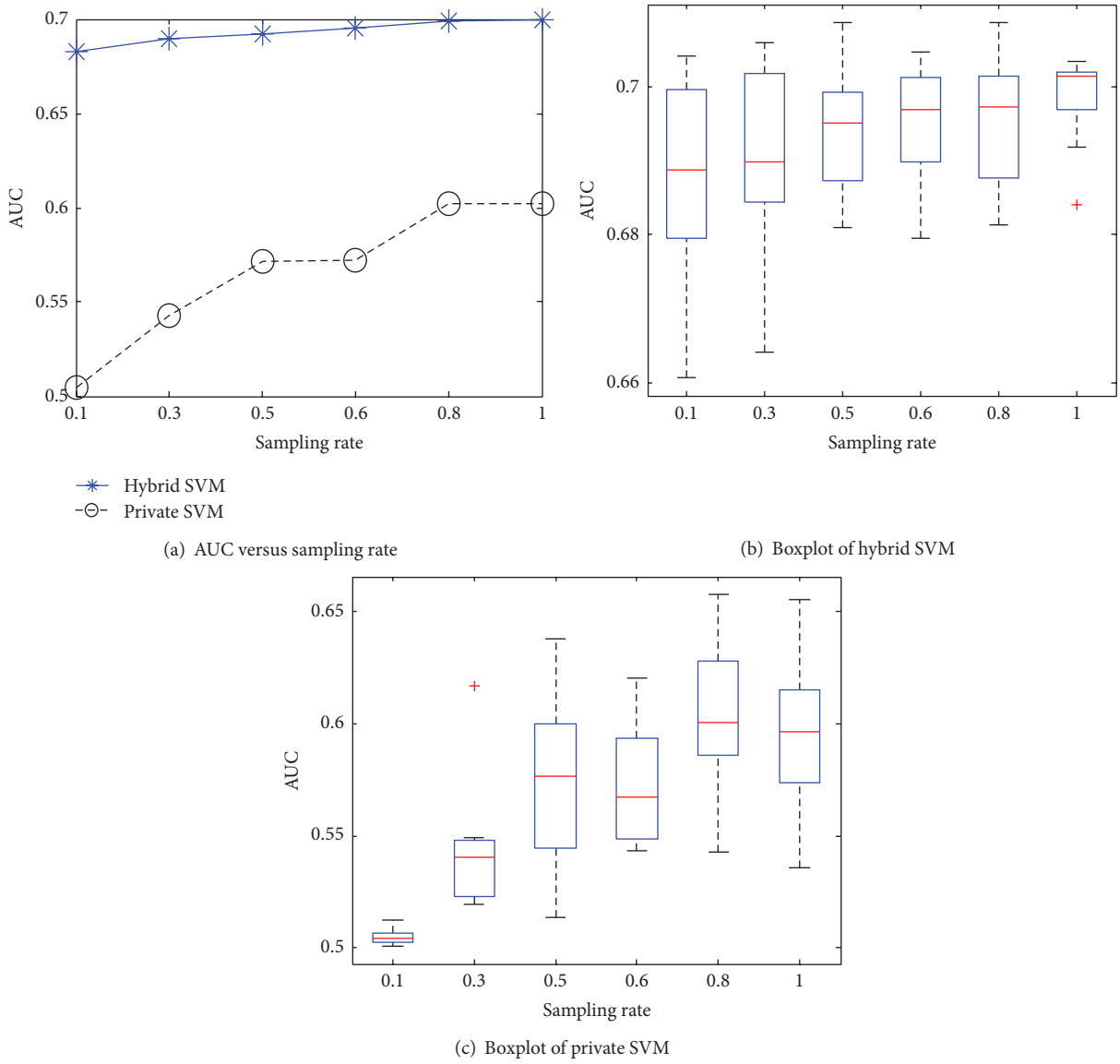
(c) Boxplot of private SVM

Figure 6: AUC versus dimensions for Brazil.

a higher number of dimensions, the AUCs of the hybrid SVM and of the SVM that uses the public data (baseline) increase. This is reasonable because the training data size with the default value being 27,000 is much larger than the number of data dimensions which are at most 14. When the number of dimensions grows, the performance improves. In contrast, the performance of the private SVM degrades in 14 dimensions with poor boxplots because more noise is introduced with higher dimensions.

5.4. AUC versus Data Cardinality. Figures 7 and 8 investigate the relationship between the sampling rate and AUC of hybrid and private SVMs. From the figures, our method consistently outperformed the private SVM at different sampling rates. It is worth mentioning that AUCs of the hybrid SVM are large even at small sampling rates and tend to stabilize when the size of training data grows (i.e., large sampling rate). The boxplots reflect that the private SVM has larger variance than the hybrid SVM, because private SVM selects the values of $\boldsymbol{\rho}$ randomly from the Fourier transform of RBF kernel. In contrast, hybrid SVM computes $\boldsymbol{\rho}$ via the public data. This helps improve the accuracy of $\boldsymbol{\rho}$ and leads to less variance.

5.5. Computation Time. Finally, Figure 9 shows the time cost of our proposed algorithm with varying dimensions and different sampling rates. We only report the results for the US dataset; the results for the Brazil dataset are greatly similar. One can notice that the dimensionality, rather than the sampling rate, determines the computational cost of the hybrid SVM. The overhead of the hybrid SVM is



(a) AUC versus sampling rate

(b) Boxplot of hybrid SVM

(c) Boxplot of private SVM

Figure 7: AUC versus sampling rate for US.

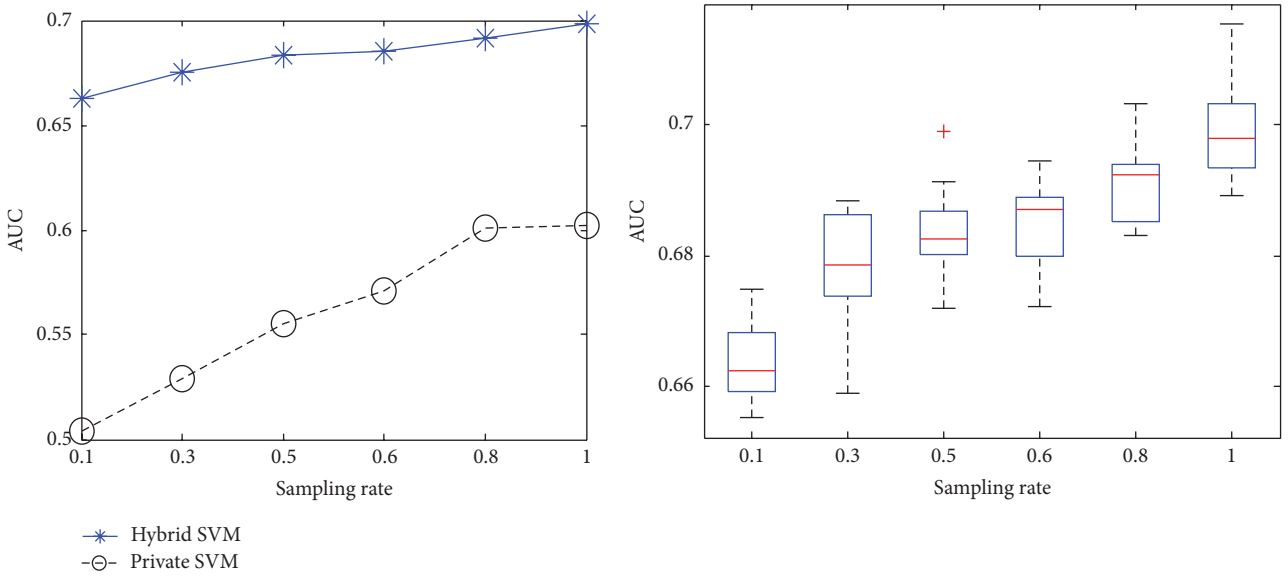
from computing $\boldsymbol{\rho}$ with the public data, since a nonlinear optimization equation needs to be solved. As the other private SVM methods, our hybrid SVM is intended for off-line use, and hence the time is generally acceptable for even 14 dimensional datasets.

6. Discussion and Conclusion

We proposed and developed a RBF kernel SVM using a small amount of public data and a large amount of private data to preserve differential privacy with improved utility. In this algorithm, we use public data to compute the parameters in an approximation form of the RBF kernel function and then train private classifiers with linear SVM after converting all private data into a new feature space defined by the approximation form. A limitation of our approach is that we used the L-BFGS method [27], which is not very efficient, to find the optimal solution. Because the objective function in (6) is not a convex function, our model is computationally intensive in order to calculate the local optimal values, especially when the size of the public data set is large. We will develop more efficient methods and test the model on clinical records in future work. Another limitation is that we assume all original data from different data sets follow some unknown joint multivariate distribution. Our assumption might now always be true in practice, and calibration is necessary for future investigation. That is, in the presence of distributional difference, we will leverage transfer learning to build the global model.

Conflict of Interests

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

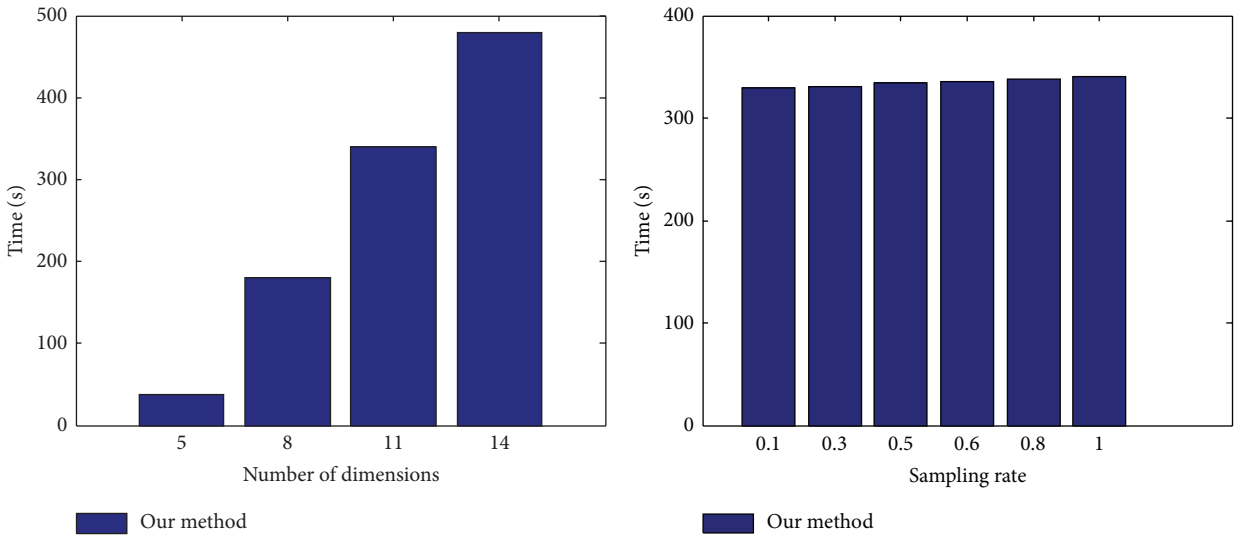


(a) AUC versus sampling rate

(b) Boxplot of hybrid SVM

(c) Boxplot of private SVM

Figure 8: AUC versus sampling rate for Brazil.



(a)

(b)

Figure 9: Time versus dimensions and sampling rate.

Acknowledgments

Lucila Ohno-Machado and Xiaoqian Jiang are partially supported by NLM (R00LM011392) and iDASH (NIH Grant U54HL108460).

References

[1] L. Ohno-Machado, V. Bafna, A. A. Boxwala et al., "iDASH: integrating data for analysis, anonymization, and sharing," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 196–201, 2012.

[2] L. Ohno-Machado, "To share or not to share: that is not the question," *Science Translational Medicine*, vol. 4, no. 165, Article ID 165cm15, 2012.

[3] N. Homer, S. Szelinger, M. Redman et al., "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genetics*, vol. 4, no. 8, Article ID e1000167, 2008.

[4] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying personal genomes by surname inference," *Science*, vol. 339, no. 6117, pp. 321–324, 2013.

[5] D. McGraw, "Building public trust in uses of Health Insurance Portability and Accountability Act de-identified data," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 29–34, 2013.

[6] L. Ohno-Machado, C. Farcas, J. Kim, S. Wang, and X. Jiang, "Genomes in the cloud: balancing privacy rights and the public good," in *AMIA Clinical Research Informatics Summit*, 2013.

[7] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: a survey of recent developments," *ACM Computing Surveys*, vol. 42, no. 4, article 14, 2010.

[8] X. Jiang, A. D. Sarwate, and L. Ohno-Machado, "Privacy technology to support data sharing for comparative effectiveness research: a systematic review," *Medical Care*, vol. 51, no. 8, pp. S58–S65, 2013.

[9] C. Dwork, "A firm foundation for private data analysis," *Communications of the ACM*, vol. 54, no. 1, pp. 86–95, 2011.

[10] C. Dwork, "Differential privacy," in *Encyclopedia of Cryptography and Security*, pp. 338–340, 2nd edition, 2011.

[11] C. Dwork, "Differential privacy: a survey of results," in *Theory and Applications of Models of Computation—TAMC*, pp. 1–19, 2008.

[12] N. Mohammed, X. Jiang, R. Chen, B. C. M. Fung, and L. Ohno-Machado, "Privacy-preserving heterogeneous health data sharing," *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 462–469, 2013.

[13] J. Gardner, L. Xiong, Y. Xiao et al., "Share: system design and case studies for statistical health information release," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 109–116, 2013.

[14] H. Li, L. Xiong, L. Zhang, and X. Jiang, "DPSynthesizer: differentially private data synthesizer for privacy preserving data sharing," in *Proceedings of the 40th International Conference on Very Large Data Bases (VLDB '14)*, Hang Zhou, China, 2014.

[15] S. A. Vinterbo, A. D. Sarwate, and A. A. Boxwala, "Protecting count queries in study design," *Journal of the American Medical Informatics Association*, vol. 19, no. 5, pp. 750–757, 2012.

[16] R. Chen, B. C. M. Fung, N. Mohammed, B. C. Desai, and K. Wang, "Privacy-preserving trajectory data publishing by local suppression," *Information Sciences*, vol. 231, pp. 83–97, 2013.

[17] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu, "Differentially private data release for data mining," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*, pp. 493–501, August 2011.

[18] C. Dwork, F. McSherry, K. Nissim, and A. Smith, ". Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference—TCC*, pp. 265–284, 2006.

[19] H. Li, L. Xiong, and X. Jiang, "Differentially private synthesization of multi-dimensional data using copula functions," in *Proceedings of the 17th International Conference on Extending Database Technology (EDBT '14)*, pp. 475–486, Athens, Greece, 2014.

[20] X. Jiang, Z. Ji, S. Wang, N. Mohammed, S. Cheng, and L. Ohno-Machado, "Differential-private data publishing through component analysis," *Transactions on Data Privacy*, vol. 6, no. 1, pp. 19–34, 2013.

[21] Z. Ji, X. Jiang, S. Wang, L. Xiong, and L. Ohno-Machado, "Differentially private distributed logistic regression using public and private biomedical datasets," in *Proceedings of the 3rd Annual Translational Bioinformatics Conference (TBC '13)*, Seoul, Korea, 2013.

[22] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft, "Learning in a large function space: privacy-preserving mechanisms for svm learning," *Journal of Privacy and Confidentiality*, vol. 4, no. 1, pp. 65–100, 2009.

[23] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *Journal of Machine Learning Research*, vol. 12, pp. 1069–1109, 2011.

[24] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS '07)*, Vancouver, Canada, December 2007.

[25] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft, "Learning in a large function space: privacy preserving mechanisms for SVM learning," *Journal of Privacy and Confidentiality*, vol. 4, no. 1, pp. 65–100, 2012.

[26] T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado, "The use of receiver operating characteristic curves in biomedical informatics," *Journal of Biomedical Informatics*, vol. 38, no. 5, pp. 404–415, 2005.

[27] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming B*, vol. 45, no. 3, pp. 503–528, 1989.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 170289, 10 pages
http://dx.doi.org/10.1155/2014/170289

Research Article

Clinic-Genomic Association Mining for Colorectal Cancer Using Publicly Available Datasets

Fang Liu,[1] Yaning Feng,[1] Zhenye Li,[2] Chao Pan,[1] Yuncong Su,[1] Rui Yang,[1] Liying Song,[1] Huilong Duan,[1] and Ning Deng[1]

[1] *Department of Biomedical Engineering, Key Laboratory for Biomedical Engineering of Ministry of Education, Zhejiang University, Hangzhou 310027, China*
[2] *General Hospital of Ningxia Medical University, Yinchuan 750004, China*

Correspondence should be addressed to Ning Deng; zju.dengning@gmail.com

Received 30 March 2014; Accepted 12 May 2014; Published 2 June 2014

Academic Editor: Degui Zhi

In recent years, a growing number of researchers began to focus on how to establish associations between clinical and genomic data. However, up to now, there is lack of research mining clinic-genomic associations by comprehensively analysing available gene expression data for a single disease. Colorectal cancer is one of the malignant tumours. A number of genetic syndromes have been proven to be associated with colorectal cancer. This paper presents our research on mining clinic-genomic associations for colorectal cancer under biomedical big data environment. The proposed method is engineered with multiple technologies, including extracting clinical concepts using the unified medical language system (UMLS), extracting genes through the literature mining, and mining clinic-genomic associations through statistical analysis. We applied this method to datasets extracted from both gene expression omnibus (GEO) and genetic association database (GAD). A total of 23517 clinic-genomic associations between 139 clinical concepts and 7914 genes were obtained, of which 3474 associations between 31 clinical concepts and 1689 genes were identified as highly reliable ones. Evaluation and interpretation were performed using UMLS, KEGG, and Gephi, and potential new discoveries were explored. The proposed method is effective in mining valuable knowledge from available biomedical big data and achieves a good performance in bridging clinical data with genomic data for colorectal cancer.

1. Introduction

Cancer is one of the major diseases that endanger human life. As American Cancer Society reported, a total of 1,660,290 new cancer cases and 580,350 cancer deaths were projected to occur in the United States in 2013 [1]. In developing countries, such as China, one person is diagnosed with cancer every six minutes, and 8550 people become cancer patients every day [2]. By 2020, the total number of cancer deaths in China is expected to reach 3 million, and the total number of prevalence will reach 6 million [2]. Worldwide, more than 20 million new cancer cases are estimated to be detected by 2030 [3]. Providing much more effective means of early detection and treatment for cancer are still great challenges faced by human beings.

Modern medicine is moving toward the direction of personalized medicine, which refers to the tailoring of medical treatment to the individual characteristics of each patient [4]. Clinical, genetic, protein, and metabolism information of patients are expected to improve the prevention, diagnosis, and treatment of disease together in this medical mode. This will have a great dependence on the successful transformation of basic research results into clinical practice.

With the development of medical informatics and molecular biology, vast amounts of biomedical data have been accumulated. These data cover multiple levels, including both clinical data in macrocosmic aspect and genomic data in microcosmic aspect. However, most clinical data have no corresponding genomic data, while most genomic data have no precise clinical annotation data. Due to the lack of effective linkages, the fruits of basic research have not been translated into clinical practice completely, and problems arising in clinical practice also have not made a big difference to the basic research directions as expected. Exploited value of

available biomedical data is far less than the intrinsic value of these data. Therefore, it can deepen our understanding of the origin and progression of disease, by mining association between clinical data and genomic data from massive available biomedical big data, which promote the bidirectional translation between clinical research and basic research, and ultimately achieve the purpose of promoting the development of personalized medicine.

In recent years, a growing number of researchers began to focus on how to establish associations between clinical data and genomic data. The association between clinical data and genomic data is named as clinic-genomic association in this paper, representing that a clinical feature may have an effect on the gene expression value or the gene may dominate the clinical feature. A persuasive research is the Human Disease Network established by Goh et al. [5]. They extracted 1284 disorders, 1777 disease genes, and associations between these disorders and genes from Online Mendelian Inheritance in Man (OMIM) [6] and then built a bipartite graph using these data. Based on the bipartite graph, they generated two biologically relevant networks, the Human Disease Network and the Disease Gene Network, by assuming that two diseases are connected if they share at least one gene and two genes are connected if they are associated with at least one common disease. Several valuable discoveries are then revealed by these two networks. Other related researches include the Phenome-Genome Network [7], Gene Expression Atlas [8], and iCOD [9]. Excellent works have been done, but there are still many aspects that are needed to be improved. Both Human Disease Network and Phenome-Genome Network take many kinds of diseases into consideration, making it difficult for them to focus on the detail of one certain disease. In addition, only conclusive data, instead of experimental data such as gene expression data, have been utilized by Human Disease Network. Gene Expression Atlas curated the original data submitted by various researches within an experiment instead of comprehensive analysis. While the iCOD only analyzed gene expression data obtained by their own experiments, without considering publicly available datasets. So the data source of iCOD is limited amount. In summary, none of these work mined clinic-genomic associations by comprehensively analysing available gene expression data for a single disease.

Colorectal cancer is the second leading cause of cancer death in the United States and the fifth leading cause in China [10]. Compared with most other cancers, the molecular mechanism of colorectal cancer is relatively clear, making it appropriate for the evaluation of research results. Besides, from the clinical perspective, the outcome of colorectal cancer depends greatly on the stage at which it is detected [11]. The 5-year survival rates of colorectal cancer patients diagnosed at distant stage decrease from 90% to 8%, compared with patients diagnosed at localized stage [11]. However, most clinical symptoms of colorectal cancer arise at a later stage, which greatly impedes the early diagnosis and treatment. If the molecular mechanism of clinical signs or symptoms can be revealed, it would be helpful to detect the molecular change before deteriorating of the disease. Therefore, mining clinic-genomic associations for colorectal cancer is a promising solution to this problem.

To this end, this paper takes colorectal cancer as a typical disease to study how to mine clinic-genomic association for a certain disease using public available datasets, aiming at facilitating the diagnosis and treatment of colorectal cancer. As well, the proposed method can provide a general way for promoting preconized medicine for other disease. The proposed method consists of three parts: extracting clinical concepts using the unified medical language system (UMLS) [12]; extracting genes through literature mining; and mining clinic-genomic associations through statistical analysis. A total of 665 colorectal cancer related clinical concepts, 8392 colorectal cancer related genes, and 23517 clinic-genomic associations were obtained using this method. To evaluate this method and interpret the results obtained, we tried different approaches to make these results more intuitive and understandable. UMLS semantic types were used for clinical concepts analysis, the Kyoto Encyclopaedia of Genes and Genomes (KEGG) [13] pathway for gene analysis, and Gephi [14] visualization for clinic-genomic association analysis. Our investigation provides some interesting findings, such as colorectal cancer related disease (osteoporosis) and related symptoms (angina pectoris), demonstrating that the proposed method can achieve a good performance in bridging clinical data with genomic data as well as mining hidden knowledge from available data.

2. Materials and Methods

2.1. Data Collection and Preprocessing. On one hand, public gene expression data repositories, such as gene expression omnibus (GEO) [15], Stanford microarray database (SMD) [16], and ArrayExpress [8], archive and distribute high-throughput gene expression data submitted by scientific community. Most of these data are accompanied with rich context information including experimental factors and clinical attributes according to the minimum information about a microarray experiment (MIAME) [17] standard, making it ideal for clinic-genomic association mining. As a large database of these, GEO provides the largest gene expression dataset and convenient query and downloading function. Therefore, we took GEO as main data source to perform association mining. On the other hand, with large number of research papers published, several databases integrating research results were established, such as OMIM and genetic association database (GAD) [18]. Data in these databases are generally acknowledged and can be used for evaluation and validation. Compared with OMIM, GAD collects data with less restrict limitations, leading to richer data records. So we take GAD as assistant data source to supplement and evaluate association mining results.

We accessed the GEO site on February 3, 2013. Colorectal cancer related GEO series (GSE) were preliminarily identified as those that passed the custom filter rule [see Supplementary Table S1 (Supplementary Material available online at http://dx.doi.org/10.1155/2014/170289)]. Search statements were constructed using GEO Datasets Advanced Search Builder [19], which aims to perform more refined queries in order to filter down to the most relevant data. A total of

628 GSE were found out and downloaded in simple omnibus format in text (SOFT) format from FTP site of GEO use Aspera Connect tool.

All sample data tables and platform data tables of a GSE are stored in a single SOFT file. Note that it is quite inconvenient and inefficient to read the generally huge line-based, plain text format file each time we parsed 628 GSE files into several sample table files and platform table files, with each sample table file holding data for a certain GEO sample (GSM) and each platform table file holding data for a certain GEO platform (GPL). Most of the clinical information is located in title, source, species, characteristics, and descriptions fields of GSM annotations. We developed in-house Perl program to extract these information into relational database for further analysis. GSM not from human beings and GSM without any keyword of colon, rectum, rectal, hepatic flexure, or sigmoid in the extracted annotations were eliminated during the extracted information. Sample data tables index expression measurements of multiple RNA transcripts with Probe Set IDs, while the external gene identifiers, names, and symbols were stored in platform data tables. In order to allow the same gene measured in different platforms to be unified, Probe Set IDs were mapped to the HUGO (human genome organization) symbols. Since platforms are made by different manufactories and platform data tables are provided correspondingly, UniGene symbols appear in different column of platform data tables irregularly or even disappear, making automatically mapping from Probe Set IDs to gene symbols quite difficult. Thus we implemented this procedure manually in order to make full use of platform data tables. As for platforms not providing HUGO symbols, IDconverter [20] was used. Platform map files were stored in Map Containers format, a data structure of MATLAB, for further use.

GAD was accessed on April 18, 2012. We used keyword-search function of GAD query tool to obtain colorectal cancer related records. Setting search field to "disease" and entering one of colorectal cancer related keywords [see Supplementary Table S2] each time, a total of 4784 records were picked out and downloaded into an excel file for subsequent processing.

2.2. Extracting Clinical Concepts Using UMLS. GEO annotations are in free-text format and a certain concept is frequently presented in different ways by different scientists, making it difficult to organize and compare data generated from different research institutions. Taking "colon cancer" as an example, it can be descripted as "colon cancer," "colon carcinoma," "human carcinoma colon cell," and so on. In this case, ontology is urgently needed to link these various descriptions together. UMLS is the largest thesaurus in the biomedical domain, collecting biomedical concepts from controlled vocabularies and classifications used in patient records, administrative health data, bibliographic databases, and so on. Each concept is annotated with at least one semantic type from a semantic network that broadly covers the medical domain [21]. In order to extract unified colorectal cancer related clinical concepts effectively, we mapped free-text annotations to UMLS concepts. The process flow is shown in Figure 1.

First, a Java program was developed to map the "characteristics" and "description" field to UMLS concepts by calling MetaMap (a program developed at National Library of Medicine) API [22]. Since "Source," "Species," and "Title" contain only some identification information, they were ignored in this process. Concept CUIs, concept names, sematic types, and corresponding original phrases were specified to be outputted and recorded in relational database. Second, we used clinical related semantic types [23] to screen for clinical concepts. Those concepts belonging to any semantic type in Supplementary Table S3 were kept, while the others were excluded. To improve the accuracy of mapping and thus reduce the burden of future work, a manual elimination of incorrectly mapped concepts was performed with the help of the recorded original phrases.

2.3. Extracting Genes through Literature Mining. The genetic factors leading to colorectal cancer have been extensively studied, and a large numbers of research papers have been published on the subject. The large body of published biomedical literature is one of the richest data sources for systematically identifying colorectal cancer related genes without microarray expression experiment. In order to obtain nontrivial knowledge quickly and accurately, we took available literature-mining achievements as a data source instead of performing literature mining algorithm directly. GAD was employed in this paper and colorectal cancer related records in GAD has already been picked out and curated in an Excel sheet in Section 2.1. Every record in GAD reflects an association between a gene with a disease through "association" attribute, with "Y" indicating "associated," "N" indicating "not associated," and blank indicating "uncertain." We extracted genes from the "Gene" column and recorded the corresponding association values.

2.4. Mining Clinic-Genomic Associations through Statistical Analysis. We proposed a statistical-analysis-based clinic-genomic association method for colorectal cancer. The procedure is shown in Figure 2. Clinical information acquired with the help of UMLS was used as data inputs to obtain the corresponding genomic information. For each concept, GSM can be divided into two groups depending on whether the concept was extracted from their annotations. Data group A contains gene expression data of GSM of which annotations contain the concept, while data group B contains data of GSM of which annotations do not contain the concept. Data from different GSEs or different GPLs were heterogeneous regarding of their measuring technologies, measured genes, and preprocessing methods, making it meaningless to analysis them together directly. So we further grouped these data into several subsets according to their GSE and GPL. Each data subset contains data from both group A and group B. Data subset with less than four GSMs involved by data group A or data group B are eliminated in the consideration of significance consideration of statistical analysis.

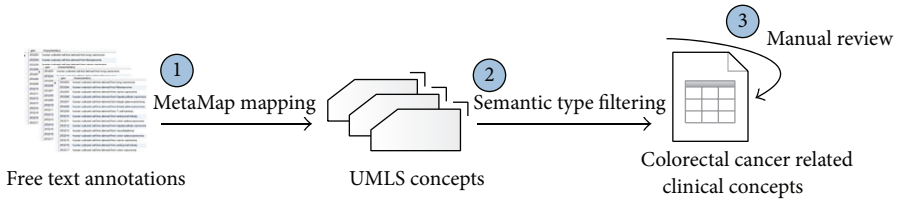


Figure 1: Extracting clinical concepts using UMLS. Three steps were involved. First, free-text annotations were mapped to UMLS concepts using MetaMap. Second, clinical concepts were screened out by semantic types. Finally, a manual review was performed to emitted mapping errors.

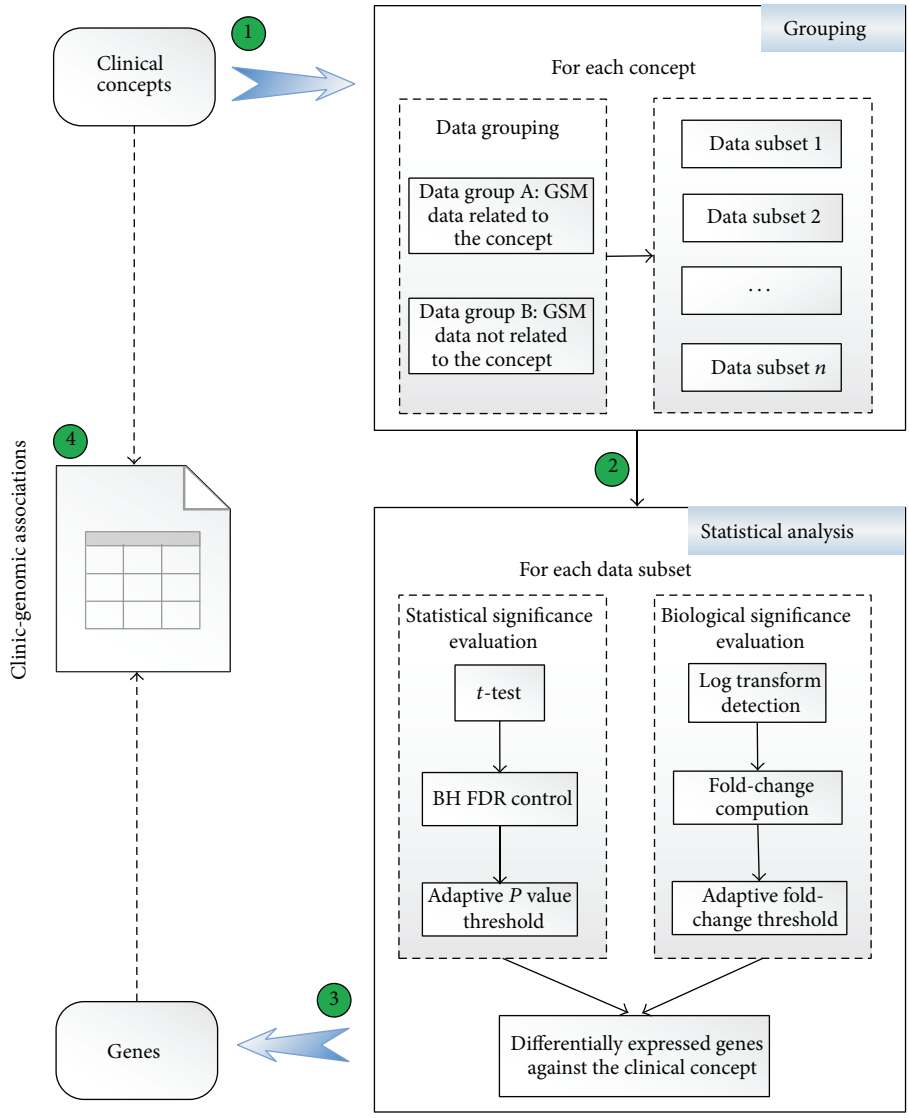


Figure 2: Statistical-analysis-based association mining flow. Four steps were involved. First, for each concept, GSM data were divided into two groups and further organized into different data subsets based on GSE and GPL. Second, for each data subset, differentially expressed genes were screened out according to statistical significance and biological significance. Third, for each concept, differentially expressed genes from every data subset were integrated. Finally, a series of associations were established between each concept with the corresponding differentially expressed genes.

Hundreds of data subsets need to be analysed and some analysis data subsets contain more than 400 samples. Such a heavy computation burden imposes great challenges on most computation tools. MATLAB is a software with powerful computing capabilities. Most importantly, bioinformatics toolbox of MATLAB is packed with a series of robust and well-tested functions, providing an integrated software environment for genome and proteome analysis. Based on above considerations, we use MATLAB to implement the proposed association mining method.

2.4.1. Statistical Significance Evaluation. To explore genes that are differentially expressed in data group A relative to data group B, we first made a hypothesis that all genes are equally expressed in these two data groups and then we used the "mattest" function, which is provided by the bioinformatics toolbox of MATLAB for classical t-test, to test our hypothesis. A list of P values presenting the significance of differential expression between data group A and data group B were figured out. To control the overall probability of type I error, these raw P values must be adjusted for multiple testing. Among all adjustment methods for multiple comparisons, the Benjamini and Hochberg procedure [24], abbreviated as "BH FDR," controlling the proportion of false positives among the genes called as differentially expressed, is probably more appropriately for datasets with very large numbers of genes [25]. Therefore, in this case study, BH FDR control was implemented using the "mafdr" function by setting the "BHFDR" parameter to be "TRUE." A series of adjusted P values was obtained. The default P value threshold was 0.05. But if more than 1% of all the measured genes are positive in this case, the threshold will automatically shift to keep only the most significant 1% genes.

2.4.2. Biological Significance Evaluation. Fold-change is defined as the average expression over all samples in a condition divided by the average expression over all samples in another conditions. The average expression should be in constant scale rather than logarithmic or exponential scale. Diverse preprocessing methods were used to obtain the preprocessed data. Some data have been logarithmic transformed, while others not. Most of the processed data did not note the used scale explicitly. So we put forward the following algorithm to detect whether the input data were in log scale. As a general rule, if the scale is around 0 to 16, it is in log scale; if it is around 0 to 40000, it is in original scale. Quantile value was computed to explore data distribution range. The original algorithm refers to GEO2R [26] and the main improvement is using mean value to avoid big noise. MATLAB code was presented in Supplementary Table S4. When input data was identified as in log scale, NeedLogC was set to false. Otherwise, NeedLogC was set to true. Fold-change was computed using the modified "mavolcanoplot" function. NeedLogC was transferred to "mavolcanoplot" function as the parameter value of "LogTrans." The returned fold-change value has been processed. Positive value means upregulated, while negative means downregulated. The default fold-change threshold was 2. But if none of the measured genes were positive, the threshold would automatically shift to keep the largest 1% absolute fold-change values (but no less than 1.5).

Only genes passing both the P value threshold and fold-change threshold were considered as differentially expressed against the clinical concept used to group data. For each concept, genes obtained from all data subsets were combined together to form a set of clinic-genomic associations.

Table 1: Top 15 concepts related to the most number of GSM.

Concept name	GSM count	Rank
Medical history	1004	15
Family history	1002	16
Instability	907	19
Microsatellite repeat	895	20
Recurrence	844	21
Protein p53	824	22
Histology procedure	643	27
Death	570	29
Microsatellite instability	547	32
Primary neoplasm	514	34
Tobacco use	446	36
Encounter due to tobacco use	446	37
Ethnic	446	38
Leukaemia	434	40
Encounter due to therapy	402	41

3. Results and Discussion

3.1. Clinical Concept Datasets. A total of 665 colorectal cancer related clinical concepts, see Supplementary Table S6, were obtained using the UMLS-based method. About 115 concepts (14.5%) resulted from incorrect mapping had been ruled out during the manual review process. The most common type of mapping errors come from abbreviations, including "pain" from "pn," "Edema" from "ED." Semantic type mistakes were also found out, such as "Dukes Disease" from "Dukes Stage."

Clustering clinical concepts based on semantic types and counting the number of concepts in each semantic type can provide us with an intuitive view about what is most concerned in clinical studies of colorectal cancer. Distribution of concepts acquired in this paper among the 20 semantic types was obtained; see Supplementary Figure S1. Neoplastic process, biologically active substance, finding, and disease or syndrome cover more than half of all concepts. The number of GSM related to each concept reflects the importance degree of the concept to some extent. The top 15 concepts relating to the most number of GSM were presented in Table 1, after ignoring general concepts like carcinoma, colon carcinoma, malignant neoplasms, and others similar. According to Table 1, medical history, family history, microsatellite instability, and tobacco use are important clinical information focused in colorectal cancer clinical research.

3.2. Genomic Datasets. From GAD, we got 904 genes, of which 247 annotated with "Y," 159 with "N," and 823 with "Blank" (overlap exists among these three cases). Association value indicates whether a gene is associated with a disease or not. However, it is not unique for some genes, due to the fact that each value in GAD depends on a single paper, while different papers may have different conclusions. Also due to this, we did not concern the specific association value in the following analysis, but just assume that these genes are related to colorectal cancer somehow. From GEO, we got 7914 genes which are extracted from our mining results of clinic-genomic associations. A total of 8392 genes [see Supplementary Table S7] were obtained from these two data sources after removing duplicates.

3.2.1. Relevance between Genes and Colorectal Cancer. Genes from GAD are extracted from the published literatures, and genes from GEO are picked out according to statistical analysis. The former is more reliable but less abundant, while the latter is just on the contrary. Genes from GEO are extracted from clinic-genomic associations, of which each one was deduced from gene expression data of one or more GSE. If we impose a different restriction on the number of association related GSE, we can get different number of genes. For instance, we got only 1687 genes after requiring more than 1 related GSE. The overlapping rate between genes from GEO and genes from GAD also varies with different restrictions. Assume genes from GAD are reliable, the overlapping rate reflects the relevance between genes from GEO with colorectal cancer to a certain extent. A method to calculate the relevance quantitatively using the overlapping rate was defined as

$$\text{Relevance score} = \frac{\text{Overlapping rate}}{\text{Proportion}}. \quad (1)$$

Here, "Relevance score" is the quantitative evaluation of association degree between genes with colorectal cancer, "Overlapping rate" is the proportion of overlapping genes in genes from GEO, and "Proportion" is the proportion of genes from GEO in all genes.

Figure 3 demonstrates the relationship between "Relevance score" and the number of relation related GSE. It can be seen that the "Relevance score" increases with the increases of related GSE numbers. This trend can be interpreted from the following perspective. Related GSE are data foundation of clinic-genomic associations. Therefore, more related GSE indicates much more reliable clinic-genomic associations about colorectal cancer and thus the closer association between genes and colorectal cancer.

3.2.2. KEGG Pathway Analysis. To further interpret the obtained results, we can link genomic information with higher order functional information. KEGG is the right knowledge base for systematic analysis of gene functions [13]. Genes are inputted in online analysis tool of KEGG pathway database. Top 10 pathways, see Supplementary Table S5, covering the most amounts of input genes are obtained, including pathways in cancer, proteoglycans in cancer, and

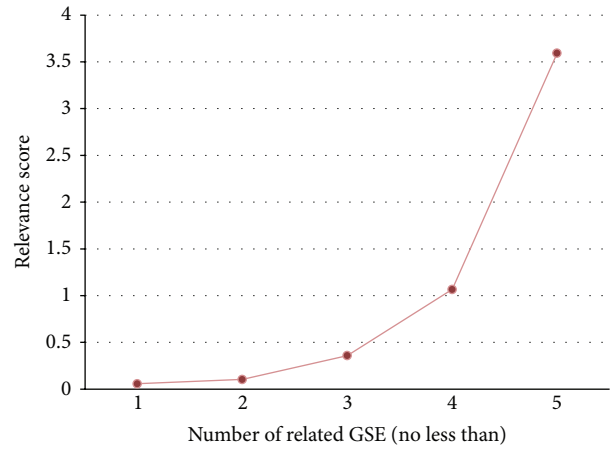


Figure 3: Quantitative evaluation of association degree between genes from GEO with colorectal cancer. The horizontal axis presents the number of association related to GSE, while the vertical axis presents the relevance score computed using formula (1).

Table 2: The 51 genes of acquired results involving the colorectal cancer pathway.

AKT1	*BIRC5*	*KRAS*	*MSH6*	*RHOA*	*TP53*
AKT2	*BRAF*	*LEF1*	*MYC*	*SMAD2*	*BCL2*
AKT3	*CASP3*	*MAP2K1*	*PIK3CA*	*SMAD3*	*JUN*
APC	*CASP9*	*MAPK1*	*PIK3CD*	*SMAD4*	*MSH3*
ARAF	*CCND1*	*MAPK10*	*PIK3R1*	*TCF7*	*RALGDS*
AXIN1	*CTNNB1*	*MAPK8*	*PIK3R2*	*TCF7L1*	*TGFBR2*
AXIN2	*CYCS*	*MAPK9*	*PIK3R5*	*TCF7L2*	*TGFBR1*
BAD	*DCC*	*MLH1*	*RAC2*	*TGFB1*	*RAC3*
BAX	*FOS*	*MSH2*			

focal adhesion pathway. As for focal adhesion pathway, the literature [27] has shown that both primary colorectal cancers and colorectal liver metastases express high levels of FAK (focal adhesion kinase) mRNA and p125 FAK protein. In addition, 51 of the acquired genes involve the colorectal cancer pathway. These genes are listed in Table 2, including important oncogenes (*KRAS* and *CTNNB1*), tumour suppressors (*APC, DCC, TP53, BAX, SMAD2, SMAD4,* and *TGFBR2*), and DNA repair genes (*MLH1, MSH2, MSH3,* and *MSH6*).

3.3. Clinic-Genomic Association Datasets. A total of 23517 associations between 7914 genes and 139 concepts were found out. All the associations are listed in Supplementary Table S8. Such a massive amount of associations is difficult to evaluate or interpret directly. Therefore, we used two methods to gain a more profound understanding of these associations. First, we use visualization as a powerful means to leverage the perceptual abilities of humans to find useful information from obtained associations. Shape, colour, distance, and other elements can all be used to corroborate understanding of network. In this paper, Gephi was performed visualization

analysis on clinic-genomic associations from different perspectives, including overall view, data-source-feature view, and semantic-type view. Clinical concepts were inputted into Gephi as source nodes, genes as destination nodes, and absolute value of fold-changes as weight of edges. Associations are clustered by the default modularize method, fast unfolding of communities in large networks, and different classes were presented in different colours. Second, the number of association related GSE was taken as a determinant to recognize highly reliable associations.

3.3.1. Overall View. In this view, all associations were imported to Gephi together. Force atlas was used as layout algorithm. The analysis result was shown in Supplementary Figure S18 and simplified version of the analysis result was shown in Figure 4. The result confirms the complexity of these associations: a gene may relate to multiple clinical concepts and also a clinical concept may associate with multiple genes. But through this visualization method, several important concepts come into sight clearly. These include malignant neoplasms, neoplasm, adenoma, adenocarcinoma, and Protein p53, suggesting that they are connected with large number of genes and they are focused concepts in colorectal cancer research.

Besides, Figure 4 reflects the development process between adenoma and adenocarcinoma. The nearest location in the figure expresses their close relationship, while the bigger size of adenoma node relative to adenocarcinoma node represents that there are more clinic-genomic associations with adenoma. Previous knowledge shows that the development process of colorectal cancer can be divided into several stages, including normal mucosa, adenoma,and adenocarcinoma, noting that adenoma appears earlier than adenoma carcinoma. In clinical practice, patients are always diagnosed based on their clinical symptoms, vital signs, and so forth in an early stage due to the slow arising of colorectal cancer symptoms, resulting in inadequate data from early patient. However, the proposed method in this paper is able to uncover more knowledge about the early stage of colorectal cancer. This lays a good knowledge foundation for the research on the early diagnosis and treatment for colorectal cancer and may reflect the significance of the proposed method to the personalized medicine.

3.3.2. Data-Source-Feature View. Due to various features of data source, statistical analysis results of certain data group subsets pass P value threshold or fold-change threshold easily, resulting in that some clinical concepts connect with lots of genes. For example, malignant neoplasm links with 1917 genes. From this point, it is inappropriate to simply unite all results together because concepts with little genes will be buried in the visualization results. To highlight the importance of each concept, we reduced the number of related genes to no more than 10. A total of 1075 associations between 139 concepts with 524 genes remained, of which the Gephi outputs was shown in Figure 5. Dual Circle Layout was used in this view for the comparable magnitude of gene and concept number. Nodes of inner circle represent clinical concepts, while nodes of outer circle represent genes. In Figure 5, certain genes, including *ADAMDEC1, ABCC2, ABCA8, ACTG2, ACSL6, LOC728448, TCERG1,* and *ENOSF1* reveal their importance. Among them, *ABCC2* was also included in genes extracted from GAD with blank association value, indicating the potential of complementing GAD with results of statistical analysis method.

3.3.3. Semantic-Type View. Classifying clinic-genomic associations based on semantic type of clinical concepts is helpful to get a deeper understanding of associations involved by each semantic type. We analysed all of the 20 semantic types, respectively. Selected results have been presented in Supplementary Figure S2~S17. Lots of meaningful rediscoveries as well as interesting new findings were obtained. In particular, two typical semantic types, "Disease or Syndrome" and "Sign or Symptom," are illustrated specifically in this paper in detail.

The "Disease or Syndrome" semantic type covers 676 associations between 10 clinical concept and 445 genes. The Gephi analysis results is shown in Supplementary Figure S6, from which the general acknowledged colorectal cancer related diseases, including inflammatory disease, irritable bowel syndrome, intestinal disease, and inflammatory bowel disease are very conspicuously. Besides, the "Osteoporosis" concept also comes into view. It is not broadly known to the public as colorectal cancer related disease, but some researchers have claimed that osteoporosis is associated with the risk of colorectal adenoma in women recently [28]. Based on statistical-analysis-based association mining method, 22 genes were identified as osteoporosis related. The top 10 genes ordering by P value are *C1orf173, TTC23, BCR, TEF, RAP1GAP, SLC45A4, CCDC66, CRISPLD2, IRX5,* and *SAPS1*. Among them, the association between *TTC23* and osteoporosis is also reported by GeneCards [29].

The "Sign or Symptom" semantic type includes 393 associations between 10 clinical concept and 393 genes. The Gephi output is shown in Supplementary Figure S17. In addition to abdominal bloating, the most obvious one, other familiar signs or symptoms, like red stools, diarrheal, constipation, change in bowel habit, and vomiting and nausea, also have a place in Supplementary Figure S2~S17. Furthermore, "Angina Pectoris" appears out of expectation. It is not a common symptom of colorectal cancer, but the eHealthMe website displays a group of data from colon cancer patients who have angina pectoris [30], indicating that angina pectoris probably has potential association with colorectal cancer.

3.3.4. Recognition of Highly Reliable Associations. Different GSE, generally submitted to GEO by different researchers, are basically irrelevant. Therefore, it is of small possibility that an association was obtained by accident if the association can be deduced from more than one GSE. This point was also illustrated in Section 3.2.1 as more related GSE indicates much more reliable clinic-genomic association. Counting the number of association related GSE for every association and then restricting the number to more than one, we got 3474 associations between 31 clinical concepts with 1689 genes. These associations are considered as highly reliable

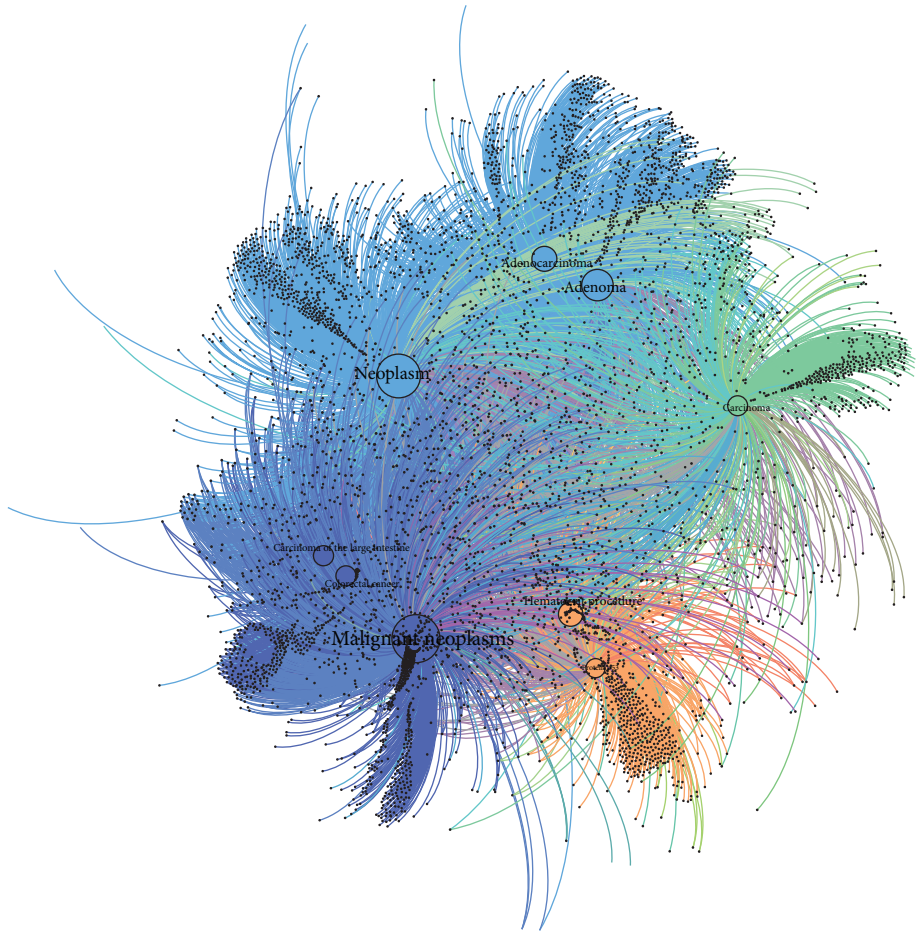


Figure 4: Simplified version of overall view of clinic-genomic associations. Complete version can be found in Supplementary Figure S18. This simplified version was generated by ignoring several unobvious concepts and genes to reflect important associations much clearer.

associations and have been marked out in Supplementary Table S8.

Generally, this paper proposed a method to mine associations between clinical data and genomic data using publicly available datasets, which is a great mission in the era of big data. We focused on a typical disease, colorectal cancer, to learn the potential of the vast amounts of existing biomedical data. Colorectal cancer related symptoms, diseases or syndromes, neoplastic processes, and other clinical features have all been covered in this research. This is a novel exploration for little researchers having done such a thorough work for a single disease using this mode. Outcome was appreciated, but there are still lots of space for improvement. First, clinical concepts are regarded as independent during the statistical analysis process to reduce the complexity. Therefore, thoughtful measures should be taken to guarantee accuracy. Second, as an exploration, we only take a representative database, GEO, as data source. Much more datasets could be involved in the future study. Last, to make good use of the association mining results and to share the association mining methods with peer researchers, a publicly available platform would be helpful. For this consideration, such a platform is in process now.

4. Conclusions

Aiming at facilitating the diagnosis and treatment of colorectal cancer and also providing a general way for promoting preconized medicine for other disease, this paper proposed a clinic-genomic association mining method for colorectal cancer, which consists of three parts: extracting clinical concepts using UMLS; extracting genes through literature mining; and mining clinic-genomic associations through statistical analysis. Using the proposed method, 23517 clinic-genomic associations between 139 clinical concepts and 7914 genes were obtained. Moreover, 3474 of all these associations, relating 31 clinical concepts with 1689 genes, were identified as highly reliable based on the number of association related GSE. Lots of results have been validated and there are also several new discoveries, including colorectal cancer related disease (osteoporosis) and related symptoms (angina pectoris), demonstrating the correctness and usefulness of the proposed method. These results can be shared with clinical researchers and basic researchers as well as translational researchers to suggest new study directions or to answer some unsettled questions. As bridges between clinical researches and genomic researches, these associations would be helpful

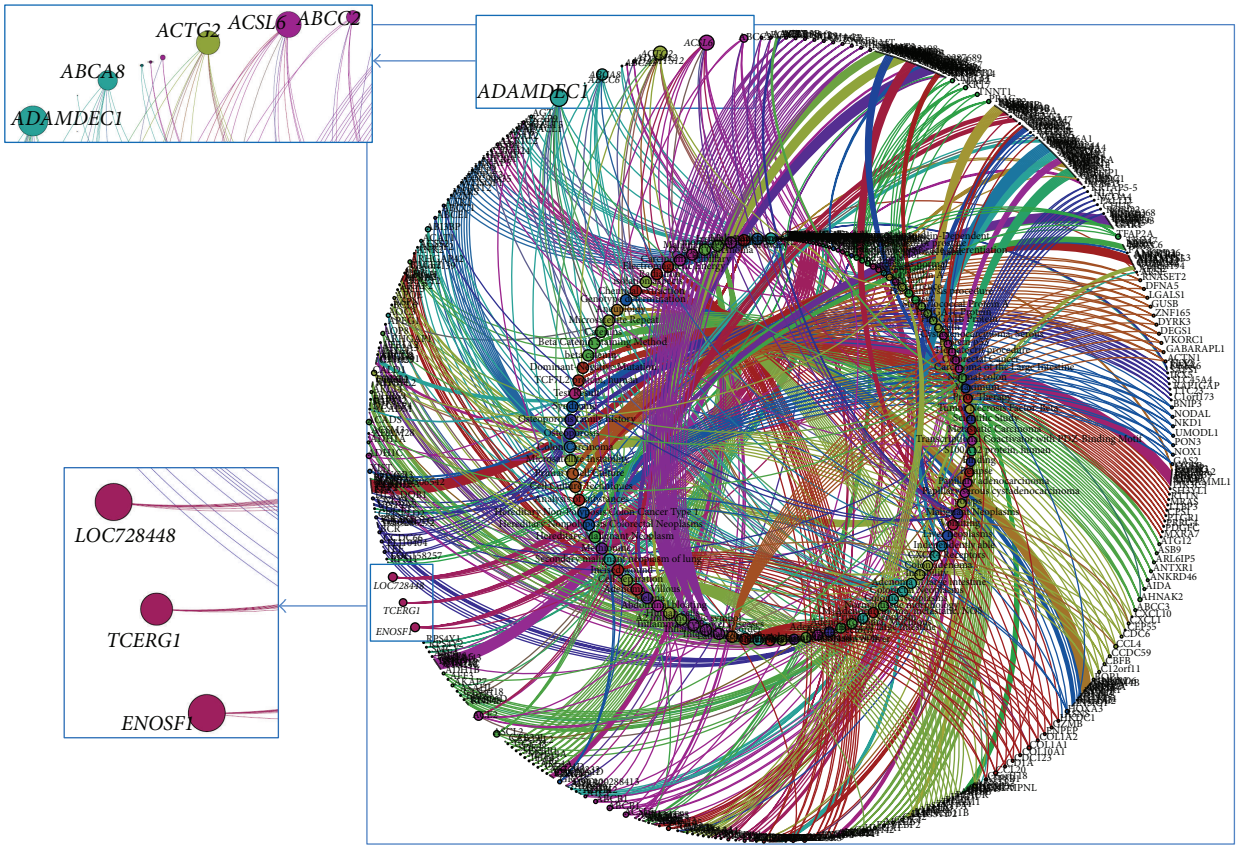


Figure 5: Data-source-feature view of obtained clinic-genomic associations. The number of related genes for each concept was limited to no more than 10. Dual Circle Layout algorithm was employed in Gephi. Several genes stood out from this view.

to accelerate the bidirectional translation between these two fields. Besides, this method can also be transplanted to analyse other diseases, such as breast cancer and liver cancer. In the future work, we will expand the data sources, blending in ArrayExpress, SMD, to enrich our results. On the other hand, expanding knowledge of clinical concepts by combining UMLS-based concepts with electronic medical records will be an appropriate direction of our research.

List of Abbreviations

GAD: Genetic association database
GEO: Gene expression omnibus
GPL: GEO Platform
GSE: GEO series
GSM: GEO sample
KEGG: Kyoto Encyclopaedia of Genes and Genomes
OMIM: Online Mendelian Inheritance in Man
SMD: Stanford microarray database
SOFT: Simple omnibus format in text
UMLS: Unified medical language system
HUGO: Human genome organization.

Conflict of Interests

The authors declare that they have no conflict of interests.

Acknowledgments

This work was supported by the National High Technology Research and Development Programs of China (863 Programs, no. 2012AA02A601 and no. 2012AA020201), the National Science and Technology Major Project of China (no. 2013ZX03005012), and the National Natural Science Foundation of China, no. 31100592.

References

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

[2] J. Hao and W. Q. Cheng, *Annual Cancer Registration of China*, Military Medical Science Press, 2012.

[3] F. Bray, A. Jemal, N. Grey, J. Ferlay, and D. Forman, "Global cancer transitions according to the Human Development Index (2008–2030): a population-based study," *The Lancet Oncology*, vol. 13, no. 8, pp. 790–801, 2012.

[4] E. Abrahams, G. S. Ginsburg, and M. Silver, "The personalized medicine coalition: goals and strategies," *American Journal of PharmacoGenomics*, vol. 5, no. 6, pp. 345–355, 2005.

[5] K. I. Goh, M. E. Cusick, D. Valle et al., "The human disease network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–8690, 2007.

[6] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, "Onlined Mendelian Inheritance in Man

(OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 30, no. 1, pp. 52–55, 2002.

[7] A. J. Butte and I. S. Kohane, "Creation and implications of a phenome-genome network," *Nature Biotechnology*, vol. 24, no. 1, pp. 55–62, 2006.

[8] A. Brazma, H. Parkinson, U. Sarkans et al., "ArrayExpress—a public repository for microarray gene expression data at the EBI," *Nucleic Acids Research*, vol. 31, no. 1, pp. 68–71, 2003.

[9] K. Shimokawa, K. Mogushi, S. Shoji et al., "ICOD: an integrated clinical omics database based on the systems-pathology view of disease," *BMC Genomics*, vol. 11, no. 4, article S19, 2010.

[10] N. Deng, L. Zheng, F. Liu, L. Wang, and H. Duan, "CrcTRP: a translational research platform for colorectal cancer," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 930362, 9 pages, 2013.

[11] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun, "Cancer statistics, 2009," *CA Cancer Journal for Clinicians*, vol. 59, no. 4, pp. 225–249, 2009.

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

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

[14] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," in *International AAAI Conference on Weblogs and Social Media*, 2009.

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

[16] G. Sherlock, T. Hernandez-Boussard, A. Kasarskis et al., "The stanford microarray database," *Nucleic Acids Research*, vol. 29, no. 1, pp. 152–155, 2001.

[17] 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.

[18] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, "The genetic association database," *Nature Genetics*, vol. 36, no. 5, pp. 431–432, 2004.

[19] S. E. Wilhite and T. Barrett, "Strategies to explore functional genomics data sets in NCBI's GEO database," *Methods in Molecular Biology*, vol. 802, pp. 41–53, 2012.

[20] A. Alibés, P. Yankilevich, A. Cañada, and R. Díaz-Uriarte, "IDconverter and IDClight: conversion and annotation of gene and protein IDs," *BMC Bioinformatics*, vol. 8, article 9, 2007.

[21] G. Divita, T. Tse, and L. Roth, "Failure analysis of MetaMap Transfer (MMTx)," *Medinfo*, vol. 11, no. 2, pp. 763–767, 2004.

[22] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," *Proceedings of the AMIA Symposium*, pp. 17–21, 2001.

[23] J. Dudley and A. J. Butte, "Enabling integrative genomic analysis of high-impact human diseases through text mining," *Pacific Symposium on Biocomputing*, pp. 580–591, 2008.

[24] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, and I. Golani, "Controlling the false discovery rate in behavior genetics research," *Behavioural Brain Research*, vol. 125, no. 1-2, pp. 279–284, 2001.

[25] D. M. Dziuda, *Data Mining For Genomics and Proteomics: Analysis of Gene and Protein Expression Data*, Wiley, 2010.

[26] 2013, NCBI, GEO2R, http://www.ncbi.nlm.nih.gov/geo/geo2r/.

[27] A. L. Lark, C. A. Livasy, B. Calvo et al., "Overexpression of focal adhesion kinase in primary colorectal carcinomas and colorectal liver metastases: immunohistochemistry and real-time PCR analyses," *Clinical Cancer Research*, vol. 9, no. 1 I, pp. 215–222, 2003.

[28] J. U. Lim, J. M. Cha, J. I. Lee et al., "Osteoporosis is associated with the risk of colorectal adenoma in women," *Diseases of the Colon and Rectum*, vol. 56, no. 2, pp. 169–174, 2013.

[29] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, "GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support," *Bioinformatics*, vol. 14, no. 8, pp. 656–664, 1998.

[30] ehealthme, Review: could Angina pectoris cause Colon cancer? 2013, http://www.ehealthme.com/cs/angina+pectoris/colon+cancer.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 901428, 9 pages
http://dx.doi.org/10.1155/2014/901428

Research Article

Identification of MicroRNAs as Potential Biomarker for Gastric Cancer by System Biological Analysis

Wenying Yan,[1,2] Shouli Wang,[3] Zhandong Sun,[4] Yuxin Lin,[4] Shengwei Sun,[5] Jiajia Chen,[4,5] and Weichang Chen[1]

[1] *Department of Gastroenterology, The First Affiliated Hospital of Soochow University, Suzhou 215006, China*
[2] *Taicang Center for Translational Bioinformatics Center for Systems Biology, Taicang 215400, China*
[3] *Department of Pathology, Soochow University School of Medicine, Suzhou 215123, China*
[4] *Center for Systems Biology, Soochow University, No. 1 Shizi Street, Suzhou 215006, China*
[5] *School of Chemistry, Biology and Material Engineering, Suzhou University of Science and Technology, Suzhou 215011, China*

Correspondence should be addressed to Jiajia Chen; njucjj@126.com and Weichang Chen; weichangchen@126.com

Received 17 January 2014; Accepted 29 March 2014; Published 28 May 2014

Academic Editor: Degui Zhi

Gastric cancers (GC) have the high morbidity and mortality rates worldwide and there is a need to identify sufficiently sensitive biomarkers for GC. MicroRNAs (miRNAs) could be promising potential biomarkers for GC diagnosis. We employed a systematic and integrative bioinformatics framework to identify GC-related microRNAs from the public microRNA and mRNA expression dataset generated by RNA-seq technology. The performance of the 17 candidate miRNAs was evaluated by hierarchal clustering, ROC analysis, and literature mining. Fourteen have been found to be associated with GC and three microRNAs (miR-211, let-7b, and miR-708) were for the first time reported to associate with GC and may be used for diagnostic biomarkers for GC.

1. Introduction

Gastric cancer (GC) or stomach cancer (SC), the fourth leading cancer worldwide, is a biologically heterogeneous disease. It is the second major contributor to mortality caused by cancer [1, 2]. GC is most common in the Asian and Pacific Islanders and the incidence and death rate are more than twice those in Whites [3]. The occurrence and development of GC is multiple step and multiple factorial processes. The risk factors for gastric cancer include *Helicobacter pylori* infection, advanced age, diet low in fruits and vegetables or high in salted, smoked, preserved foods, chronic atrophic gastritis, and family history of gastric cancer [4–7].

MicroRNAs are small, single-stranded, and noncoding RNAs that negatively regulate gene expression at the post-transcriptional level [8]. Multiple studies have shown differential expression of microRNAs between cancer and normal tissues. Aberrant changes in microRNAs expression have been shown to be associated with lung cancers [9], breast cancers [10], prostate cancers [11], and others. MicroRNAs are therefore the promising candidates as diagnostic, prognostic, and predictive biomarkers in cancers. Various studies have investigated important role of the microRNAs in gastric cancers [12–17].

However, gastric cancers are systems biology diseases and the heterogeneity and complexity of carcinogenesis complicate the marker identification process. Herein we employed an integrated systems biology approach to identifying candidate miRNAs as biomarkers that could differentiate patient with gastric cancer from healthy controls. The analysis pipeline of this paper is shown in Figure 1.

2. Methods

2.1. Dataset Collection and Outlier Differential Expressed Genes Detection. We explored expression profiles (GSE36968 from NCBI GEO) for gastric cancer (GC) and noncancerous gastric tissue samples [18]. The dataset was generated by the AB SOLiD System 3.0 (Homo sapiens). The dataset includes 30 transcriptomic profiles, 6 from noncancerous gastric tissues and 24 from gastric tumor samples. Among the 30

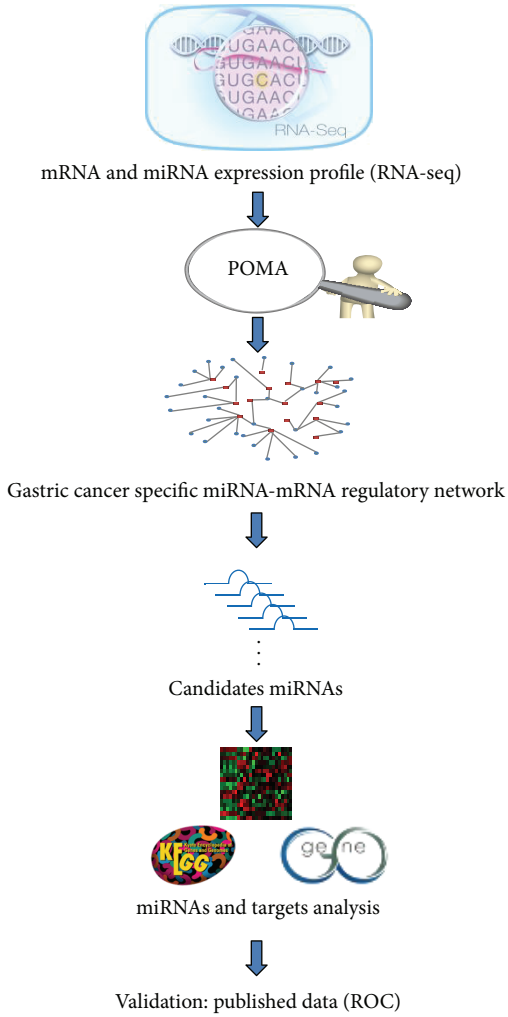


Figure 1: Analysis pipeline in this study.

samples, 25 of them have paired miRNA and mRNA expression profiles. These 25 samples, which contain 6 noncancerous gastric tissue samples and 19 gastric tumor samples, were selected for further analysis. The clinical information of the samples was summarized in Table 1 and the detailed information was listed in Additional file 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/901428. We directly downloaded the processed expression data and used log transformation of the expression values for the following analysis.

Outlier microRNAs and genes were detected with least sum of ordered subset square t-statistic (LSOSS) [19] and implemented in R scripts by Karrila et al. [20]. We used the spearman correlation method to detect negative correlations between outlier miRNAs and outlier genes. The cutoff for correlation coefficient was chosen to be -0.6 and P value < 0.05. Thus we got the significant inverse expression pattern specific for the gastric cancer.

2.2. Refinement of Candidate Gastric Cancer MicroRNAs with the Pipeline of Outlier MicroRNA Analysis (POMA). We employed an in-house prediction model POMA to identify the candidate GC miRNAs from the outlier miRNAs detected by LSOSS. POMA is an integrative method to identify candidate cancer miRNA biomarkers from the miRNA regulatory network by linking paired miRNA and gene expression data and highly reliable miRNA-mRNA interaction data [21]. The main hypothesis of POMA is that if the deregulated genes are targeted exclusively by certain miRNA, this very miRNA is more likely to show regulatory activity. Based on the in-depth exploration of miRNA regulatory network, we conclude that miRNAs with greater independent regulatory power were more likely to be potential biomarkers in human. POMA has successfully identified miRNAs as potential biomarkers in prostate cancer [21], clear cell renal cell carcinoma [22], and sepsis [23].

Using POMA, we mapped the inverse expression pattern of miRNAs and targets to a human miRNA-mRNA interaction network to construct a GC-specific miRNA-mRNA interaction subnetwork. The human miRNA-mRNA interaction network was reconstructed by a comprehensive search of experimentally validated interactions extracted from 4 databases (miRecords, miRTarbase, miR2Disease, and TarBase) and computational prediction from HOCTAR, starBase, and ExprTargetDB.

Then the Z-score was calculated to measure the probability of miRNA having regulatory role in cancer. Z-score was the ratio of number of genes targeted exclusively by a specific miRNA (α) and number of all the genes targeted by that miRNA (β). The Z-score was calculated for each miRNA in the GC-specific miRNA-mRNA interaction subnetwork. Using thresholds 0.3 of Z-score and significantly larger α ($\alpha > 1$, P value < 0.05), we identified candidate miRNAs with potential regulatory role in GC.

2.3. Evaluation of the Performance of MicroRNAs. We employed the heat map and ROC analysis to evaluate the quality of candidate miRNA as GC biomarkers. Heat map and hierarchical clustering were performed by the R package

Table 1: Clinical information of 25 samples.

Characteristic		Sample ($n = 25$)
Age	Median	66
	Range	32–83
Sex	Male	20
	Female	5
Stage	Stage I	5
	Stage II	5
	Stage III	5
	Stage IV	4
	Normal	6
Histology	Mixed	2
	Diffuse	9
	Intestine	5
	Unknown	3
	Normal	6

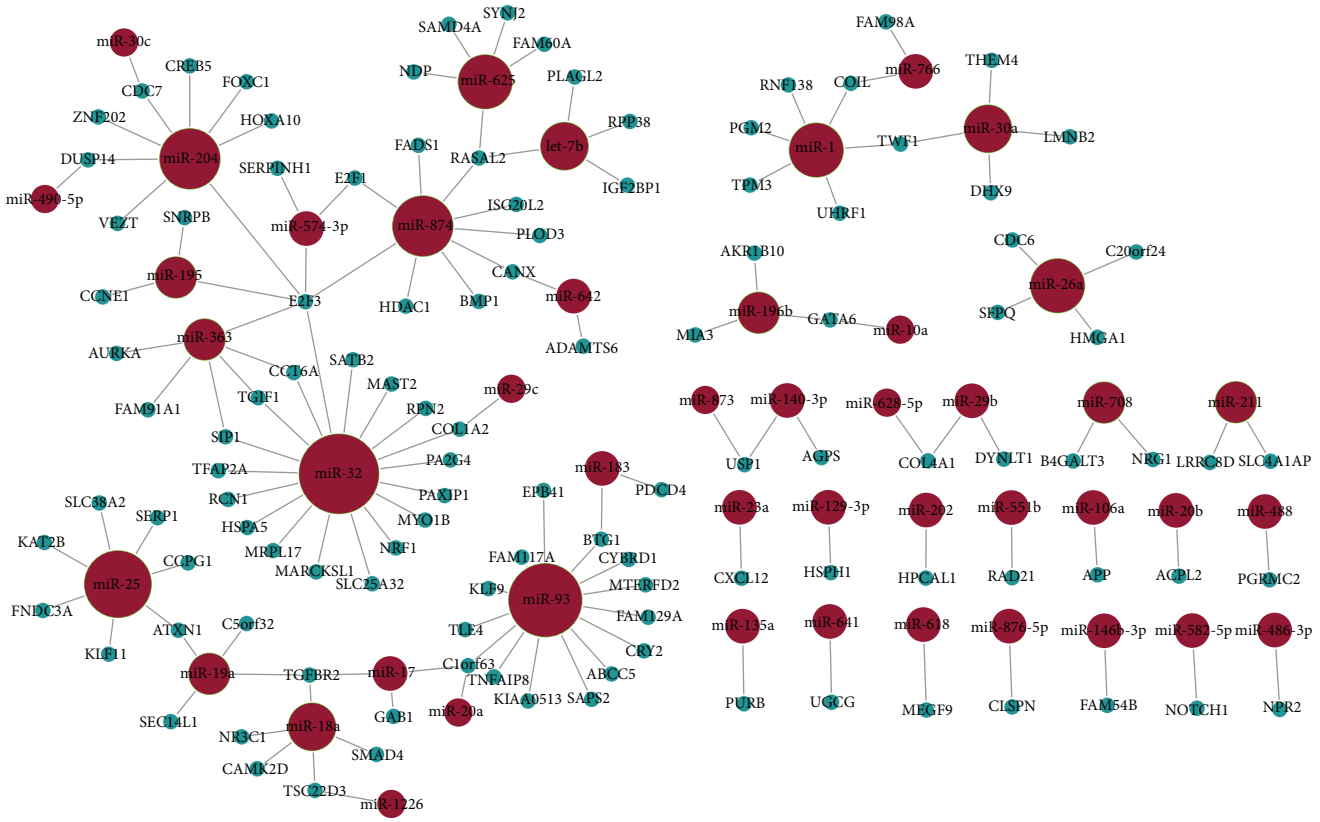


FIGURE 2: Gastric cancer specific miRNA-mRNA subnetwork. Red nodes and blue nodes denote miRNAs and target genes, respectively. miRNAs nodes with green border are candidate miRNAs as biomarkers.

"gplots" [24]. Receiver-operating characteristic (ROC) curves were constructed and the area under ROC curves (AUC) was calculated to evaluate the performance of each candidate miRNA. The parameters sensitivity, specificity, and accuracy were also provided for the miRNAs.

2.4. Functional Enrichment of MicroRNA Targets. The target genes of candidate miRNAs were mapped to different databases, including gene ontology (GO), KEGG, and MetaCore pathway maps and diseases (by Biomarkers) ontology for functional enrichment analysis. GO and KEGG pathway enrichment were performed using Database for Annotation, Visualization and Integrated Discovery (DAVID) [25, 26]. MetaCore pathway maps and diseases (by Biomarkers) ontology enrichment analysis were performed by MetaCore from Thomson Reuters. MetaCore calculates the P value by hypergeometric distribution to evaluate the statistical significance of the enriched pathways and diseases (by Biomarkers) and uses false discovery rate adjustment for multiple test correction.

3. Results

3.1. Detection of Gastric Cancer Specific miRNA-mRNA Subnetwork. Using POMA, we mapped the significant inverse expression pattern from the gastric cancer data to the human miRNA-mRNA interaction network to get the GC specific miRNA-mRNA subnetwork as illustrated in Figure 2 and the edge list of the network was shown in Additional file 2. The subnetwork contains 161 nodes, 46 miRNAs (red nodes), 115 target genes (green nodes), and 142 interactions.

3.2. Identifying Candidate miRNA as Biomarkers for GC. We retrieved a set of seventeen candidate miRNAs (Table 2) through POMA and evaluated their performance as biomarkers in three aspects. First, using hierarchical clustering based on the candidate miRNAs expression, we successfully separated the 19 samples of cancer tissue and 6 normal samples into discrete groups (Figure 3). The ROC curves for the 17 candidate miRNAs are presented in Figure 4. The areas under the ROC curve (AUC) for the 17 miRNAs are 0.833–0.986, and overall accuracy is 73.2%–94.3%. The identified miRNAs are able to effectively differentiate patients with the GC from controls.

Furthermore, we also performed the literature search of the seventeen miRNAs to validate their role in the regulation of GC; see Table 2. Fourteen of them have been reported for their roles in gastric cancer by the low-throughput experiment, such as miR-204 targets Bcl-2 [27] and SIRT1 [28] with downregulation in gastric cancer. Although the remaining four miRNAs were not proved in low-throughput experiment, they have their role in gastric cancer or other cancers. miRNA-30a is ensembled with other six miRNAs

Table 2: Aberrantly expressed miRNAs in gastric cancer detected by low-throughput methods.

miRNA	Expression in GC	Detection technology	Study design	PMID
miR-204	Down	RT-PCR/QRT-PCR RT-PCR QRT-PCR	Cell lines Tissue Tissue	23768087 23152059 21416062
miR-211	—	—	—	—
miR-196b	Up	QRT-PCR	Tissue	21416062 24222951
let-7b	—	—	—	—
miR-18a	Up	QRT-PCR	Tissue	21671476
miR-19a	Up	QRT-PCR	Tissue Cell lines	23621248
miR-25	Up	Northern blotting	Tissue	19153141
miR-874	Down	QRT-PCR	Cell lines	23800944
miR-625	Down	QRT-PCR	Tissue	22677169
miR-30a	—	—	—	—
miR-363	Up	QRT-PCR	Cell lines	23975832
miR-93	Up	QRT-PCR	Tissue	18328430
miR-32	Up	QRT-PCR	Tissue	21874264
miR-26a	Down	QRT-PCR	Tissues Cell lines	24015269
miR-195	Down	RT-PCR	Tissue	21987613
miR-708	—	—	—	—
miR-1	Up	QRT-PCR	Serum	21112772

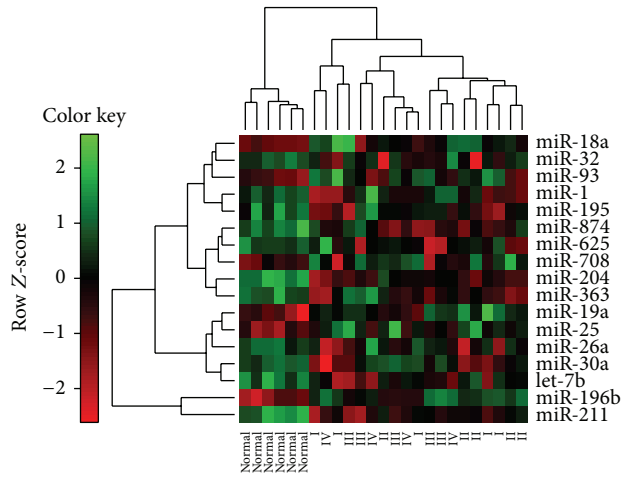


Figure 3: Hierarchical clustering of 19 cancer samples and 6 normal samples with the 17 candidate miRNAs. Every row represents individual miRNA, and each column represents individual sample.

as a seven-miRNA signature which is closely associated with relapse-free and overall survival among patients with gastric cancer [29]. miRNA-211 has contribution to colorectal cancer cell growth [30], melanoma cell invasion [31], and head and neck carcinomas [32]. The expression level of miR-708 reflects differences between colorectal carcinogenesis and normal samples [33] and it may play an important role as a tumor suppressor in human glioblastoma cells [34]. let-7b was upregulated in the acute myeloid leukemia when compared to healthy controls [35]. let-7b in GC patients with low HMGA2 (high mobility group A2) expression was significantly higher than in those with high HMGA2 expression and high expression of HMGA2 in GC correlates was an independent prognostic factor [36]. Therefore, the four miRNAs may be the potential biomarkers for gastric cancer.

3.3. Function Enrichment of Candidate miRNAs Target Genes. The candidate miRNAs, along with their regulated genes, provide potential miRNA-mRNA target pairs in gastric cancer. The targets of these miRNAs were mapped to functional databases, including GO, KEGG, and MetaCore (Figure 5 and Additional file 3). The significantly enriched GO terms (P value < 0.05 and FDR < 0.05) include regulation of RNA metabolic process, regulation of transcription from RNA polymerase II promoter, regulation of transcription, DNA-templated and regulation of transcription. KEGG pathways that are significantly enriched with the candidate miRNAs targets are associated with cancer, for example, cell cycle, pancreatic cancer, pathways in cancer, and prostate cancer.

The enriched (P value < 0.05 and FDR < 0.05) MetaCore pathway maps converge on cell cycle, development, and transcription, as shown in Figure 5 and Table 3. Then we searched the PubMed for published papers describing their constituent network objects in GC to evaluate the relevance of these pathway maps in gastric cancer. All of the enriched pathways have at least ten objects related to gastric cancer; see Additional file 4.

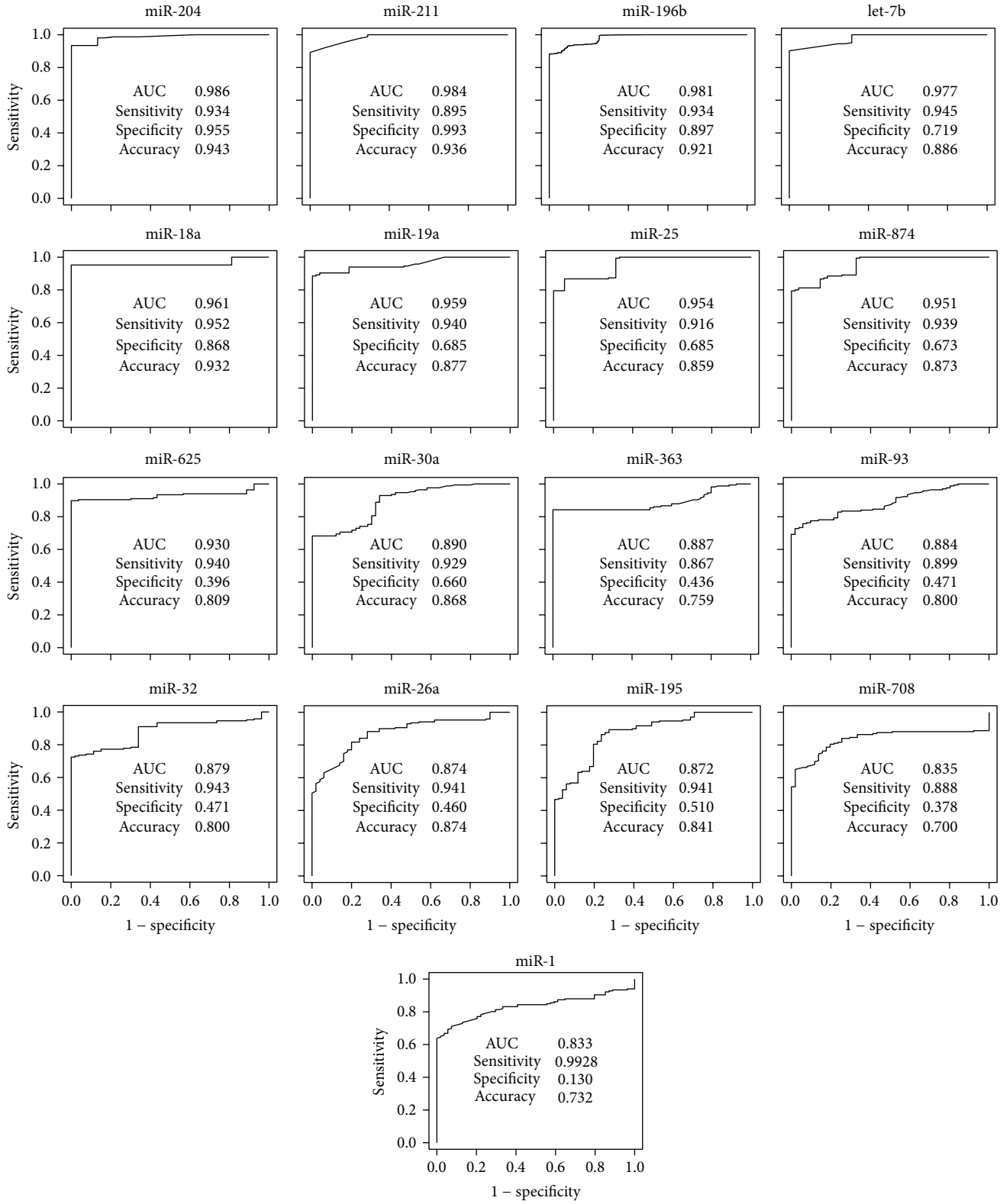


Figure 4: ROC curve of candidate GC miRNAs. AUC: area under the curve.

Disease (biomarkers) ontology in MetaCore is created based on the classification in Medical Subject Headings (MeSH), a controlled vocabulary of medical terms created by the National Library of Medicine (http://www.nlm.nih.gov). Each disease in diseases ontology has its corresponding biomarker gene or a set of genes. The stomach neoplasms disease term ranked top three among the enriched diseases. There are 41 objects in the stomach neoplasms that were

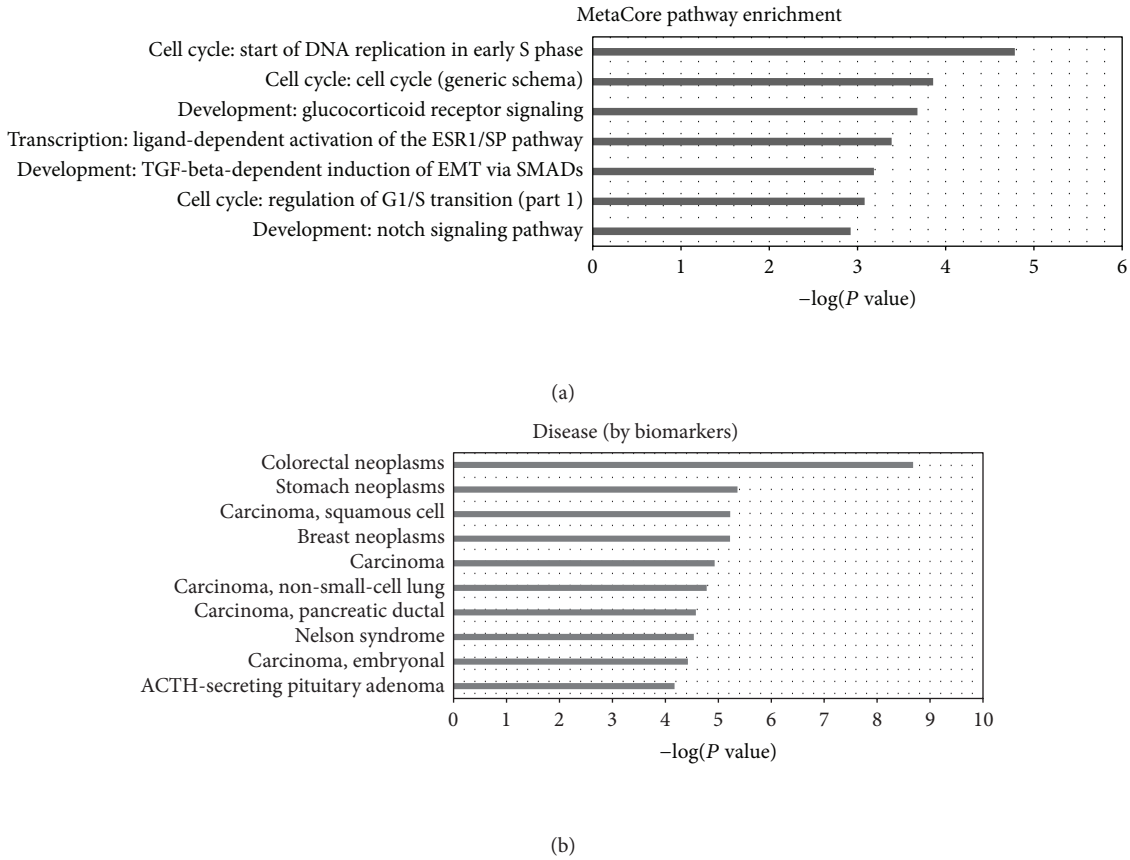


Figure 5: Functional enrichment analysis of target genes. (a) is the significantly enriched MetaCore pathway map. (b) is the significantly enriched disease (biomarkers) ontology.

mapped by the candidate miRNAs target genes. All these results further confirmed the correlation between these target genes and GC and, hence, testified the reliability of our predicted miRNAs as gastric cancer biomarkers.

4. Discussion

In this study, we identified 17 miRNAs using a systematic and integrative method POMA from RNA-seq based expression profile. We first applied LSOSS to detect differentially expressed microRNAs and genes from the RNA-seq data. LSOSS generally outperforms the t-statistics and is more competent for cancer data analysis, as our previous studies indicated [22, 37]. Then the inverse expression pattern of miRNAs and genes was predicted by the spearman correlation.

Using POMA, we got gastric cancer specific miRNA-mRNA subnetwork and 17 candidate GC miRNAs for biomarkers with regulatory roles. The performance of the identified miRNAs was evaluated by hierarchical clustering and ROC curve. Moreover, literature mining confirmed that 14 out of the 17 candidate miRNAs have been reported to have aberrant expression in GC, which lends credibility to our finding. The remaining three miRNAs, miR-211, miR-708, and let-7b, have no previous annotation in GC, but their role in other digestive systems cancers has been reported. miR-211 expression promotes colorectal cancer cell growth [30] and could be a prognostic factor in resected pancreatic ductal adenocarcinoma [38]. miR-708 undergoes aberrant expression in colorectal carcinogenesis samples [33] and pancreatic intraepithelial neoplasias samples [39]. In colorectal liver metastases, invasion front-specific downregulation of let-7b plays a pivotal role in tumor progression [40]. let-7 (let-7b and let-7c) expression has relationship with response to chemotherapy in patients with esophageal cancer and can be potentially used to predict the response to cisplatin-based chemotherapy in esophageal cancer [41]. To our best knowledge, this is the first report that the three microRNAs (miR-211, miR-708, and let-7b) could be the candidate biomarkers for human gastric cancers.

Functional enrichment analysis of the candidate miRNAs target genes revealed some important biological process and pathway maps. Most GO biological process terms are about the regulation processes, for example, the regulation of RNA metabolic process and regulation of transcription. The enriched GO terms in molecular function were also associated with transcription activity such as microRNA regulation activity. The GO enrichment results agree well with the regulatory concepts of microRNAs. MicroRNAs

Table 3: The significant GeneGo pathway maps enriched with candidate miRNAs target genes.

Pathway maps	Pathway map category	Ration of mapped targets	P value	PubMed citation number
Start of DNA replication in early S phase	Cell cycle	4/32	$1.650E-05$	37
Cell cycle (generic schema)	Cell cycle	3/21	$1.387E-04$	75
Glucocorticoid receptor signaling	Development	3/24	$2.089E-04$	76
Ligand-dependent activation of the ESR1/SP pathway	Transcription	3/30	$4.105E-04$	319
TGF-beta-dependent induction of EMT via SMADs	Development	3/35	$6.505E-04$	292
Regulation of G1/S transition (part 1)	Cell cycle	3/38	$8.298E-04$	238
Notch signaling pathway	Development	3/43	$1.193E-03$	29

repress their target genes to fine-tune distinct gene regulatory programs. In cancer, microRNAs play either oncogenic or tumor suppressive role. Oncogenic microRNAs downregulate tumor suppressor genes directly, whereas tumor suppressor microRNAs might lead to the upregulation of oncogenes. In this way, microRNAs regulate cancer progression and dictate specific disease phenotypes. It is also observed that microRNAs are tightly related to other families of regulators, such as transcription factors in gene regulatory networks. They work synergistically to regulate gene expression. So it is not surprising that the targets of GC-related microRNAs converge in gene regulatory processes.

KEGG pathways that are significantly enriched with candidate miRNA targets were all associated with cancers, for example, chronic myeloid leukemia, pancreatic cancer, pathways in cancer, and prostate cancer. It is worth noting that the enriched pathways from both KEGG and MetaCore are involved in cell cycle. For example, in the MetaCore, the top two significantly enriched pathways: the start of DNA replication in early S phase and cell cycle (generic schema) belong to the cell cycle category. The remaining pathways also have important roles in gastric cancer, such as the famous TGF-beta signaling pathway [42–44]. We further evaluated the relevance of the enriched MetaCore pathway maps to gastric cancer by performing the text mining at the objects levels in each pathway and found that all these pathways contain at least ten critical genes in gastric cancers.

According to the disease ontology enrichment analysis, the stomach neoplasm was the second most enriched disease ontology in MetaCore pathways, colorectal neoplasm being the top enriched one. The reason may be that the colorectal neoplasms category incorporates more genes (8014 genes) than stomach neoplasms (3101 genes) in MetaCore database. Thus genes are more likely to be enriched in the colorectal neoplasms. Additionally, colorectal neoplasms and stomach neoplasms share some genes.

The experimental validation is a necessary task to be done after the identification of putative gastric cancer related miRNAs. This is our research plan for the future. Since we did not verify the miRNAs directly in this study, we provided some "indirect evidences" to validate our result by text mining. Although not perfect, text mining helps us to mine previously discovered differential miRNAs and pathways from large volumes of literature, which can help reduce the number of our predicted cancer associated pathways, and to expedite the biological validation of the pathways of interest.

Because the main goal of this research is to identify viable biomarkers of GC diagnosis, we only grouped the samples into 2 major categories: cancer versus noncancerous. Such a binary classification has not fully considered the clinical aspects of each sample. As a future perspective, patients could be subdivided into well-defined small groups according to their unique clinical features, for example, stage, histologic, and therapeutic response. In this manner, the individual difference of cancer mechanism is accounted. This kind of investigation will help to find population-specific biomarkers and facilitate personalized diagnosis, prognosis, or treatment of gastric cancer.

In conclusion, we identified 17 microRNAs that are associated with gastric cancers and 3 of them (miR-211, let-7b, and miR-708) could be potential novel biomarkers for gastric cancer diagnosis and treatment. The candidates predicted herein need further wet-lab validation.

Conflict of Interests

The authors declare that there is no conflict of interests.

Authors' Contribution

Wenying Yan and Shouli Wang contributed equally to this work.

Acknowledgments

This work was supported by the Natural Science Foundation for Colleges and Universities in Jiangsu Province

(13KJB180021) and Natural Science Foundation of USTS (XKQ201315), National Natural Science Foundation of China (Grant nos. 31170795 and 91230117), 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. P. Hamilton and S. J. Meltzer, "A review of the genomics of gastric cancer," *Clinical Gastroenterology and Hepatology*, vol. 4, no. 4, pp. 416–425, 2006.

[2] J. Ferlay, H.-R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin, "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008," *International Journal of Cancer*, vol. 127, no. 12, pp. 2893–2917, 2010.

[3] A. Jemal, R. Siegel, E. Ward et al., "Cancer statistics, 2006," *Ca-A Cancer Journal for Clinicians*, vol. 56, no. 2, pp. 106–130, 2006.

[4] R. C. Kurtz and P. Sherlock, "The diagnosis of gastric cancer," *Seminars in Oncology*, vol. 12, no. 1, pp. 11–18, 1985.

[5] J. M. Scheiman and A. F. Cutler, "Helicobacter pylori and gastric cancer," *The American Journal of Medicine*, vol. 106, no. 2, pp. 222–226, 1999.

[6] C. M. Fenoglio-Preiser, A. E. Noffsinger, J. Belli, and G. N. Stemmermann, "Pathologic and phenotypic features of gastric cancer," *Seminars in Oncology*, vol. 23, no. 3, pp. 292–306, 1996.

[7] S. A. Navarro Silvera, S. T. Mayne, H. A. Risch et al., "Principal component analysis of dietary and lifestyle patterns in relation to risk of subtypes of esophageal and gastric cancer," *Annals of Epidemiology*, vol. 21, no. 7, pp. 543–550, 2011.

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

[9] N. Yanaihara, N. Caplen, E. Bowman et al., "Unique microRNA molecular profiles in lung cancer diagnosis and prognosis," *Cancer Cell*, vol. 9, no. 3, pp. 189–198, 2006.

[10] M. V. Iorio, M. Ferracin, C.-G. Liu et al., "MicroRNA gene expression deregulation in human breast cancer," *Cancer Research*, vol. 65, no. 16, pp. 7065–7070, 2005.

[11] M. Ozen, C. J. Creighton, M. Ozdemir, and M. Ittmann, "Widespread deregulation of microRNA expression in human prostate cancer," *Oncogene*, vol. 27, no. 12, pp. 1788–1793, 2008.

[12] J. L. Wang, Y. Hu, X. Kong et al., "Candidate microRNA biomarkers in human gastric cancer: a systematic review and validation study," *PLoS ONE*, vol. 8, no. 9, Article ID e73683, 2013.

[13] W. K. K. Wu, C. W. Lee, C. H. Cho et al., "MicroRNA dysregulation in gastric cancer: a new player enters the game," *Oncogene*, vol. 29, no. 43, pp. 5761–5771, 2010.

[14] D. Madhavan, K. Cuk, B. Burwinkel, and R. Yang, "Cancer diagnosis and prognosis decoded by blood-based circulating microRNA signatures," *Frontiers in Genetics*, vol. 4, article 116, 2013.

[15] J. Gong, J. Li, Y. Wang et al., "Characterization of microRNA-29 family expression and investigation of their mechanistic roles in gastric cancer," *Carcinogenesis*, vol. 35, no. 2, pp. 497–506, 2014.

[16] H. W. Pan, S. C. Li, and K. W. Tsai, "MicroRNA dysregulation in gastric cancer," *Current Pharmaceutical Design*, vol. 19, no. 7, pp. 1273–1284, 2013.

[17] J. Wang, Q. Wang, H. Liu, B. Hu, W. Zhou, and Y. Cheng, "MicroRNA expression and its implication for the diagnosis and therapeutic strategies of gastric cancer," *Cancer Letters*, vol. 297, no. 2, pp. 137–143, 2010.

[18] Y. H. Kim, H. Liang, X. Liu et al., "AMPKalpha modulation in cancer progression: multilayer integrative analysis of the whole transcriptome in Asian gastric cancer," *Cancer Research*, vol. 72, no. 10, pp. 2512–2521, 2012.

[19] Y. Wang and R. Rekaya, "LSOSS: detection of cancer outlier differential gene expression," *Biomarker Insights*, vol. 2010, no. 5, pp. 69–78, 2010.

[20] S. Karrila, J. H. E. Lee, and G. Tucker-Kellogg, "A comparison of methods for data-driven cancer outlier discovery, and an application scheme to semisupervised predictive biomarker discovery," *Cancer Informatics*, vol. 10, pp. 109–120, 2011.

[21] W. Zhang, J. Zang, X. Jing et al., "Identification of candidate miRNA biomarkers from miRNA regulatory network with application to prostate cancer," *Journal of Translational Medicine*, vol. 12, article 66, 2014.

[22] J. Chen, D. Zhang, W. Zhang et al., "Clear cell renal cell carcinoma associated microRNA expression signatures identified by an integrated bioinformatics analysis," *Journal of Translational Medicine*, vol. 11, article 169, 2013.

[23] J. Huang, Z. Sun, W. Yan et al., "Identification of microRNA as sepsis biomarker based on miRNAs regulatory network analysis," *BioMed Research International*, vol. 2014, Article ID 594350, 12 pages, 2014.

[24] G. R. Warnes, B. Bolker, L. Bonebakker et al., "gplots: various R 12 programming tools for plotting data," The Comprehensive R Archive Network. R Package Version 2.6.0, http://cran.r-project.org/package=gplots.

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

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

[27] A. Sacconi, F. Biagioni, V. Canu et al., "miR-204 targets Bcl-2 expression and enhances responsiveness of gastric cancer," *Cell Death & Disease*, vol. 3, aticle e423, 2012.

[28] L. Zhang, X. Wang, and P. Chen, "MiR-204 down regulates SIRT1 and reverts SIRT1-induced epithelial-mesenchymal transition, anoikis resistance and invasion in gastric cancer cells," *BMC Cancer*, vol. 13, article 290, 2013.

[29] X. Li, Y. Zhang, Y. Zhang, J. Ding, K. Wu, and D. Fan, "Survival prediction of gastric cancer by a seven-microRNA signature," *Gut*, vol. 59, no. 5, pp. 579–585, 2010.

[30] C. Cai, H. Ashktorab, X. Pang et al., "MicroRNA-211 expression promotes colorectal cancer cell growth in vitro and in vivo by targeting tumor suppressor CHD5," *PLoS ONE*, vol. 7, no. 1, Article ID e29750, 2012.

[31] C. Margue, D. Philippidou, S. E. Reinsbach, M. Schmitt, I. Behrmann, and S. Kreis, "New target genes of MITF-induced microRNA-211 contribute to melanoma cell invasion," *PLoS ONE*, vol. 8, no. 9, Article ID e73473, 2013.

[32] T. H. Chu, C. C. Yang, C. J. Liu, M. T. Lui, S. C. Lin, and K. W. Chang, "miR-211 promotes the progression of head and neck carcinomas by targeting TGFbetaRII," *Cancer Letters*, vol. 337, no. 1, pp. 115–124, 2013.

[33] S. Pizzini, A. Bisognin, S. Mandruzzato et al., "Impact of microRNAs on regulatory networks and pathways in human

colorectal carcinogenesis and development of metastasis," *BMC Genomics*, vol. 14, article 589, 2013.

[34] P. Guo, J. Lan, J. Ge, Q. Nie, Q. Mao, and Y. Qiu, "miR-708 acts as a tumor suppressor in human glioblastoma cells," *Oncology Reports*, vol. 30, no. 2, pp. 870–876, 2013.

[35] H. Fayyad-Kazan, N. Bitar, M. Najar et al., "Circulating miR-150 and miR-342 in plasma are novel potential biomarkers for acute myeloid leukemia," *Journal of Translational Medicine*, vol. 11, article 31, 2013.

[36] K. Motoyama, H. Inoue, Y. Nakamura, H. Uetake, K. Sugihara, and M. Mori, "Clinical significance of high mobility group A2 in human gastric cancer and its relationship to let-7 MicroRNA family," *Clinical Cancer Research*, vol. 14, no. 8, pp. 2334–2340, 2008.

[37] Y. Tang, W. Yan, J. Chen, C. Luo, A. Kaipia, and B. Shen, "Identification of novel microRNA regulatory pathways associated with heterogeneous prostate cancer," *BMC Systems Biology*, vol. 7, no. 3, pp. 1–9, 2013.

[38] E. Giovannetti, A. van der Velde, N. Funel et al., "High-throughput microRNA (miRNAs) arrays unravel the prognostic role of MiR-211 in pancreatic cancer," *PLoS ONE*, vol. 7, no. 11, Article ID e49145, 2012.

[39] J. Yu, A. Li, S. M. Hong, R. H. Hruban, and M. Goggins, "MicroRNA alterations of pancreatic intraepithelial neoplasias," *Clinical Cancer Research*, vol. 18, no. 4, pp. 981–992, 2012.

[40] C. Kahlert, F. Klupp, K. Brand et al., "Invasion front-specific expression and prognostic significance of microRNA in colorectal liver metastases," *Cancer Science*, vol. 102, no. 10, pp. 1799–1807, 2011.

[41] K. Sugimura, H. Miyata, K. Tanaka et al., "Let-7 expression is a significant determinant of response to chemotherapy through the regulation of IL-6/STAT3 pathway in esophageal squamous cell carcinoma," *Clinical Cancer Research*, vol. 18, no. 18, pp. 5144–5153, 2012.

[42] O. Shinto, M. Yashiro, T. Toyokawa et al., "Phosphorylated Smad2 in advanced stage gastric carcinoma," *BMC Cancer*, vol. 10, article 652, 2010.

[43] M. R. Kano, Y. Bae, C. Iwata et al., "Improvement of cancer-targeting therapy, using nanocarriers for intractable solid tumors by inhibition of TGF-β signaling," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 9, pp. 3460–3465, 2007.

[44] L. Mishra, V. Katuri, and S. Evans, "The role of PRAJA and ELF in TGF-beta signaling and gastric cancer," *Cancer Biology & Therapy*, vol. 4, no. 7, pp. 694–699, 2005.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 375980, 7 pages
http://dx.doi.org/10.1155/2014/375980

Research Article

Simulated Annealing Based Algorithm for Identifying Mutated Driver Pathways in Cancer

Hai-Tao Li,[1] Yu-Lang Zhang,[2] Chun-Hou Zheng,[1,3] and Hong-Qiang Wang[4]

[1] *College of Information and Communication Technology, Qufu Normal University, Rizhao 276826, China*
[2] *College of Jia Sixie Agriculture, Weifang University of Science and Technology, Shouguang 262700, China*
[3] *College of Electrical Engineering and Automation, Anhui University, Hefei 230000, China*
[4] *Intelligent Computing Lab, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230000, China*

Correspondence should be addressed to Chun-Hou Zheng; zhengch99@126.com

Received 20 March 2014; Accepted 13 May 2014; Published 26 May 2014

Academic Editor: Bairong Shen

With the development of next-generation DNA sequencing technologies, large-scale cancer genomics projects can be implemented to help researchers to identify driver genes, driver mutations, and driver pathways, which promote cancer proliferation in large numbers of cancer patients. Hence, one of the remaining challenges is to distinguish functional mutations vital for cancer development, and filter out the unfunctional and random "passenger mutations." In this study, we introduce a modified method to solve the so-called maximum weight submatrix problem which is used to identify mutated driver pathways in cancer. The problem is based on two combinatorial properties, that is, coverage and exclusivity. Particularly, we enhance an integrative model which combines gene mutation and expression data. The experimental results on simulated data show that, compared with the other methods, our method is more efficient. Finally, we apply the proposed method on two real biological datasets. The results show that our proposed method is also applicable in real practice.

1. Introduction

Cancer is a fatal disease which is extremely complex. Researchers have found that cancer should be arisen by single-nucleotide mutations, larger copy-number aberrations, or structural aberrations [1]. The dreadful feature of cancer cells is infinite proliferation. These abnormal cells can spread to other tissues through blood circulation or lymphatic system [2]. Hence, cancer is very difficult to be treated.

Clinical diagnostics, prognostics, and targeted therapeutics of cancer need across-the-board comprehending molecular mechanisms of cancer cells. One of the remaining challenges is to distinguish functional mutations vital for cancer development, which is so-called "driver mutations," and filter out the unfunctional and random "passenger mutations" [3]. With the development of next-generation DNA sequencing technologies, large-scale cancer genomics projects have been implemented to help researchers to identify driver genes, driver mutations, and driver pathways which promote cancer proliferation in large numbers of cancer patients [4–6]. Hence, it is necessary to find efficient methods for identifying mutated driver pathways in cancer cells, which can be further used to aid in designing effective drugs to treat cancer [7, 8].

In the past years, in gene level, several studies have been devoted to predict driver mutation with significantly higher mutation rate than background mutation rate in a large cohort of cancer patients. These methods have detected several gene mutations in cancer progression. However, even cancer genomes from the same type of cancer, no two genomes exhibit exactly the same complement of somatic aberrations. In other words, these approaches cannot capture the heterogeneity of genome mutations [9, 10].

As it is well known, same pathway may result from different genome aberrations [11, 12]. Hence, it is significant to study gene in pathway level, rather than in gene level. In pathway level, it is easy to capture the heterogeneous phenomenon in cancer cells [13, 14]. Until now, most of the studies analyze known pathway for enrichment of somatic mutations [9, 10,

15]. Though several pathways find out significantly perturbed genes [16–18], unfortunately, knowledge of pathways remains incomplete, and many pathway databases contain overlap and unavailable data. Therefore, taking into account these obvious limitations, it is indispensable to develop de novo discovery of mutated driver pathways without relying on prior knowledge.

In the whole genome, there are a huge number of gene sets if testing exhaustively. For instance, there are more than 10^{26} sets of seven human genes [19]. Therefore, testing all the groups up to a reasonable size seems implausible. However, in recent years, several studies have provided some methods to solve this problem [12, 20]. In these studies, the researchers find that there are two constraints on combinatorial patterns of mutations in cancer. First, generally, a driver mutation is rare. Particularly, researchers find that a single mutation is frequently enough to perturb one way. In other words, there is a phenomenon of mutual exclusivity between driver mutations. Second, a significant cancer pathway should cover a great majority of patients. Thus, the mutations should be contained by most patients in the pathway. This property is called high coverage. Lately, based on these two constraints, Vandin et al. [19] proposed a new and effective method, which defined a novel scoring function using the above two properties to detect the mutated driver pathway using the cancer data detected by next-generation DNA sequencing technologies. They defined the maximization of this method as the maximum weight submatrix problem. However, this problem is computationally difficult to solve.

In order to solve this problem, in this paper, based on GA method introduced by Zhao et al. [21], we propose the simulated annealing hybrid genetic algorithm (SAGA) method for mutated driver pathway detecting. In particular, we incorporate the gene expression data to improve GA to detect mutated driver pathway, and the experimental results on both simulated and real data show that the proposed method is effective.

The rest of this paper is organized as follows. In Section 2, some materials and methods used throughout this paper are introduced. Then, in Section 3, to test the efficiency of our methods, we apply our methods onto simulated data and two biological datasets. The results show that our methods are more efficient. Finally, we draw our conclusions in Section 4.

2. Materials and Methods

2.1. A Brief Introduction. Identifying driver pathway is extremely difficult. Considering this point, some researchers transformed this problem into maximum weight submatrix problem using two criteria [19], that is, "high coverage" and "high exclusivity." However, this problem is NP-hard. In other words, no algorithm efficient in every case awaits a satisfactory result. Hence, many researchers use stochastic search methods to solve this problem. Particularly, Vandin et al. [19] proposed a method using these two properties (Figure 1). The first one is "high coverage," which means the majority of samples have at least one mutation in driver pathway; the second one is "high exclusivity," which means that lots of samples have no more than one mutation in one

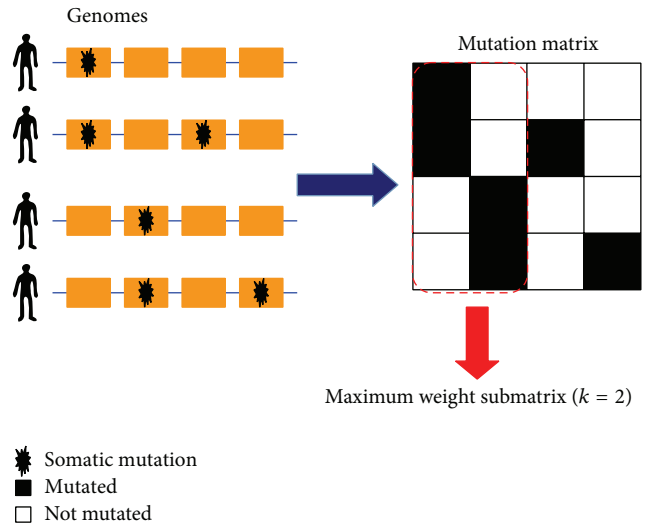


Figure 1: Somatic mutations in samples (patients) are represented in a mutation matrix.

driver pathway. They reflect these two properties using a mutation matrix and a scoring function. A binary mutation matrix A is constructed by m rows (samples) and n columns (genes). The maximum weight submatrix problem is defined as selecting a submatrix M of size $m \times k$ in the mutation matrix A by calculating maximizing the scoring function:

$$W(M) = |\Gamma(M)| - \omega(M) = 2|\Gamma(M)| - \sum_{g \in M} |\Gamma(g)|, \quad (1)$$

where $\Gamma(g) = \{i : A_{ig} = 1\}$ denotes that gene g in ith row (sample) is mutated. $\Gamma(M) = \bigcup_{g \in M} \Gamma(g)$ represents the set of patients, in which at least one of the genes in M is aberrations. So, $|\Gamma(M)|$ indicates the coverage of $M \cdot \omega(M) = \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|$ denotes the coverage overlap weight. In order to solve this problem, Vandin et al. [19] proposed the Markov chain Monte Carlo (MCMC) method. After that, Zhao et al. [21] used the genetic algorithm (GA) to solve this problem and achieved good experimental results. However, to avoid tripping in a local solution, local search method proposed by them is not good enough to solve this problem.

2.2. Simulated Annealing Hybrid Genetic Algorithm (SAGA). As Zhao et al. [21] discussed, the genetic algorithm (GA) is a stochastic and powerful technique that can be effective in solving the maximum weight submatrix problem. However, there is a phenomenon called "premature" that maybe appear in the optimal solutions of GA. In other words, the result may be trapped in a local solution. Taking into account this situation, in this paper, we propose to use simulated annealing hybrid genetic algorithm (SAGA) to solve this problem. Simulated annealing (SA) as an optimization and heuristic algorithm mimics certain thermodynamic principles of producing an ideal crystal, which solve large-scale optimization problems in order to achieve a global optimal solution [22]. SA has been widely used in operational research problems. For example, Chu et al. [23] used SA to analyze the network

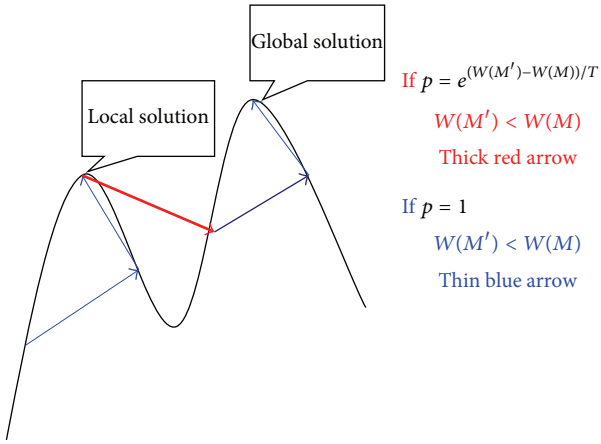


FIGURE 2: Simulated annealing: escape from local maximum solution.

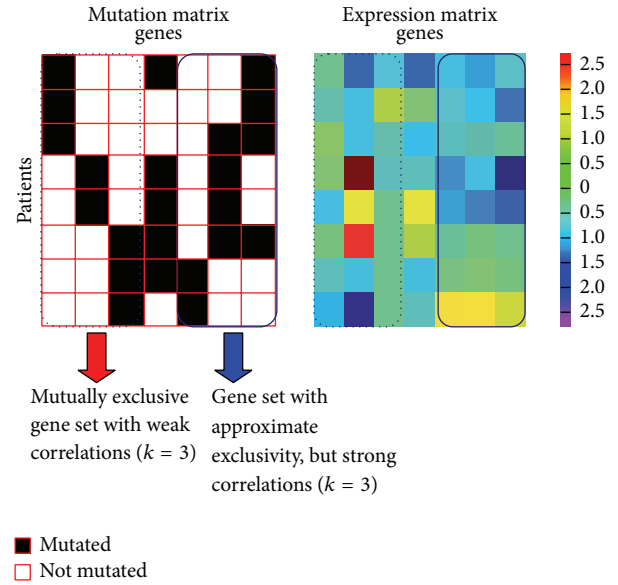


FIGURE 3: Illustration of the advantage of the integrative model. It utilizes the phenomenon that the expression profiles of gene pairs in same pathway have stronger correlations than those in different pathways to detect the driver mutation pathways. In blue dashed box, the genes have very weak expression correlations between each other, while, in the blue real line box, the genes with approximate exclusivity are strongly correlated with each other.

of interacting genes. They detected the genes which control embryonic development and other biological processes. The details of our implementation, named simulated annealing hybrid genetic algorithm (SAGA), for the maximum weight submatrix problem based on SA are described as follows.

Step 1. Initialize the temperature S_0.

Step 2. Use GA method to generate initial solution submatrix M, and generate the scoring function $W(M)$.

Step 3. Using GA method to generate a new solution submatrix M', in the neighborhood of current solution X, reevaluate the scoring function $W(M')$.

Step 4. If the generated solution submatrix scoring $W(M')$ is larger than former $W(M)$, put $M = M'$. Update the existing optimal solution and go to Step 6.

Step 5. Else accepts M' with probability

$$p = e^{\Delta S/T}, \tag{2}$$

where

$$\Delta S = W\left(M'\right) - W\left(M\right). \tag{3}$$

If the solution is accepted, then $M = M'$. Update the existing optimal solution.

Step 6. Decrease the temperature periodically.

Step 7. Repeat Step 2 through 6 until stopping criterion is met.

Figure 2 shows the process of SA. It can be seen clearly that we can solve the global maximum solution by using SA.

2.3. Integrating with Gene Expression Data. In biology, generally, there is noise and/or other factors contained in the data. On the other hand, multiple optimal solutions maybe occur. Taking into account this situation, Zhao et al. [21] proposed a new method called integrative model to deal with this problem. Their new method is based on a phenomenon: the expression profiles of gene pairs in same pathway have stronger correlations than that in different pathways (Figure 3). Hence, they combine the mutation submatrix and the gene expression data, which can distinguish the same score for selecting mutation pathway. They define the integrative model function as follows:

$$F_{\mathrm{ME}} = W\left(M\right) + \lambda * R\left(E_M\right), \tag{4}$$

where $R(E_M) = \sum_{j1 \neq j2}(|\mathrm{pcc}(x_{j1}, x_{j2})|/(k(k-1)/2))$, E_M is the gene expression submatrix which corresponds to the same gene set with the mutation submatrix M, and $\mathrm{pcc}(\cdot)$ is the Pearson correlation coefficient. x_{j1} and x_{j2} are the expression data, which correspond to j_1 and j_2 in E_M. Therefore, $R(E_M)$ is an additional term which enhances the biological correlation. λ is a coefficient. When $\lambda = 1$, F_{ME} will distinguish driver mutation from the same $W(M)$. When $\lambda \geq 1$, F_{ME} will detect the gene set with high correlation and exclusivity. In our study, we set $\lambda = 1$ and $\lambda = 10$. We apply SAGA into the integrative model, and it is more efficient compared with GA method for solving the maximum weight submatrix problem.

3. Results

We first tested the ability of the SAGA to detect the set M of maximum weight submatrix and compared the result with the MCMC and GA methods.

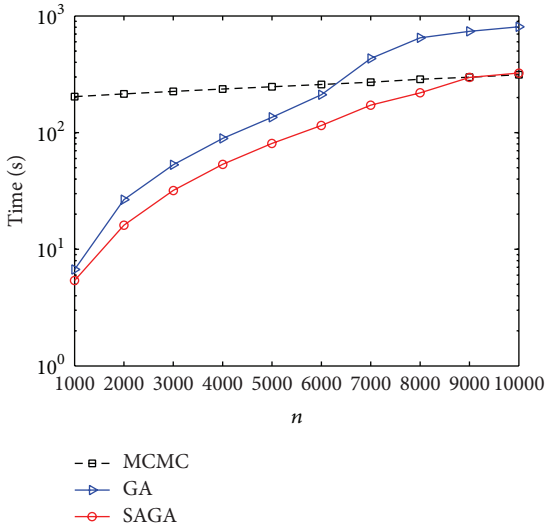


FIGURE 4: Comparison of computational time of SAGA, GA, and MCMC in terms of gene number from 1000 to 10000. In this plot, we use semilog coordinate (the y-axis) to show the computational time in seconds. All the makers correspond to the results of an average over 10 times.

3.1. Simulation Study. We adopt the method represented by Zhao et al. [21]. The details of their implementation of simulated mutation data start with five gene sets $M_1, M_2, \ldots, M_k$. Each set has k members ($k = 5$ has been used in this study). For each row, we set the number to 1 (chosen uniformly at random) in M_i ($i = 1, 2, \ldots, 5$) with probability p_i ($p_i = 1 - i \cdot \Delta$, $\Delta = 0.05$ has been used in this study), and if gene is 1 already, after that, with probability p_0 we set the others to 1 in M_i ($p_0 = 0.04$ has been used in this study). We can see that p_i indicates the coverage of M_i and p_0 indicates exclusivity of M_i. The others in M_i are mutated using a random model based on the observed characteristics of the glioblastoma data. This is the background mutation rate in M.

We have compared the time complexity of MCMC, GA, and SAGA on selecting the submatrix of maximum weight (Figure 4). From this picture, we can see clearly that the GA is faster than MCMC when n is less than about 5000. Particularly, SAGA is always faster than GA from $n = 1000$ to $n = 10000$. In fact, it is well know that, for almost all of real applications, the n is smaller than 5000. On the other hand, the results of SAGA are the same as GA method; that is, they can both detect the five pathways.

Then we use an exact approach to test the accuracy of these methods, which is called binary linear programming (BLP) model proposed by Zhao et al. [21]. We run the BLP method to compare MCMC and GA performance with SAGA. After processing the data, the accuracy of GA and SAGA is equal, which is 95%, but higher than that of MCMC, which is 44%. In summary, our SAGA method has competitive efficiency with GA and MCMC.

3.2. Biological Applications. In this subsection, we applied our SAGA method onto lung adenocarcinoma and glioblastoma datasets. It needs to be emphasized that we consider the mutations in the same samples as one "metagene." We use this criterion when we solve the maximum weight submatrix problem for further analysis. Using the same methods as Zhao et al. [21], we adopt the permutation test to assess the significance of the identified gene patterns. Not only do we get "best" results, but also we check the second optimal patterns, which move the "optimal" submatrix and then detect the "optimal" results in the new matrix.

We first apply our SAGA method onto lung adenocarcinoma. Comparing with GA method, we found that both of them can get the exact same "optimal" submatrix. However, the time using our method is less than that of GA. Afterwards, we apply SAGA-integrative model onto mutation matrix and gene expression matrix of glioblastoma. Compared with the integrating method in Zhao et al., like the former experiment, our method has the same results but using less time.

3.2.1. Lung Adenocarcinoma. We applied our SAGA method to analyzing a dataset of 1013 somatic mutations identified in 188 lung adenocarcinoma patients' 623 sequenced genes from the Tumor Sequencing Project [9]. According to statistics, there are 365 genes mutated in at least one patient. We run the SAGA for sets of size $2 \leq k \leq 10$. After running this algorithm, when $k = 2$, the pair EGFR and KBAS is the maximum weight submatrix. When $k = 3$, the most significant triplet is EGFR, KRAS, and STK11. When $k \geq 4$, all sets are sampled with frequency < 0.3%. Then we perform a permutation test, as described in Vandin et al. [19]. The P value obtained is 0.018, which is larger than that of the triplet (EGFR, KRAS, and STK11). In other words, the triplet (EGFR, KRAS, and STK11) is at least as significant as the pair (EGFR and KRAS). In biology, we find that EGFR, KRAS, and STK11 are all involved in the pathway of mTOR (Figure 5). In Ding et al. [9], the mTOR pathway is very important for lung adenocarcinoma. Hence, our method can seek out driver pathway.

We remove the above three genes and apply the method to detect the additional gene sets. On the remaining genes, when $k = 2$, we identify the gene set (ATM, TP53) that is mutated with frequency 56% and find that the weight of the pair is significant ($P < 0.01$). Previous studies have shown that both ATM and TP53 are in the cell cycle checkpoint control and direct interaction [24, 25] (Figure 6).

3.2.2. Glioblastoma. We next analyzed a collection of somatic single-nucleotide mutations and gene expression profiles identified from 206 glioblastoma multiforme samples from The Cancer Genome Atlas [15]. After processing these data, we established two matrices, that is, a mutation matrix and an expression matrix, which cover 90 samples and 1126 genes.

Firstly, we discover the mutation pattern only depending on the mutation matrix. When $k = 2$, we identify gene pairs (CDKN2A and TP53), and the other is CDK2B and one "metagene" containing TSPAN31 and CDK4. However, using previous methods to solve the original maximum weight submatrix problem, we cannot solve this problem, because there are two same score "optimal" gene sets. Then, we apply integration model onto these "optimal gene sets." Running

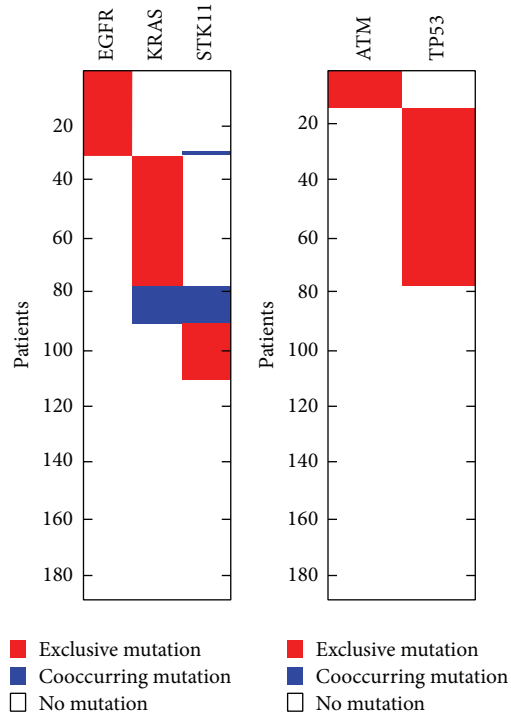


Figure 5: The high weight submatrices of the "optimal" gene set in the lung adenocarcinoma data. In this picture: red, exclusive mutation; blue, cooccurring mutation; white, no mutation. It is similar to Figure 7.

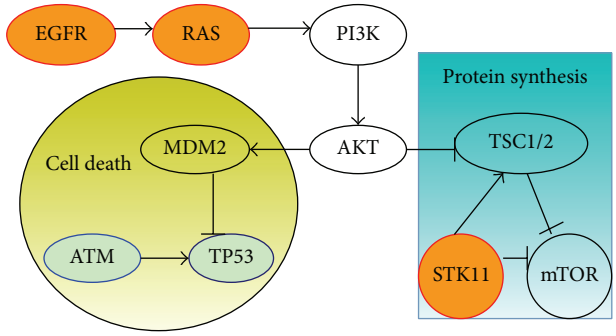


Figure 6: In mTOR signaling pathway, there is the triplet of genes codes for proteins (orange nodes), and the pair (ATM and TP53) corresponds to interacting proteins in the cell cycle pathway (light blue nodes). These two pathways are reported in Ding et al. [9].

the process of integrative method with mutation matrix and gene expression matrix, we find that the correlation between CDK4 and CDKN2B has high score compared to that between TSPAN31 and CDK4. In other words, CDK4 is stronger correlation than TSPAN31 with CDK2B. In biological research, we find that the genes CDK4, CDKN2B are part of RB signaling pathway; however, there is no evidence to discover the relation between TSPAN31 and CDKN2B. In another point of view, it proves the advantages of integrative method. When $k = 3$, the optimal solution is CDK4, CDKN2B, and RB1. After that, we perform a permutation test, as described in Vandin et al. [19]. We find that the triplet

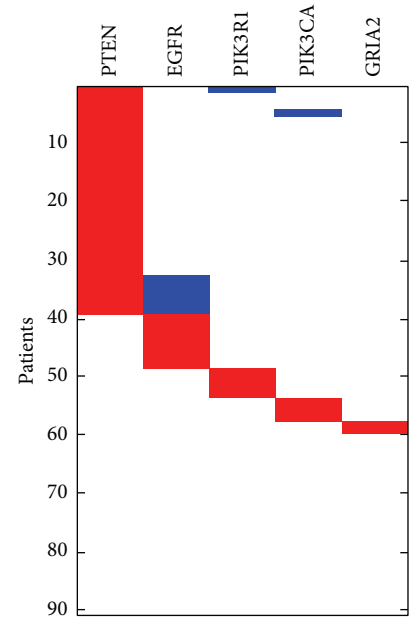


Figure 7: The high weight submatrix of "optimal" gene set after moving the sets (PTEN, EGFR, PIK3R1, PIK3CA, and GRIA2) in glioblastoma data.

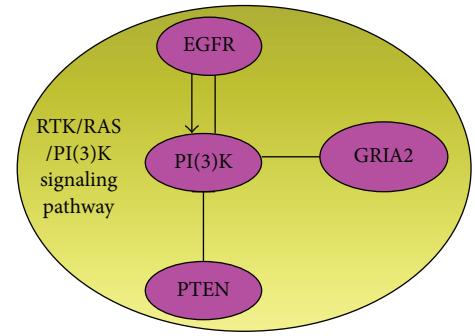


Figure 8: The genes sets (PTEN, EGFR, PIK3R1, and PIK3CA) are involved in the RTK/RAS/PI(3)K signaling pathway, which is reported in TCGA [15].

(RB1, CDKN2B, and CDK4) is more significant than the pair (CDK4 and CDKN2B).

We remove these five genes (RB1, CDKN2A, CDKN2B, CDK4, and TP53) from the mutation matrix and then apply SAGA to discover the others genes. When $k = 5$, the optimal result is PTEN, EGFR, PIK3R1, PIK3CA, and GRIA2, which is significant in the other solutions (Figure 7). The set (PTEN, EGFR, PIK3R1, and PIK3CA) is all part of RTK/RAS/PI(3)K signaling pathway (Figure 8). In biology, gene GIRA2 is very important in glioma cells [26, 27].

4. Discussion and Conclusion

In bioinformatics, it is important to detect mutated driver pathway in cancer cells. In this paper, we introduce an algorithm for discovering mutated driver patterns de novo using somatic mutation data from biological datasets, which

is based on recent exploration made by Vandin et al. [19] and Zhao et al. [21]. We proposed an optimization and heuristic algorithm, that is, simulated annealing hybrid genetic algorithm, which is named SAGA. By means of simulation study, we proved that our SAGA method had completive efficiency with GA and MCMC. Then, we applied our method onto lung adenocarcinoma and glioblastoma. Particularly, we considered incorporating the gene expression data into SAGA method to improve its performance, which achieved satisfactory results. Not only are the results the same as GA, but also the arithmetic speed of SAGA is faster than that of GA.

Although the proposed method can find mutated driver pathway without relying on prior knowledge, we should note that the assumption of high exclusivity and high coverage is too strict for selecting the driver pathway. In biological application, mutual exclusivity is a fairly strong assumption, which holds only for driver mutations in the same pathway. It is well known that driver mutations may be caused by multiple pathways, such as cooccurring and possibly cooperative. For example, acute myeloid leukemia is caused by CBF translocations and kinase mutations [28]. So, we emphasize that assumption of mutual exclusivity occurs only in the same driver pathway. In the future, we will study the biological data, such as DNA methylation and copy-number variant (CNV), exploring the regular pattern of cooccurring and the other mutated driver pathways.

Conflict of Interests

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

Acknowledgments

This work was supported by the National Science Foundation of China under Grant nos. 61272339, 61271098, and 61374181 and the Key Project of Anhui Educational Committee under Grant no. KJ2012A005.

References

[1] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, 2000.

[2] I. J. Fidler, "The pathogenesis of cancer metastasis: the "seed and soil" hypothesis revisited," *Nature Reviews Cancer*, vol. 3, no. 6, pp. 453–458, 2003.

[3] C. Greenman, P. Stephens, R. Smith et al., "Patterns of somatic mutation in human cancer genomes," *Nature*, vol. 446, pp. 153–158, 2007.

[4] E. R. Mardis and R. K. Wilson, "Cancer genome sequencing: a review," *Human Molecular Genetics*, vol. 18, no. 2, pp. R163–R168, 2009.

[5] International cancer genome consortium, "International network of cancer genome projects," *Nature*, vol. 464, pp. 993–998, 2010.

[6] M. Meyerson, S. Gabriel, and G. Getz, "Advances in understanding cancer genomes through second-generation sequencing," *Nature Reviews Genetics*, vol. 11, no. 10, pp. 685–696, 2010.

[7] J. B. Overdevest, D. Theodorescu, and J. K. Lee, "Utilizing the molecular gateway: the path to personalized cancer management," *Clinical Chemistry*, vol. 55, no. 4, pp. 684–697, 2009.

[8] C. Swanton and C. Caldas, "Molecular classification of solid tumours: towards pathway-driven therapeutics," *British Journal of Cancer*, vol. 100, no. 10, pp. 1517–1522, 2009.

[9] L. Ding, G. Getz, D. A. Wheeler et al., "Somatic mutations affect key pathways in lung adenocarcinoma," *Nature*, vol. 455, pp. 1069–1075, 2008.

[10] S. Jones, X. Zhang, D. W. Parsons et al., "Core signaling pathways in human pancreatic cancers revealed by global genomic analyses," *Science*, vol. 321, no. 5897, pp. 1801–1806, 2008.

[11] W. C. Hahn and R. A. Weinberg, "Modelling the molecular circuitry of cancer," *Nature Reviews Cancer*, vol. 2, no. 5, pp. 331–341, 2002.

[12] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control," *Nature Medicine*, vol. 10, no. 8, pp. 789–799, 2004.

[13] S. M. Boca, K. W. Kinzler, V. E. Velculescu, B. Vogelstein, and G. Parmigiani, "Patient-oriented gene set analysis for cancer mutation data," *Genome Biology*, vol. 11, no. 11, article R112, 2010.

[14] S. Efroni, R. Ben-Hamo, M. Edmonson, S. Greenblum, C. F. Schaefer, and K. H. Buetow, "Detecting cancer gene networks characterized by recurrent genomic alterations in a population," *PLoS ONE*, vol. 6, no. 1, Article ID e14437, 2011.

[15] The Cancer Genome Atlas Research Network (TCGA), "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061–1068, 2008.

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

[17] L. J. Jensen, M. Kuhn, M. Stark et al., "STRING 8—a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Research*, vol. 37, no. 1, pp. D412–D416, 2009.

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

[19] F. Vandin, E. Upfal, and B. J. Raphael, "De novo discovery of mutated driver pathways in cancer," *Genome Research*, vol. 22, no. 2, pp. 375–385, 2012.

[20] C.-H. Yeang, F. McCormick, and A. Levine, "Combinatorial patterns of somatic gene mutations in cancer," *FASEB Journal*, vol. 22, no. 8, pp. 2605–2622, 2008.

[21] J. Zhao, S. Zhang, L.-Y. Wu, and X.-S. Zhang, "Efficient methods for identifying mutated driver pathways in cancer," *Bioinformatics*, vol. 28, no. 22, pp. 2940–2947, 2012.

[22] R. Sharda, S. Vob, D. L. Woodruff, and A. Fink, *Optimization Software Class Libraries*, Springer, Berlin, Germany, 2003.

[23] K.-W. Chu, Y. Deng, and J. Reinitz, "Parallel simulated annealing by mixing of states," *Journal of Computational Physics*, vol. 148, no. 2, pp. 646–662, 1999.

[24] K. K. Khanna, K. E. Keating, S. Kozlov et al., "ATM associates with and phosphorylates p53: mapping the region of interaction," *Nature Genetics*, vol. 20, no. 4, pp. 398–400, 1998.

[25] N. H. Chehab, A. Malikzay, M. Appel, and T. D. Halazonetis, "Chk2/hCds1 functions as a DNA damage checkpoint in G1 by stabilizing p53," *Genes and Development*, vol. 14, no. 3, pp. 278–288, 2000.

[26] F. Beretta, S. Bassani, E. Binda et al., "The GluR2 subunit inhibits proliferation by inactivating Src-MAPK signalling and induces

apoptosis by means of caspase 3/6-dependent activation in glioma cells," *European Journal of Neuroscience*, vol. 30, no. 1, pp. 25–34, 2009.

[27] S. Maas, S. Patt, M. Schrey, and A. Rich, "Underediting of glutamate receptor Glur-B mRNA in malignant gliomas," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 25, pp. 14687–14692, 2001.

[28] K. Deguchi and D. G. Gilliland, "Cooperativity between mutations in tyrosine kinases and in hematopoietic transcription factors in AML," *Leukemia*, vol. 16, no. 4, pp. 740–744, 2002.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 601869, 8 pages
http://dx.doi.org/10.1155/2014/601869

Research Article

The Analysis of the Disease Spectrum in China

Xin Zhang, Xiaoping Zhou, Xinyi Huang, Shumei Miao, Hongwei Shan, Shenqi Jing, Tao Shan, Jianjun Guo, Jianqiu Kou, Zhongmin Wang, and Yun Liu

Department of Information, The First Affiliated Hospital, Nanjing Medical University, Nanjing 210029, China

Correspondence should be addressed to Yun Liu; liuyun@njmu.edu.cn

Received 25 March 2014; Accepted 30 April 2014; Published 22 May 2014

Academic Editor: Junfeng Xia

Analysis of the related risks of disease provides a scientific basis for disease prevention and treatment, hospital management, and policy formulation by the changes in disease spectrum of patients in hospital. Retrospective analysis was made to the first diagnosis, age, gender, daily average cost of hospitalized patients, and other factors in the First Affiliated Hospital of Nanjing Medical University during 2006–2013. The top 4 cases were as follows: cardiovascular disease, malignant tumors, lung infections, and noninsulin dependent diabetes mellitus. By the age of disease analysis, we found a younger age trend of cardiovascular disease, and the age of onset of cancer or diabetes was somewhat postponed. The average daily cost of hospitalization and the average daily cost of the main noncommunicable diseases were both on the rise. Noncommunicable diseases occupy an increasingly important position in the constitution of the disease, and they caused an increasing medical burden. People should pay attention to health from the aspects of lifestyle changing. Hospitals should focus on building the appropriate discipline. On the other hand, an integrated government response is required to tackle key risks. Multiple interventions are needed to lower the burden of these diseases and to improve national health.

1. Introduction

Currently, the burden of global disease has changed greatly with the development of the national economy, the improvement of people's living standard, the deterioration of environment, the increasing pressure of work, the transformation of lifestyle, and other changes. The main diseases affecting human health have switched from acute and chronic infectious diseases to chronic noncommunicable diseases [1].

The study of the burden of disease provides an overall guidance for disease prevention and treatment, determines the level of medical technology and community medical demands, and is a reliable basis on which one can distribute the healthcare resources appropriately. The study of changes of disease spectrum has significance to find the disease regional characteristics, guide health policy, and solve the problem of shortage of medical resources. In view of what is mentioned above, we established a clinical database to conduct a further analysis of the changes of disease spectrum by the process of the clinical data of patients of the First Affiliated Hospital of Nanjing Medical University.

2. Materials and Methods

2.1. Used Dataset. Our research objects are these patients who have been discharged from January 2006 to December 2013, and their data we used are from EMR (electronic medical record) and HIS (hospital information system) of the First Affiliated Hospital of Nanjing Medical University.

2.2. Methods. According to the primary diagnosis of hospitalized patients with ICD10 code [2], we established a clinical database to calculate the volume of patients of each disease and the percentage of the total cases of discharge, to rank the diseases and then to compare the percentages of different diseases in different years. There are many types of malignant tumors, so we coanalyzed all the malignant tumors and calculated various malignant tumors cases accounting for all malignant tumor cases proportion. The main types of cardiovascular diseases are coronary heart disease, cerebral infarction, hypertension, and arrhythmia, and we coanalyzed these four types of cardiovascular diseases. We made a count

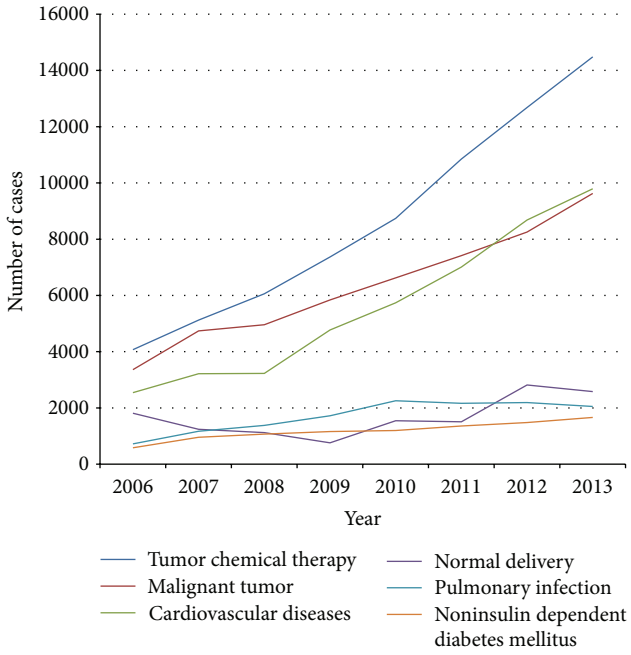


Figure 1: Trend of disease, 2006–2013. The horizontal axis represents year and the vertical axis represents discharges of each disease every year.

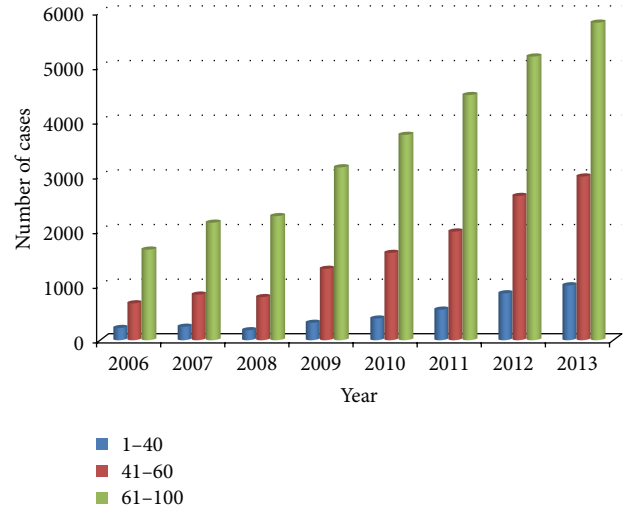


Figure 2: Age distribution of cardiovascular diseases. The horizontal axis represents year and the vertical axis represents discharges of each group of age every year.

on the age and gender of the patients having these noncommunicable diseases. The age variable was categorized into three age groups: <40 years old, 41–60 years old, and >61 years old. We calculated the daily average hospitalization costs discharged each year and the daily average hospitalization costs of cardiovascular diseases and noninsulin dependent diabetes mellitus.

3. Results

3.1. General Review. There are 531718 discharges during 2006–2013. The average number is 66465 every year. Women account for 50.64%. The number of discharged patients increased year by year from 37105 in 2006 to 93040 in 2013.

3.2. The Change of Disease Sequence during 2006–2013. Table 1 provides an overall comparative view of disease sequence between 2006 and 2013. The top ten diseases in our hospital in 2013 were tumor chemical therapy, cardiovascular diseases, malignant tumor, normal delivery, other specified medical care (e.g., Z51.8 code in ICD-10. We can access the URL: "http://apps.who.int/classifications/apps/icd/icd10online2005/fr-icd.htm" or "http://apps.who.int/classifications/icd10/browse/2010/en#/Z51.8". It was coded by WHO.), pulmonary infection, noninsulin dependent diabetes mellitus, cataract, chronic renal failure, and colon polyps. The trend of top six diseases is shown in Figure 1. The horizontal axis represents year and the vertical axis represents discharges of each disease every year.

Tumor chemical therapy is in the first place in disease sequence in all the 8 years and the number of discharges increased from 4072 in 2006 to 14481 in 2013. Since every patient of tumor chemical therapy needs to be treated in hospital several times, it is not the first disease. Malignant tumor is in the second place in disease sequence during 2006–2011 and in the third place during 2012-2013. It is a main disease of patients in our hospital.

Cardiovascular diseases were in the third place in disease sequence during 2006–2011 and have risen to the second place during 2012-2013. It is a main disease of elderly people in China. The most popular diseases are coronary heart disease, cerebral infarction, hypertension, and arrhythmia.

Pulmonary infection and noninsulin dependent diabetes mellitus both constituted the largest number for disease sequence.

3.3. Age Distribution of Cardiovascular Diseases. The age distribution of cardiovascular diseases is shown in Figure 2. The horizontal axis represents year and the vertical axis represents discharges of each group of age every year. We noted that most patients are in the age group above 61 years. It is a common disease of the elderly. Recently, especially after 2009, patients in the age group below 40 years and 41–60 years increased year by year. There is no patient under 30 having coronary heart disease before 2009, while there are more and more patients under 30 having coronary heart disease since 2009. It provides the trend of younger age of cardiovascular diseases, which is coincident with the Global Burden of Diseases, Injuries, and Risk Factors Study 2010 (GBD 2010) by Yang et al. [3].

Coronary heart disease is a major cardiovascular disease, whose age distribution is shown in Figure 3. The horizontal axis represents year, and the vertical axis represents discharges of each group of age every year. We noted that most patients are in the age group above 61 years every year. By removing the extreme differences in individual year, it is

Table 1: Cases and sequence of diseases leaving our hospital each year, 2006–2013.

Diagnosis	Year															
	2013		2012		2011		2010		2009		2008		2007		2006	
	Number of cases (%)	Pos.	Number of cases (%)	Pos.	Number of cases (%)	Pos.	Number of cases (%)	Pos.	Number of cases (%)	Pos.	Number of cases (%)	Pos.	Number of cases (%)	Pos.	Number of cases (%)	Pos.
Total	93040		87273		77621		69301		62205		53903		51270		37105	
Tumor chemical therapy	14481 (15.56)	1	12684 (14.53)	1	10849 (13.98)	1	8739 (12.61)	1	7366 (11.84)	1	6061 (11.24)	1	5122 (9.99)	1	4072 (10.97)	1
Cardiovascular diseases	9790 (10.52)	2	8682 (9.95)	2	7011 (9.03)	3	5739 (8.28)	3	4774 (7.67)	3	3226 (5.98)	3	3217 (6.27)	3	2546 (6.86)	3
Malignant tumor	9629 (10.35)	3	8256 (9.46)	3	7418 (9.56)	2	6624 (9.56)	2	5842 (9.39)	2	4960 (9.20)	2	4742 (9.25)	2	3363 (9.06)	2
Normal delivery	2580 (2.77)	4	2817 (3.23)	4	1510 (1.95)	6	1547 (2.23)	5	758 (1.22)	7	1123 (2.08)	5	1243 (2.42)	4	1813 (4.89)	4
Other specified medical care	2054 (2.21)	5	1855 (2.13)	6	1570 (2.02)	5	1220 (1.76)	6	1034 (1.66)	6	753 (1.40)	7	578 (1.13)	7	506 (1.36)	8
Pulmonary infection	2053 (2.21)	6	2193 (2.51)	5	2164 (2.79)	4	2255 (3.25)	4	1720 (2.77)	4	1380 (2.56)	4	1171 (2.28)	5	725 (1.95)	5
Noninsulin dependent diabetes Mellitus	1661 (1.79)	7	1480 (1.70)	7	1361 (1.75)	7	1198 (1.73)	7	1160 (1.86)	5	1071 (1.99)	6	959 (1.87)	6	582 (1.57)	6
Cataract	1080 (1.16)	8	1153 (1.32)	8	736 (0.95)	8	513 (0.74)	11	482 (0.77)	10	474 (0.86)	9	330 (0.64)	14	271 (0.73)	10
Chronic Renal Failure	908 (0.98)	9	907 (1.04)	9	599 (0.77)	10	584 (0.84)	8	568 (0.91)	8	409 (0.76)	11	355 (0.69)	13	267 (0.72)	11
Colon Polyps	854 (0.92)	10	739 (0.85)	10	343 (0.44)	17	214 (0.31)	35	235 (0.38)	28	118 (0.22)	68	113 (0.22)	77	70 (0.19)	73

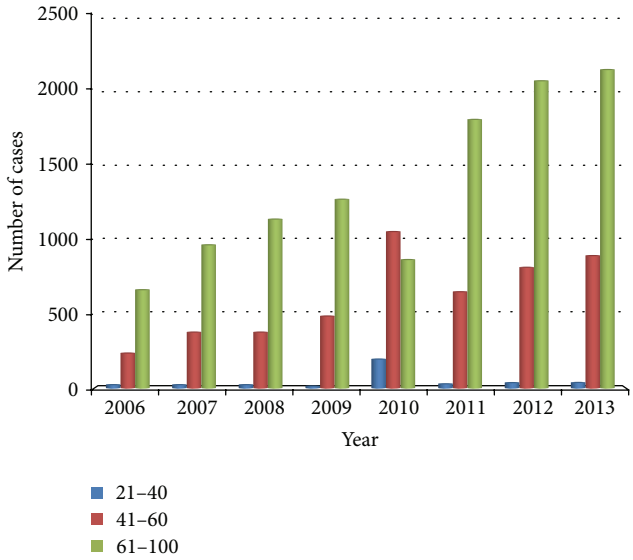


Figure 3: Age distribution of coronary heart diseases. The horizontal axis represents year and the vertical axis represents discharges of each group of age every year.

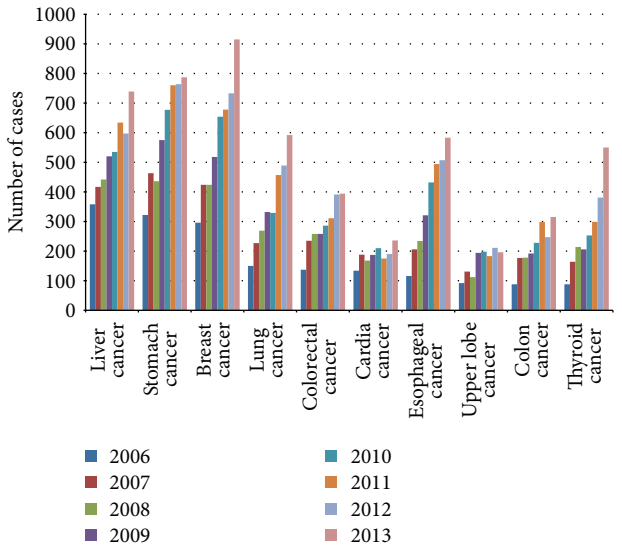


Figure 4: Trend of malignant tumor, 2006–2013. The horizontal axis represents the various cancers and the vertical axis represents the number of cancer patients in different years.

shown that the volume of patients in the age group of 41–60 years has increased year by year, while that of patients in the age group above 60 years has reduced relatively. It provides the trend of younger age of coronary heart diseases and is consistent with the trend of younger age of cardiovascular diseases.

3.4. Age Distribution of Malignant Tumor. Figure 4 provides a status of malignant tumor between 2006 and 2013 of which the horizontal axis expresses the various cancers and the vertical axis expresses the number of cancer patients in

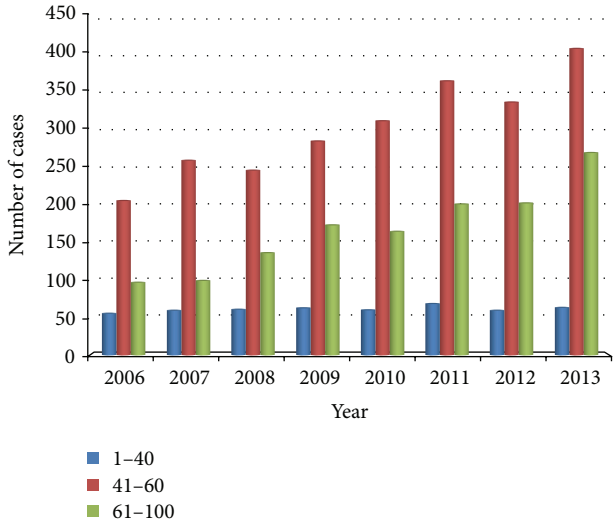


Figure 5: Age distribution of liver cancer. The horizontal axis represents year and the vertical axis represents the number of liver cancer patients in different age groups.

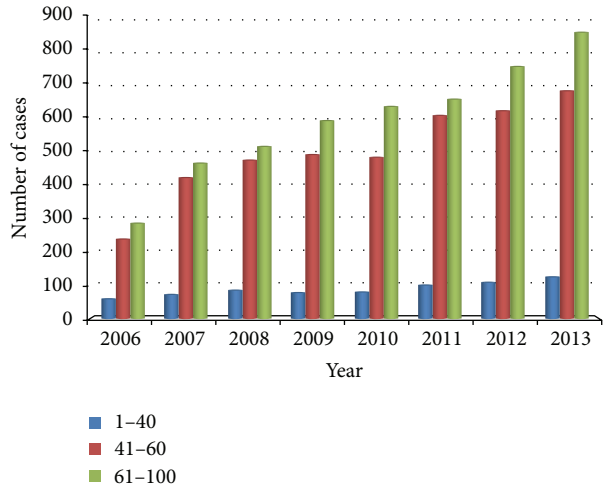


Figure 6: Age distribution of noninsulin dependent diabetes mellitus. The horizontal axis represents year and the vertical axis represents the volume of discharges of each group of age every year.

different years. The top three cancers in sequence are stomach cancer, liver cancer, and breast cancer. An age distribution for liver cancer is shown in Figure 5 of which the horizontal axis expresses years and the vertical axis expresses the number of liver cancer patients in different age groups. We noted that, recently, especially after 2010, the volume of patients in the age group below 40 years and 41–60 years has decreased slowly while that of patients in the age group above 61 years increased dramatically.

3.5. Age Distribution of Noninsulin Dependent Diabetes Mellitus. The age distribution of noninsulin dependent diabetes mellitus is shown in Figure 6 of which the horizontal axis represents year and the vertical axis represents discharges of

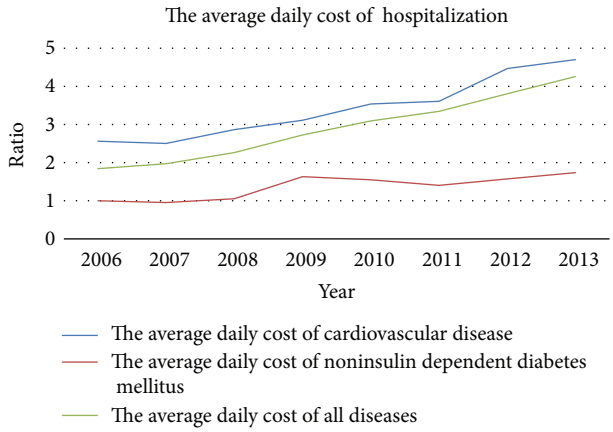


Figure 7: The average daily cost of hospitalization. The horizontal axis represents year and the vertical axis represents the ratio between the average daily cost of cardiovascular diseases or the average daily cost of noninsulin dependent diabetes mellitus each year and the average daily cost of noninsulin dependent diabetes mellitus in 2006.

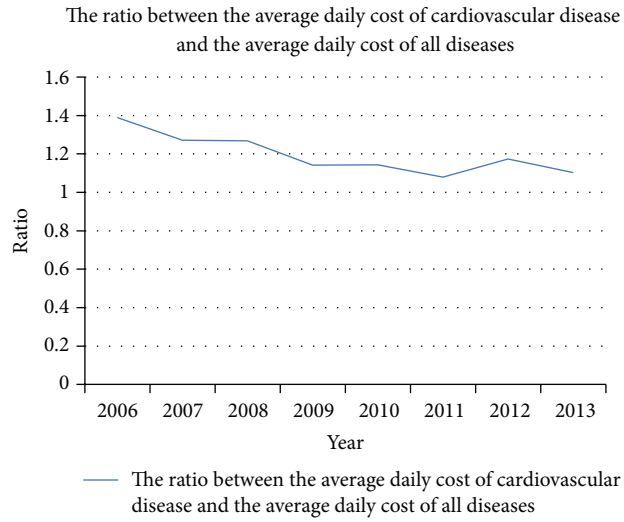


Figure 8: The ratio between the average daily cost of cardiovascular disease and the average daily cost of all diseases. The horizontal axis represents year and the vertical axis represents the ratio between the average daily cost of cardiovascular disease and the average daily cost of all diseases.

each group of age every year. We noted that the volume of patients in the age group below 40 years has reduced from 9.97% in 2006 to 7.53% in 2013 and the volume of patients in the age group of 41–60 years has reduced from 41.07% in 2006 to 40.00% in 2013 in comparison with the fact that the volume of patients in the age group above 61 years has increased from 48.97% in 2006 to 51.48% in 2013.

3.6. The Cost of Major Noncommunicable Diseases Analysis. The average daily cost of hospitalization between 2006 and 2013 is shown in Figure 7. The horizontal axis represents year and the vertical axis represents the ratio between the average daily cost of cardiovascular diseases or the average daily cost of noninsulin dependent diabetes mellitus each year and the average daily cost of noninsulin dependent diabetes mellitus in 2006. The average daily cost has increased gradually year by year. The average daily cost of cardiovascular diseases and the average daily cost of noninsulin dependent diabetes mellitus have increased during recent years, which is consistent with the average daily cost of hospitalization. The ratio of average daily costs of cardiovascular diseases and average daily costs for all diseases in the same year is shown in Figure 8. The horizontal axis represents year and the vertical axis represents the ratio between the average daily cost of cardiovascular disease and the average daily cost of all diseases. We noted that there is a decreasing trend.

3.7. Gender Distribution Analysis. The gender distribution of cardiovascular diseases and noninsulin dependent diabetes mellitus is shown in Tables 2 and 3. We noted that men carry a higher burden of disease than women in both cardiovascular diseases and noninsulin dependent diabetes mellitus. Furthermore, there is an increasing trend of the men's higher burden in recent years.

4. Conclusion

The volume of patients in most of the major hospitals increased year by year with the improvements of people's living standard and healthcare demands. According to our data, the total volume of patients having been discharged from our hospital in 2013 was about 2.5 times as much as that in 2006 (Table 1). On one hand, it is because of extension of our hospital scale and establishment of the Group of Hospital; thus increasing sickbeds would also improve the ability of hospital medical services. On the other hand, the health policy "national basic medical insurance" covers more widely and benefits more people. We are getting closer to the goal: to assure that every citizen has equal access to affordable basic health care [4].

WHO (World Health organization) in Global Status Report on Noncommunicable Diseases 2010 has pointed out that noncommunicable diseases (NCDs) are the leading causes of death globally, killing more people each year than all other causes together [5]. Of the 57 million deaths that occurred globally in 2008, 36 million—almost two-thirds—were due to NCDs, comprising mainly cardiovascular diseases, cancers, diabetes, and chronic lung diseases [6]. The analysis results of disease composition in our hospital show that cardiovascular disease, cancer, diabetes, and lung infection are the most important diseases (Table 1, Figure 1), which are consistent with the above reports.

Cardiovascular diseases (CVDs) are the number one cause of death globally: more people die annually from CVDs than from any other cause [7, 8]. An estimated number of 17.3 million people died from CVDs in 2008, representing 30% of all global deaths. Of these deaths, an estimated number of 7.3 million were due to coronary heart disease and 6.2 million were due to stroke [9]. Our research found that, since 2012, cardiovascular disease has leapt to the first disease in our hospital (except chemotherapy for cancer treatment, Figure 1). In 2013, the volume of hospital discharged patients of cardiovascular disease is 10.52% of the total volume of

Table 2: Gender distribution of cardiovascular diseases, 2006–2013.

Gender		2006	2007	2008	2009	2010	2011	2012	2013
Male	Number of cases	1640	2084	2135	3165	3628	4282	5254	5991
	Percentage (%)	64.4	64.8	66.2	66.3	63.2	61.1	60.5	61.2
Female	Number of cases	906	1133	1091	1609	2111	2729	3428	3799
	Percentage (%)	35.6	35.2	33.8	33.7	36.8	38.9	39.5	38.8
Total		2546	3217	3226	4774	5739	7011	8682	9790

Table 3: Gender distribution of noninsulin dependent diabetes mellitus, 2006–2013.

Gender		2006	2007	2008	2009	2010	2011	2012	2013
Male	Number of cases	337	559	600	655	718	835	918	1053
	Percentage (%)	57.9	58.3	56.0	56.5	59.9	61.4	62.0	63.4
Female	Number of cases	245	400	471	505	480	526	562	608
	Percentage (%)	42.1	41.7	44.0	43.5	40.1	38.6	38.0	36.6
Total		582	959	1071	1160	1198	1361	1480	1661

hospital discharged patients (Table 1). Cardiovascular diseases, including coronary heart disease, have shown a trend of younger age (Figures 2 and 3). Therefore, the improvement of the prevention and treatment for cardiovascular disease cannot be delayed [10, 11].

Cancer is another leading cause of death worldwide, accounting for 8.2 million deaths in 2012 [12] (IARC). Freddie Bray predicts an increase in the incidence of all-cancer cases from 12.7 million new cases in 2008 to 22.2 million by 2030 [13]. Malignant tumors have been in the first place in the sequence of disease in our hospital during 2006 to 2011 and in the second place during 2012 to 2013 (Figure 1). In 2013, the discharged patients of malignant tumors from our hospital are 10.35% of the total discharged patients from our hospital (Table 1). Lung, liver, stomach, colorectal, and breast cancers cause the most cancer deaths each year [14, 15]. In 2013, the volume of breast cancer discharged patients is 9.5% of the total volume of hospital discharged patients, which is in the first place, while the volume of stomach cancer, liver cancer, and lung cancer discharged patients presents 8.17%, 7.67%, and 6.15% of the total volume of hospital discharged patients (Figure 4). Ageing is a fundamental factor for the development of cancer. The incidence of cancer rises dramatically with age, most likely due to a build-up of risks for specific cancers that increase with age. The overall risk accumulation is combined with the tendency for cellular repair mechanisms to be less effective as a person grows older [14]. In the age distribution of liver cancer in our hospital, the age group below 40 years has the least discharged patients; the volume in this group is 8.53% of the total volume of discharged patients of liver cancer. The age group of 41–60 years and the age group above 61 years both have a lot of patients. The volumes in these groups are 55.07% and 36.40% of the volume of total discharged patients of liver cancer, respectively (Figure 5).

Liver cancer is the second cause of death from cancer worldwide. The prognosis for liver cancer is very poor (overall ratio of mortality to incidence of 0.95) [12]. Our data shows that the survival time of liver cancer is 0.69 year. The number may be lower than the fact because of the lack of follow-up data. Most of liver cancer patients find their cancer too late, because they have no early symptom. So the high risk group should check regularly.

The findings of Ravi Prakash Upadhyay pointed out a high burden of diabetes and prediabetes [16]. The prevalence of diabetes in the United States has nearly tripled in the past couple of decades. From 1990 to 2010, the prevalence of diabetes in those aged >18 years increased from 6.6 million in 1990 to 20.7 million in 2010 [17]. The volume of discharges of noninsulin dependent diabetes mellitus increased from 582 in 2006 to 1661 in 2013; it is 2.85 times. Programs should be implemented to educate the community regarding the disease, its signs/symptoms, importance of early detection, and treatment along with ensuring availability of trained staff and well equipped health facilities.

The epidemic of these noncommunicable diseases is being driven by powerful forces now touching every region of the world: demographic ageing, rapid unplanned urbanization, and the globalization of unhealthy lifestyles. Gonghuan Yang pointed that dietary risk factors, high blood pressure, and tobacco exposure are the risk factors that constituted the largest number of attributable DALYs (disability-adjusted life years) in China [3]. Lim et al. pointed that, in 2010, the three leading risk factors for global disease burden were high blood pressure, tobacco smoking including second-hand smoke, and household air pollution from solid fuels. Dietary risk factors (diets low in fruits and those high in sodium) and physical inactivity are also important [18]. Nowadays, a menu of options are set out for addressing these diseases through both population-wide interventions, largely aimed at prevention, and individual interventions, aimed at early detection and treatment that can reduce progression to severe and costly illness and complications [5]. In our research findings, the ratio that the noninsulin dependent diabetes mellitus patients aged <40 accounting for the diabetes patients reduces from 9.97% in 2006 to 7.53% in 2013 (Figure 6) and reveals that the age of onset of noninsulin dependent diabetes mellitus is postponed due

to these effective intervention solutions mentioned above possibly. The rise of inpatient expenditures is caused by the rise of hospital costs, the rise of labor costs, the emerging of new medic methods with the development of science and technology, and the rising demand for health care jointly. The literature shows that health care spending has grown faster than that of income during 1993 to 2003 [4]. According to our analysis of data, the mean inpatient expenditures rise increasingly year to year, that the average cost of 2013 was 2.3 times as much as that of 2006 (Figure 2).

WHO pointed that the costs of health-care systems from noncommunicable diseases are high and increasing [5]. In recent years, the average daily hospitalization costs of cardiovascular disease and noninsulin dependent diabetes mellitus in our hospital are both rising. The average daily hospitalization cost of cardiovascular disease in 2013 was 1.83 times as much as that in 2006. The average daily hospitalization cost of noninsulin dependent diabetes mellitus in 2013 was 1.74 times as much as that in 2006 (Table 2). The average daily hospitalization cost of cardiovascular disease is much higher than the average daily hospitalization cost of all the diseases every year. The average daily hospitalization cost of cardiovascular disease was 1.39 times as much as that of the average daily hospitalization cost of all the diseases in 2006. It showed that the costs from noncommunicable diseases are high. Comparing the ratio between the average daily hospitalization cost of cardiovascular disease and the average daily hospitalization cost, we found that the ratio has decreased. It may show that people paid more attention to the cardiovascular disease in recent years. Vigorous propaganda on cardiovascular disease, preventing it through lifestyle changing, and all the other various interventions could improve the disease burden of cardiovascular disease [5].

These noncommunicable diseases are the focus of hospital medical operation management. Hospitals should construct the appropriate discipline in accordance with the rules and characteristics of the disease, adjust settings or optimize related specialties, and strengthen discipline construction.

On the other hand, people often have the point of view on noncommunicable diseases as problems solely resulting from harmful individual behaviours and lifestyle choices and blame victims. The influence of socioeconomic circumstances on risk and vulnerability to noncommunicable diseases and the impact of health-damaging policies are not always fully understood. Reduction of population exposures from poor diet, high blood pressure, tobacco use, cholesterol, and fasting blood glucose are public policy priorities for China, as are the control of ambient and household air pollution [5]. These changes will require an integrated government response to improve primary care and undertake required multisectoral action to tackle key risks.

Conflict of Interests

The authors have no conflict of interests regarding the publication of this paper.

Authors' Contribution

Xin Zhang and Xiaoping Zhou contributed equally to this study.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (Grant no. 81270952), the Jiangsu Province's Key Provincial Talents Program (BE 2011802), the Projects in the Jiangsu Science and Technology Pillar Program (BE 2011802), the Project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions, the Program for Development of Innovative Research Team in the First Affiliated Hospital of NJMU (no. 20113012), the Special Scientific Research Project from the Ministry of Health, China (Grant no. 201002002), and Nanjing Medical University Science and Technology Development Foundation (2012NJMU122).

References

[1] Y. Liu, G. Yang, Y. Zeng, R. Horton, and L. Chen, "Policy dialogue on China's changing burden of disease," *The Lancet*, vol. 381, pp. 1961–1962, 2013.

[2] World Health Organization, *ICD-10*, vol. 2, World Health Organization, Geneva, Switzerland, 2nd edition, 2005.

[3] G. Yang, Y. Wang, Y. Zeng et al., "Rapid health transition in China, 1990–2010: findings from the Global Burden of Disease Study 2010," *The Lancet*, vol. 381, no. 9882, pp. 1987–2015, 2013.

[4] W. Yip and W. Hsiao, "China's health care reform: a tentative assessment," *China Economic Review*, vol. 20, no. 4, pp. 613–619, 2009.

[5] WHO, *Global Status Report on Noncommunicable Diseases*, WHO Press, 2010.

[6] WHO, *Global Strategy For the Prevention and Control of Noncommunicable Diseases*, WHO, Geneva, Switzerland, 2008.

[7] WHO, "Cardiovascular diseases (CVDs)," Fact Sheet 317, http://www.who.int/mediacentre/factsheets/fs317/en/.

[8] R. Lozano, M. Naghavi, K. Foreman et al., "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010," *The Lancet*, vol. 380, pp. 2095–2128, 2012.

[9] WHO, *Global Atlas on Cardiovascular Disease Prevention and Control*, World Health Organization, Geneva, Switzerland, 2011.

[10] A. Moran, M. Forouzanfar, U. Sampson et al., "The epidemiology of cardiovascular diseases in Sub-Saharan Africa: the global burden of diseases, injuries and risk factors 2010 study," *Progress in Cardiovascular Diseases*, vol. 56, pp. 234–239, 2013.

[11] V. L. Feigin, M. H. Forouzanfar, R. Krishnamurthi et al., "Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010," *The Lancet*, vol. 383, pp. 245–255, 2014.

[12] IARC, "Globocan 2012," http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx.

[13] F. Bray, A. Jemal, N. Grey, J. Ferlay, and D. Forman, "Global cancer transitions according to the Human Development Index (2008–2030): a population-based study," *The Lancet Oncology*, vol. 13, pp. 790–801, 2012.

[14] WHO, "Cancer," Fact Sheet 297, http://www.who.int/media-centre/factsheets/fs297/en/.

[15] J. Traebert, I. Jayce Ceola Schneider, C. Flemming Colussi, and J. Telino de Lacerda, "Burden of disease due to cancer in a Southern Brazilian state," *Cancer Epidemiology*, vol. 37, pp. 788–792, 2013.

[16] R. Prakash Upadhyay, P. Misrab, and V. G. Chellaiyan, "Burden of diabetes mellitus and prediabetes in tribal population of India: a systematic review," *Diabetes Research and Clinical Practice*, vol. 102, pp. 1–7, 2013.

[17] J. M. Lopez, R. A. Bailey, M. F. T. Rupnow, and K. Annunziata, "Characterization of type 2 diabetes mellitus burden by age and ethnic groups based on a nationwide survey," *Clinical Therapeutics*, vol. 36, no. 4, pp. 494–506, 2014.

[18] S. S. Lim, T. Vos, A. D. Flaxman et al., "A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010," *The Lancet*, vol. 380, pp. 2224–2260, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 594350, 12 pages
http://dx.doi.org/10.1155/2014/594350

Research Article

Identification of MicroRNA as Sepsis Biomarker Based on miRNAs Regulatory Network Analysis

Jie Huang,[1] Zhandong Sun,[1,2] Wenying Yan,[2,3,4] Yujie Zhu,[2] Yuxin Lin,[2] Jiajai Chen,[2,4] Bairong Shen,[2] and Jian Wang[1]

[1] *Systems Sepsis Biology Team, Soochow University Affiliated Children's Hospital, Suzhou 215003, China*
[2] *Center for Systems Biology, Soochow University, Suzhou 215006, China*
[3] *Suzhou Zhengxing Translational Biomedical Informatics Ltd., Taicang 215400, China*
[4] *Taicang Center for Translational Bioinformatics, Taicang 215400, China*

Correspondence should be addressed to Jian Wang; wangjian_sdfey@sina.com

Received 17 January 2014; Accepted 3 March 2014; Published 6 April 2014

Academic Editor: Junfeng Xia

Sepsis is regarded as arising from an unusual systemic response to infection but the physiopathology of sepsis remains elusive. At present, sepsis is still a fatal condition with delayed diagnosis and a poor outcome. Many biomarkers have been reported in clinical application for patients with sepsis, and claimed to improve the diagnosis and treatment. Because of the difficulty in the interpreting of clinical features of sepsis, some biomarkers do not show high sensitivity and specificity. MicroRNAs (miRNAs) are small noncoding RNAs which pair the sites in mRNAs to regulate gene expression in eukaryotes. They play a key role in inflammatory response, and have been validated to be potential sepsis biomarker recently. In the present work, we apply a miRNA regulatory network based method to identify novel microRNA biomarkers associated with the early diagnosis of sepsis. By analyzing the miRNA expression profiles and the miRNA regulatory network, we obtained novel miRNAs associated with sepsis. Pathways analysis, disease ontology analysis, and protein-protein interaction network (PIN) analysis, as well as ROC curve, were exploited to testify the reliability of the predicted miRNAs. We finally identified 8 novel miRNAs which have the potential to be sepsis biomarkers.

1. Introduction

Sepsis is among the common causes of death in the intensive care units' patients [1]. A well-defined reason for sepsis is the clinical syndrome resulting from the presence of both systemic inflammatory response and bacterial infection [2]. Sepsis may represent a pattern of response by the immune system to injury, with changes in the activity of thousands of endogenous mediators of inflammation, coagulation, complement, and metabolism [3]. The death toll caused by severe sepsis is of the same range as those from acute myocardial infarction [4]. The need for a timely diagnosis and accurate stratification of the severity of sepsis is no less essential, reducing mortality from sepsis [5].

Over the past decade, sepsis has been considered as a hidden public health disaster [6]. A large number of biomarkers have been proposed as candidates for sepsis diagnosis, prognosis, and therapeutic guidance. The biomarkers aim at recognizing sepsis early, so that supportive measures may be implemented as soon as possible [7, 8]. The most commonly used biomarkers of sepsis in routine clinical diagnostics are procalcitonin (PCT) and C-reactive protein(CRP) [9]. However, it is difficult to diagnose sepsis with high sensitivity and specificity at present due to the limitations of these biomarkers. MicroRNAs (miRNAs) are small noncoding RNAs that pair to sites in mRNAs to regulate gene expression in eukaryotes and play important roles in a variety of cellular functions as well as in several diseases [10–13]. Like other protein-based regulators, miRNAs have been reported as related factors to disease [14, 15]. The abnormal expression of miRNAs leads to malignant phenotypes and implicates changes in a wide array of cellular and developmental processes of disease initiation, progression, and transcriptional regulation network, such as cell proliferation, cell differentiation, apoptosis, invasion, and

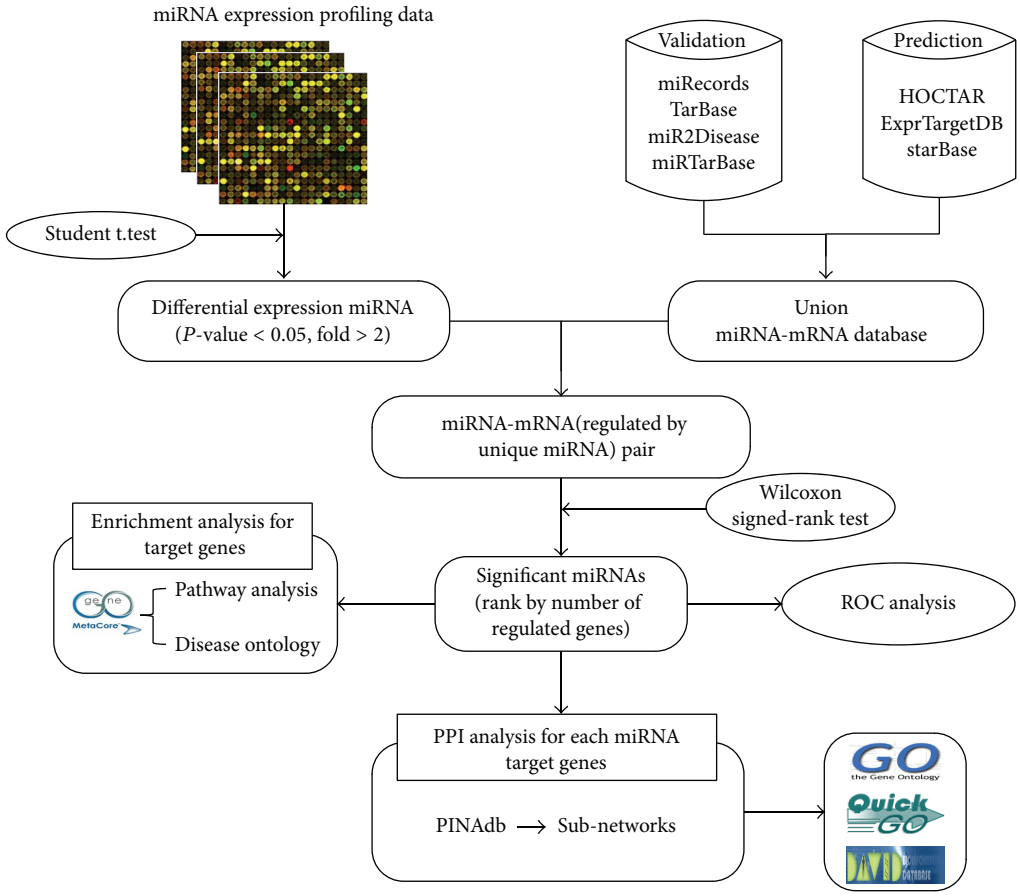


Figure 1: The schematic workflow in our study for identifying miRNAs as potential sepsis biomarkers.

metastasis [10, 16, 17]. MicroRNAs are isolatable from a set of sepsis patient peripheral blood, measured by performing genome-wide profiling by microarray in leukocytes, and have been proposed to be potential sepsis biomarkers [18]. Receiver operating characteristic curves showed that miR-15a has an area under the curve of 0.858 in distinguishing sepsis patients from normal controls [19]. Serum miR-16, miR-193b*, and miR-483-5p are associated with death from sepsis and are identified as prognostic predictors of sepsis patients [20].

Until now, there are many works reported to identify putative microRNA biomarkers [21–27]. Most of them detected the putative microRNA biomarkers by the analysis of differentially expressed microRNA and then verified these candidates by real-time PCR and bioinformatics analysis; they paid much attention to the multiple-multiple interaction between microRNAs and mRNAs. Few of them analyzed the substructure of microRNA-mRNA network with considering the independent regulation power of specific microRNAs. In this study, we applied an integrative analysis of miRNA regulatory networks and microarray expression profiles to identify microRNAs as sepsis biomarker. The procedure of sepsis-related miRNAs identification and analysis is illustrated in Figure 1. We previously analyzed the microRNA regulatory network [28, 29] and defined a novel out degree (NOD) to indicate the independent regulation power for an individual miRNA in the miRNA-mRNA interaction network, that is, the number of genes targeted exclusively by a specific microRNA. It means that miRNAs with larger NOD values are statistically more likely to be candidate disease biomarkers. We exploited different methods to verify the reliability of our candidate miRNA for sepsis diagnosis, and the final result reveals that these miRNAs have the potential to serve as new biomarkers for sepsis.

2. Materials and Methods

2.1. Data Collection. We conducted exhaustive search in Medline database with the key words "sepsis or severe sepsis or septic shock," "miRNA or microRNA," and "biomarker or marker or indicator." Publication date (before October 31, 2013) and human studies were used as filters. We then extracted from each paper the relevant information of biomarkers, for example, microRNA name, accession number in miRBase [30], biomarker type, detection technology, study design, expression in sepsis patients, and PMID.

2.2. MiRNA Microarray Profiles Analysis. The miRNA expression profiles were retrieved from EBI ArrayExpress (http://www.ebi.ac.uk/arrayexpress/). The accession number

is E-TABM-713 [31], produced by Vasilescu et al. The dataset contains 8 normal samples and 8 sepsis samples. We downloaded the normalized miRNA expression data directly and these profiles consist of the expression information of 556 miRNAs.

2.3. Statistical Methods. To identify miRNAs of interest, we adopted student t-test for the statistical analysis. Considering the fact that sample size is not big, we used a threshold of 0.05 for the P value and selected only those probe sets which showed a fold change ≥2 [26]. The miRNAs with differential expression were further ranked by their NOD values, and then Wilcoxon signed-rank test was applied to assign each miRNA a statistic significance value P value, indicating whether the NOD value of an individual miRNA was significantly greater than the median level of all these candidate miRNAs. We take P value < 0.05 as the threshold to select significant miRNAs. The ability to distinguish sepsis group and control group was characterized by the receiver operating characteristic (ROC) curve. We applied ROC analysis on the selected miRNA array data to evaluate the reliability of a biological maker or a classifier. R package epicalc [32] was used to plot the ROC curve and calculate the area under curve (AUC).

2.4. Union miRNA-mRNA Interactions Database. We created union miRNA-mRNA interactions for human, which combine experimentally validated targeting data and computational prediction data. The experimentally validated data were extracted from miRecords [33], TarBase [34], miR2Disease [35], and miRTarBase [36], while the computational prediction data consisted of miRNA-mRNA target pairs residing in no fewer than 2 datasets from HOCTAR [37], ExprTargetDB [38], and starBase [39]. In total, there were 32739 regulation pairs between 641 miRNAs and 7706 target genes.

2.5. Functional Enrichment Analysis. Herein, we mapped the genes uniquely regulated by candidate miRNAs to GeneGo database for analysis of enriched signaling pathway and disease ontology [40–42]. GeneGo database was from MetaCore. In GeneGo, hypergeometric tests were used to evaluate the statistical significance of the enriched pathways and disease. The gene ontology analysis was performed using DAVID Bioinformatics Resources 6.7 [43] and QuickGO [44].

3. Results and Discussion

3.1. Analysis of Known Sepsis miRNA Biomarker. Text mining in NCBI PubMed was used to identify miRNAs as sepsis biomarker. By setting the specific key words, we collated 10 miRNAs that were already proven to be helpful for diagnosis or prognosis of sepsis. To analyze common characteristics of 10 known biomarkers, the number of genes targeted exclusively by a specific microRNA in union miRNA-mRNA interactions database was conducted and we termed it as a novel out degree (NOD) to indicate the independent regulation power for an individual miRNA [28, 29]. Wilcoxon signed-rank test was applied to measure statistical significance of an individual miRNA targets count. We found that 8 of 10 (80%) known miRNA biomarkers were significantly greater than median level of all miRNAs in database; it means that miRNAs with larger NOD values are more likely to be potential sepsis biomarker. Additionally, our previous analysis of identification of cancer miRNA biomarker also suggested that miRNAs with greater independent regulation power tend more likely to be potential cancer miRNA biomarker [28, 29]. Based on this result, we can identify novel miRNA biomarker in sepsis disease. The distribution of NOD value was compared between known miRNA biomarkers and all miRNAs in database, illustrated in Figure 2. Table 1 gives detailed information of known miRNAs biomarker which was extracted from the literature.

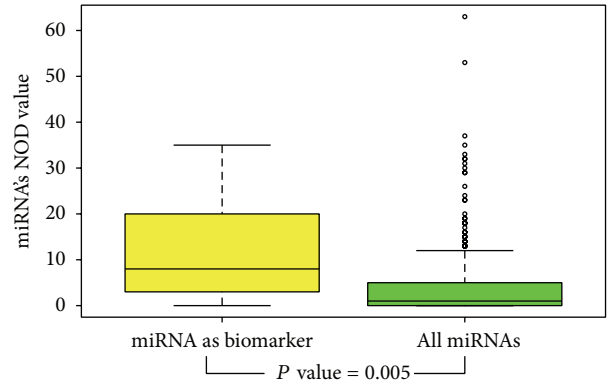


Figure 2: The distribution of NOD value was compared between known miRNA biomarkers and all miRNAs in database. Though we constructed miRNA-mRNA interactions network, the number of genes targeted exclusively by a specific microRNA can be computed. So each miRNA has a NOD value. Kolmogorov-Smirnov test (K-S test) was used to test whether two underlying one-dimensional probability distributions differ. The above boxplot really highlights the difference between two samples. The P value is 0.005 and illustrates that known miRNA biomarkers have more genes uniquely regulated by it.

3.2. Prediction of Candidate Sepsis miRNA Biomarkers. With the result above, we exploited miRNA expression profiles to predict disease biomarker. As described in Methods, we identified 10 significantly and differentially expressed miRNAs to be candidate sepsis miRNA biomarkers from our selected miRNA expression dataset. Among these miRNAs, miR-16 [19] and miR-146a [45] have been previously reported to be sepsis biomarkers. There are some well-known miRNA biomarkers that are not presented in our list; the reason may be the heterogeneity of experimental samples and the stringent threshold we used when selecting differentially expressed miRNAs.

The diagnostic potential of candidate miRNAs was evaluated by ROC curve analysis and the discriminatory accuracy was presented by AUC values. We found that the minimum of AUC is 0.81, the maximum is 0.97, and the average of 5

Table 1: The details of sepsis miRNA biomarkers extracted from the literature.

MicroRNA name (Hsa-)	Accession number (MIMAT)	Biomarker type	Detection technology	Study design	Expression in sepsis patients	PMID	Reference
miR-15a	0000068	Diagnosis	qRT-PCR	Serum	Up	22868808	[19]
miR-16	0000069	Diagnosis	qRT-PCR	Serum	Up	22868808	[19]
miR-122	0000421	Diagnosis	qRT-PCR	Serum	Down	23026916	[18]
miR-146a	0000449	Diagnosis	qRT-PCR	Serum	Down	20188071	[45]
miR-223	0000280	Diagnosis	qRT-PCR	Serum	Down	20188071	[45]
miR-483-5p	0004761	Prognosis	qRT-PCR	Serum	Downregulated in survivors	22719975	[20]
miR-499-5p	0002870	Diagnosis	qRT-PCR	Serum	Down	23026916	[18]
miR-574-5p	0004795	Prognosis	qRT-PCR	Serum	Upregulated in survivors	22344312	[46]
miR-150	0000451	Diagnosis	qRT-PCR	Serum	Down	19823581	[31]
miR-193b*	0004767	Prognosis	qRT-PCR	Serum	Downregulated in survivors	22719975	[20]

Table 2: Candidate miRNAs with outlier activity in sepsis.

MicroRNA name (Hsa-)	Accession number (MIMAT)	P value (sepsis patients versus controls)	Fold change (log 2)	NOD value	P value (NOD statistical significant value)	AUC value (95% CI)
let-7b	0000063	0.020	85.93	53	$2.4E-07$	0.81
miR-16	0000069	0.030	55.79	35	$3.12E-07$	0.84
miR-15b	0000417	0.001	192.07	33	$3.82E-07$	0.95
miR-146a	0000449	0.002	−6.89	20	$1.84E-05$	0.90
miR-210	0000267	0.023	1.64	15	0.0006	0.97
miR-340	0004692	0.021	−1.18	11	0.0021	0.88
miR-145	0000437	0.021	13.03	11	0.0021	0.83
miR-484	0002174	0.002	3.74	11	0.0021	0.92
miR-324-3p	0000762	0.021	2.45	10	0.0041	0.84
miR-486-5p	0002177	0.019	102.49	8	0.0151	0.97

miRNAs' AUC is above 0.90. Because the property of ROC is measured as area under the curve (AUC), the ROC curve comparing sepsis patients and healthy controls provides a graphical demonstration of the superiority of candidate miRNA as sepsis marker. Finally, we plot the false positive rate (1−specificity) versus true positive rate (sensitivity) of a test (see Figure 3) for individual miRNA's ROC analysis. The detailed information on candidate miRNAs is given in Table 2.

3.3. Enrichment Analysis for Target Genes of the Candidate miRNAs. Previous researches have revealed that microRNAs emerged as key gene regulators in diverse biological pathways [47] and aberrant miRNA expression can contribute to human diseases [48]. It means that if a miRNA is abnormally expressed in sepsis patients, the target gene regulated by it should also change in sepsis patients. Accordingly, in order to explore the property of miRNA biomarker, we mapped the uniquely regulated genes of candidate miRNAs to GeneGo database (MetaCore) for pathway and disease ontology analysis [49, 50].

For pathway analysis, we retrieved 29 significantly enriched pathways (P value < 0.05) from GeneGo database. These pathways mapped converge on "immune response," "cell cycle," "apoptosis," and "development," which are well known to play a part in sepsis development. There are 11 pathways related to immune response; it is clear that the endotoxins of reducing sepsis interact with host cells via specific receptors on the cell surface and trigger a dysregulated immune response [51]. We also found 2 pathways for apoptosis, an important factor impacting programmed cell death and a major contributor to the pathophysiology of sepsis [52]. Among development pathways, 3 pathways about angiopoietin or cell proliferation, angiopoietin plays divergent roles in mediating inflammation and vascular quiescence [53], and cell proliferation is concomitantly observed in human severe infections [54]. The cell cycle pathways mainly contained chromosome condensation, chromosome separation, and DNA replication. The other pathways included cell adhesion, cytoskeleton remodeling, DNA damage, and metabolism. According to pathway analysis, the result well confirmed that

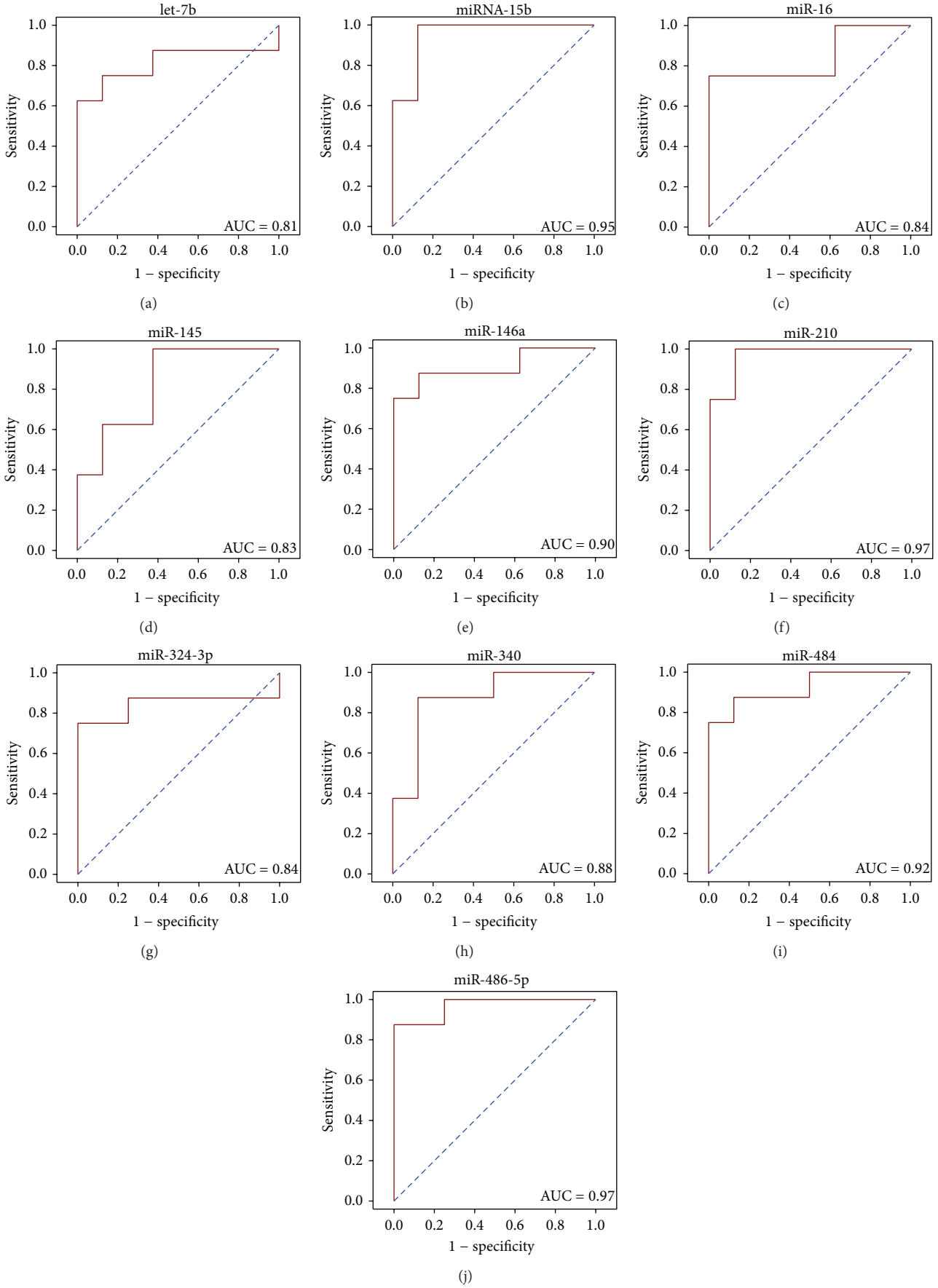


Figure 3: Receiver operating characteristic (ROC) curves of the 10 candidate miRNAs for their performance of diagnosis of sepsis.

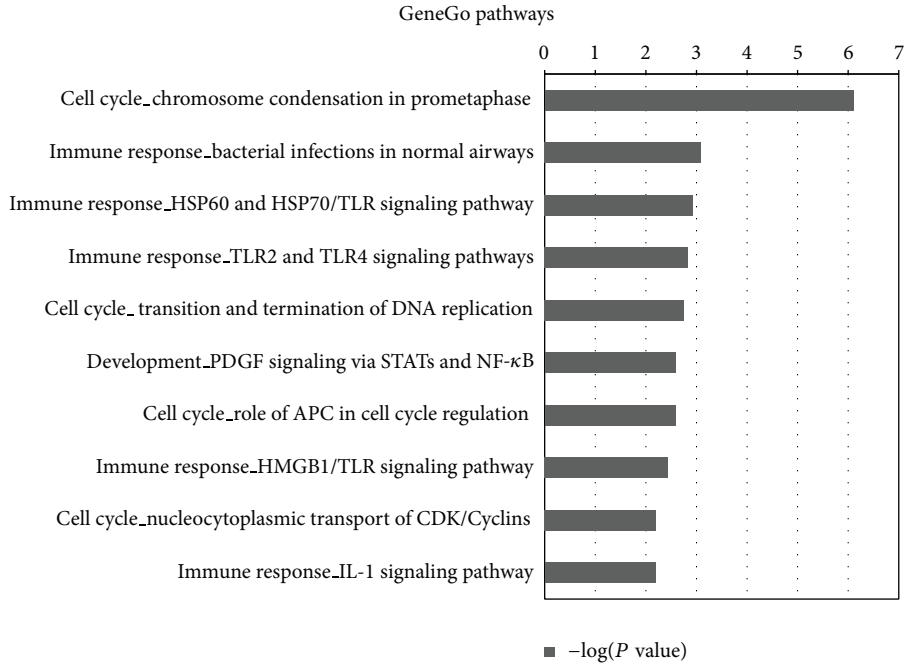


Figure 4: Pathway enrichment analysis for the target genes of the 10 candidate sepsis miRNA biomarkers. The uniquely regulated and targeted genes of the candidate sepsis miRNA biomarkers from our method were retrieved and annotated with analysis of pathway enrichment in GeneGo database. In total, 207 genes are uniquely regulated and targeted by the 10 candidate miRNA biomarkers. The statistical significance level P value was negative 10-based log transformed. Top 10 significantly enriched pathways were listed.

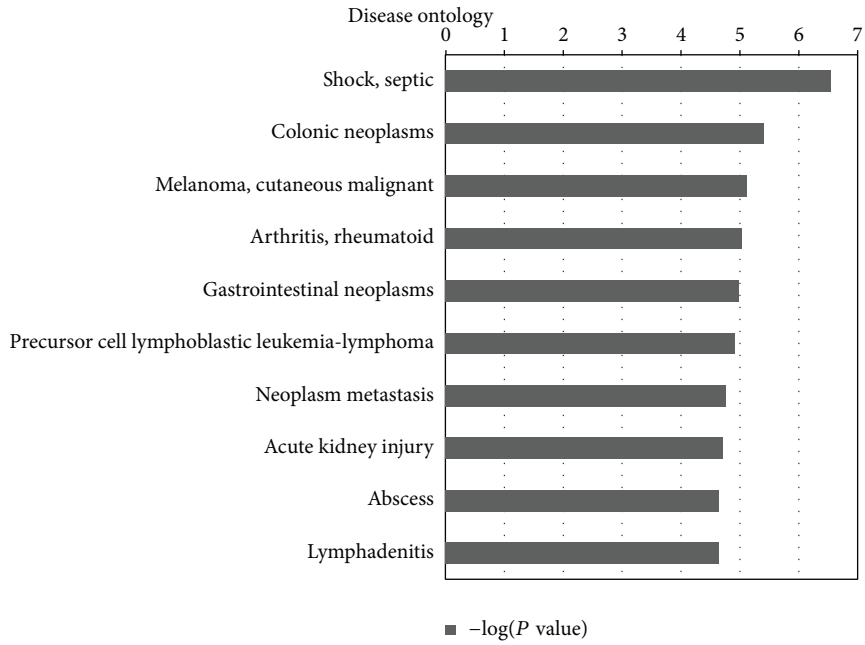


Figure 5: Disease ontology analysis for uniquely regulated and targeted genes of the 10 candidate sepsis miRNA biomarkers. The uniquely regulated and targeted genes of the candidate sepsis miRNA biomarkers from our method were retrieved and annotated with disease ontology analysis. In total, 207 genes are uniquely regulated and targeted by the 10 candidate miRNA biomarkers. The statistical significance level (P value) was negative 10-based log transformed. The top 10 significantly enriched diseases were shown.

Table 3: Summary of constructed 10 miRNA regulated PINs. N0: gene was included in PINA database; N1: the extended subnetwork of N0 gene directly connected to N0 gene; N2: the total genes of miRNA regulated subnetwork.

MicroRNA name (Hsa-)	Accession number (MIMAT)	NOD count	N0 count	N1 count	N2 count
let-7b	0000063	53	42	424	466
miR-15b	0000417	33	26	201	227
miR-16	0000069	35	28	384	412
miR-145	0000437	11	8	256	264
miR-146a	0000449	20	13	202	215
miR-210	0000267	15	10	39	49
miR-324-3p	0000762	10	10	121	131
miR-340	0004692	11	9	124	133
miR-484	0002174	11	11	246	257
miR-486-5p	0002177	8	6	26	32

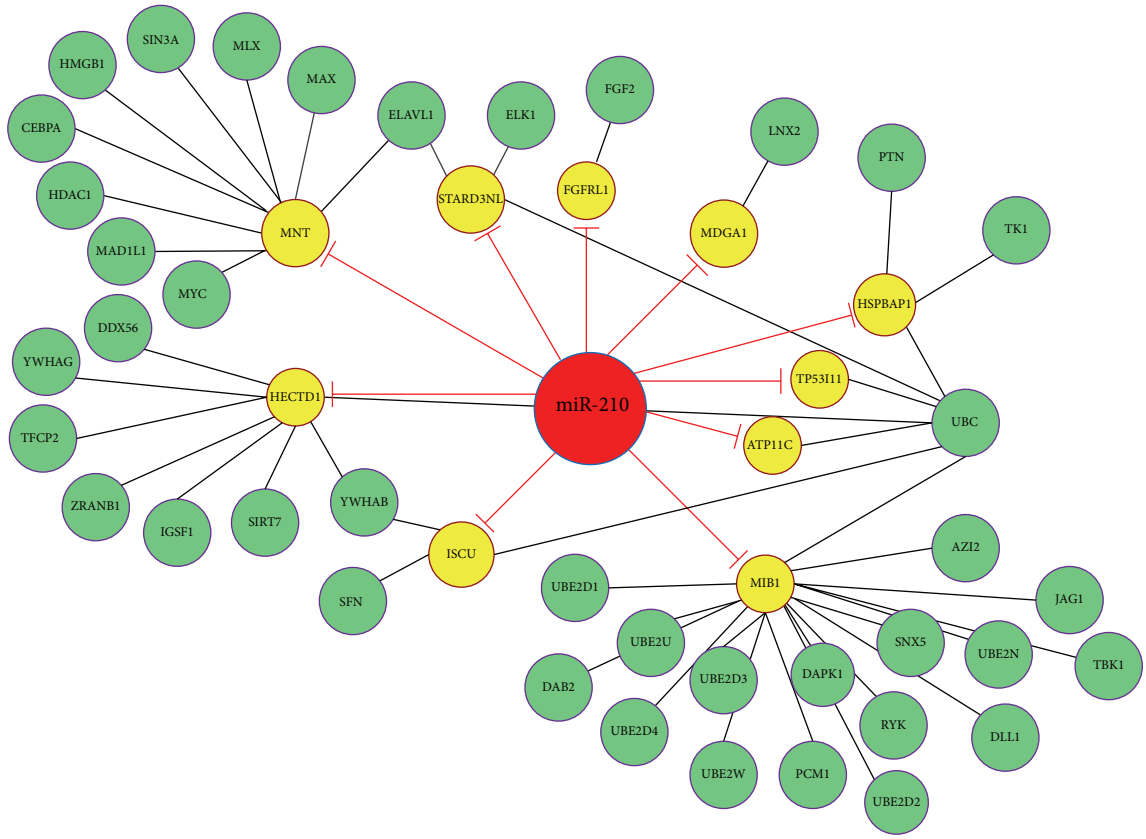


Figure 6: The miRNA-210 regulated protein-protein interaction network (PPIN). In this network, red node denotes the miRNA, yellow nodes denote miRNA directly targeted genes, and green nodes denote genes connected with target genes. The red lines represent a negative regulatory relationship initiated by miRNAs. The black lines represent interactions between protein and protein.

the abnormal expression of candidate miRNAs can cause specific signaling pathway to be active in sepsis progress, and their target genes are closely related to sepsis. Therefore, our predicted candidate miRNAs are reliable for sepsis. The top 10 significant GeneGo pathways enriched with the target genes of the predicted candidate sepsis miRNAs are shown in Figure 4.

Disease ontology is created based on the classification in medical subject headings (MeSH). Each disease in disease ontology has its corresponding biomarker gene or set of genes. After mapping the uniquely regulated and targeted genes of candidate miRNA biomarkers, we noted that the most significant disease is septic shock. Septic shock is severe sepsis plus a state of acute circulatory failure characterized by persistent arterial hypotension unexplained by other causes despite adequate volume resuscitation [55]. Based on the principle of disease ontology in GeneGo, the enriched genes are disease-related biomarkers. However, these genes

Table 4: GO analysis results of miR-15b regulated PIN. The common GO terms for miR-15b were listed.

MIMAT0000417 (Hsa-miR-15b)		
GO term	Genes	*P* value
GO:0006916~antiapoptosis	BFAR, HSP90B1, GSK3B, BCL2, HIPK3, TGFBR1, NPM1, UBC, SERPINB2, FAIM3, BCL2L1, HSPA5	$2.96E-04$
GO:0009891~positive regulation of biosynthetic process	DVL3, HRAS, THRB, GRIP1, PCBD1, RXRB, RXRA, TGFBR1, PPARG, DDX5, CALR, POT1, SREBF2, ATXN1, MAPK1, MEIS2, PSMC5, NCOA2, HNF4A, ATXN7, NPM1, UBC, YAP1	$8.10E-04$
GO:0010557~positive regulation of macromolecule biosynthetic process	DVL3, HRAS, THRB, GRIP1, PCBD1, RXRB, RXRA, TGFBR1, PPARG, DDX5, CALR, POT1, SREBF2, ATXN1, MAPK1, MEIS2, PSMC5, NCOA2, HNF4A, ATXN7, UBC, YAP1	$8.92E-04$
GO:0010604~positive regulation of macromolecule metabolic process	HRAS, THRB, GRIP1, PPARG, PSMD1, PSMD2, PSMD3, H2AFX, PSMD4, YAP1, PSMD6, PSMD7, PRKCA, PCBD1, RXRB, RXRA, PSMA2, UBE2N, MAPK1, NCOA2, HNF4A, PSMA6, PSMA3, UBC, MDM2, CALR, POT1, PIN1, PSMB5, MEIS2, BCL2, UBE2D1, DVL3, TGFBR1, DDX5, FURIN, SREBF2, ATXN1, PSMC6, PSMD14, PSMD13, PSMC5, PSMD12, PSMC4, PSMC3, PSMD11, PSMD10, ATXN7, PSMC2, PSMC1	$1.54E-16$
GO:0010605~negative regulation of macromolecule metabolic process	THRB, TSG101, PPARG, BCL2L1, TERF2IP, CALR, POT1, PSMB5, MEIS2, NPM1, PSMD1, PSMD2, PSMD3, PSMD4, UBE2D1, PSMD6, PSMD7, PRKCA, RXRA, ZNF24, UBE2I, FURIN, CDK5, SIRT3, PSMA2, ATXN1, PSMD14, PSMC6, PSMD13, NCOA2, PSMC5, PSMA6, HNF4A, PSMD12, PSMC4, PSMC3, PSMD11, PSMD10, PSMC2, PSMA3, PSMC1, UBC, BUB1B, MDM2, FABP4, SMURF2	$3.72E-16$
GO:0010628~positive regulation of gene expression	DVL3, THRB, GRIP1, RXRB, PCBD1, RXRA, TGFBR1, PPARG, DDX5, SREBF2, ATXN1, MAPK1, MEIS2, PSMC5, NCOA2, HNF4A, ATXN7, UBC, YAP1	0.0031
GO:0010941~regulation of cell death	HRAS, BCAR1, BCL2L1, CALR, ITSN1, DYNLL1, BCL2, SOS1, CASP8, RAC1, NPM1, POU4F1, HSPA5, PRKCA, VAV3, TP53BP2, TGFBR1, TMBIM6, RXRA, ACTN1, ACTN2, FURIN, VAV1, CDK5, CASP10, MAPK1, BFAR, HSP90B1, PSMC5, GSK3B, HIPK3, UBC, SERPINB2, ERN1, FAIM3, MAPK8, CACNA1A	$4.80E-09$
GO:0031328~positive regulation of cellular biosynthetic process	DVL3, HRAS, THRB, GRIP1, PCBD1, RXRB, RXRA, TGFBR1, PPARG, DDX5, CALR, POT1, SREBF2, ATXN1, MAPK1, MEIS2, PSMC5, NCOA2, HNF4A, ATXN7, NPM1, UBC, YAP1	$6.69E-04$
GO:0042981~regulation of apoptosis	HRAS, BCAR1, BCL2L1, CALR, ITSN1, DYNLL1, BCL2, SOS1, CASP8, RAC1, NPM1, POU4F1, HSPA5, PRKCA, VAV3, TP53BP2, TGFBR1, TMBIM6, RXRA, ACTN1, ACTN2, FURIN, VAV1, CDK5, CASP10, MAPK1, BFAR, HSP90B1, GSK3B, HIPK3, UBC, SERPINB2, ERN1, FAIM3, MAPK8, CACNA1A	$1.18E-08$
GO:0043066~negative regulation of apoptosis	HRAS, TMBIM6, TGFBR1, BCL2L1, ITSN1, FURIN, BFAR, HSP90B1, GSK3B, HIPK3, BCL2, NPM1, UBC, SERPINB2, FAIM3, MAPK8, HSPA5, CACNA1A	$2.61E-05$
GO:0043067~regulation of programmed cell death	HRAS, BCAR1, BCL2L1, CALR, ITSN1, DYNLL1, BCL2, SOS1, CASP8, RAC1, NPM1, POU4F1, HSPA5, PRKCA, VAV3, TP53BP2, TGFBR1, TMBIM6, RXRA, ACTN1, ACTN2, FURIN, VAV1, CDK5, CASP10, MAPK1, BFAR, HSP90B1, PSMC5, GSK3B, HIPK3, UBC, SERPINB2, ERN1, FAIM3, MAPK8, CACNA1A	$4.35E-09$
GO:0043069~negative regulation of programmed cell death	HRAS, TMBIM6, TGFBR1, BCL2L1, ITSN1, FURIN, BFAR, HSP90B1, PSMC5, GSK3B, HIPK3, BCL2, NPM1, UBC, SERPINB2, FAIM3, MAPK8, HSPA5, CACNA1A	$8.38E-06$
GO:0045941~positive regulation of transcription	DVL3, THRB, GRIP1, RXRB, PCBD1, RXRA, TGFBR1, PPARG, DDX5, SREBF2, ATXN1, MAPK1, MEIS2, PSMC5, NCOA2, HNF4A, ATXN7, UBC, YAP1	0.0022
GO:0060548~negative regulation of cell death	HRAS, TMBIM6, TGFBR1, BCL2L1, ITSN1, FURIN, BFAR, HSP90B1, PSMC5, GSK3B, HIPK3, BCL2, NPM1, UBC, SERPINB2, FAIM3, MAPK8, HSPA5, CACNA1A	$8.76E-06$

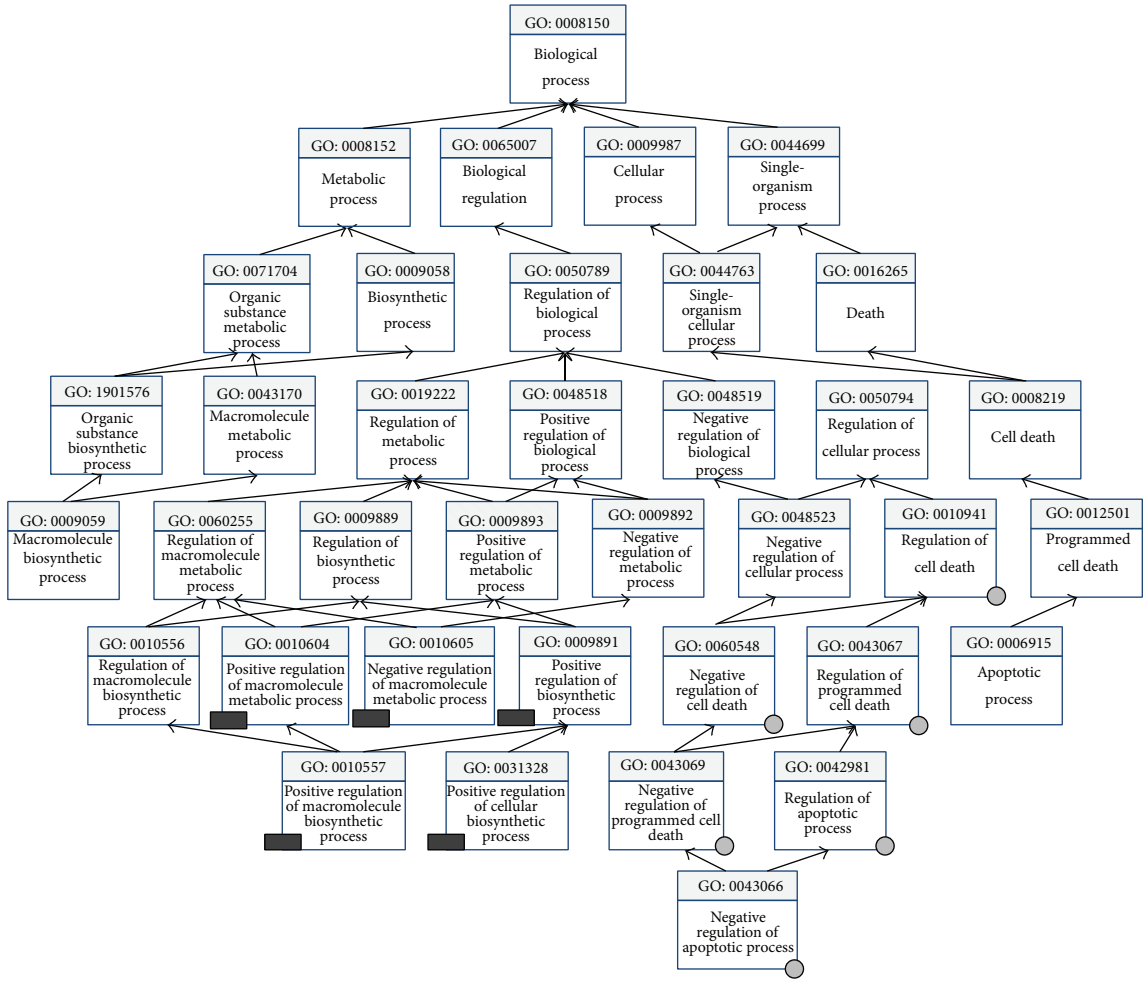


Figure 7: The ancestor chart for common GO terms obtained from the GO analysis of the 10 candidate miRNAs. The grey circle represents GO term related to cell death process. The black rectangle represents GO term related to macromolecule biosynthetic process. All marked GO terms are included in the common GO terms.

are all targeted exclusively by our candidate miRNAs. This fully proves the accuracy and effectiveness of our candidate miRNAs to distinguish sepsis from healthy population. The top 10 significant disease enrichment results are listed in Figure 5.

3.4. The Functions of the PINs Regulated by the Candidate miRNAs. MicroRNAs implement their function by regulating their target genes, thereby directly affecting expression of their target genes at the posttranscriptional level and the related protein-protein interaction network [56]. A fundamental view is that aberrant miRNA can regulate disease progression-related biological processes [57]. If a miRNA could be the useful diagnostic marker for sepsis, the biology function of PIN regulated by it will highly relate to sepsis progression. In order to demonstrate the regulation of miRNA in sepsis crucial biological processes, we applied gene ontology analysis for miRNA regulated PIN and then validated the reliability of our candidate miRNAs.

We constructed candidate miRNAs regulatory networks, containing miRNAs, genes exclusively targeted by them, and the genes directly connected to the targets. The extended network nodes were obtained by appending known interactions form the PINA database. Protein interaction network analysis (PINA) platform integrated protein-protein interaction data from six public curated databases containing 108477 binary interactions [58]. The details of 10 miRNA regulated PINs are listed in Table 3. Figure 6 shows miR-210 regulated protein-protein interaction network, which is one of the 10 miRNA regulated PINs constructed in our work. After the construction of the 10 PINs, GO enrichment analysis was applied to elucidate their functions. We exploited DAVID to select highly significantly enriched GO terms in biology process for each miRNA regulated PIN (P value < 0.05). We summarized the result of GO analysis and noted that the number of nodes in individual miRNA regulated PINs was different; in addition, the number of enriched GO terms for each miRNA was also different. By extracting the common GO term of the 10 candidate miRNAs, we found that a total of 14 GO terms were included in all candidate miRNAs. The result of the GO analysis for miR-15b regulated PIN was listed in Table 4 (common GO terms for each miRNA were listed

only) and result of all miRNA regulated PINs could be found in Supplementary Table S1 (see the Supplementary Material available online at http://dx.doi.org/10.1155/2014/594350).

Further studies are needed to confirm the relationship between 14 GO terms and sepsis. The 12 of 14 terms could be divided into two processes: one is cell death and the other is macromolecule biosynthetic process. As shown in Figure 7, QuickGO was applied to build ancestor chart for the common terms. The term GO~0006916 (antiapoptosis) is the same as GO~0043066 (negative regulation of apoptosis); two other terms are related to gene expression and transcription. The pathomechanism of organ failure and death in patients with sepsis remain elusive, but programmed cell death (or apoptosis) is a key feature in sepsis, especially as it involves the lymphoid system with resulting immunoparalysis [59]. Meanwhile, macromolecule biosynthetic and metabolic process is also prominent feature in sepsis; it is related to activation and release of bacterial endotoxin, which is a macromolecule engaged in initiation of cytokine cascade [60]. The results above fully testified our candidate miRNAs by targeting specific genes to affect important biology process of sepsis progression and further illuminate the reliability of miRNA as sepsis biomarker.

4. Conclusions

In this study, we applied an integrative approach to identify microRNAs as sepsis biomarkers from miRNA expression profiles. Comparing with the work by Vasilescu et al., we identified 10 novel and reliable miRNA biomarkers for sepsis, supported by our pathways analysis, disease ontology analysis, and protein-protein interaction network analysis, as well as ROC curve comparison. These putative miRNA biomarkers could hopefully promote the precision diagnosis of sepsis.

Conflict of Interests

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

Authors' Contribution

Jie Huang and Zhandong Sun contribute equally to this work.

Acknowledgments

This work was supported by Grants from Key Medical Subjects of Jiangsu Province (XK201120), Innovative Team of Jiangsu Province (LJ201114), Special Clinical Medical Science and Technology of Jiangsu Province (BL2012050 and BL2013014), Key Laboratory of Suzhou (SZS201108, SZS201307), and National Natural Science Foundation (81100371, 81370627, 81300423, and 81272143).

References

[1] F. B. Mayr, S. Yende, and D. C. Angus, "Epidemiology of severe sepsis," *Virulence*, vol. 5, no. 1, 2013.

[2] M. M. Levy, M. P. Fink, J. C. Marshall et al., "2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference," *Critical Care Medicine*, vol. 31, no. 4, pp. 1250–1256, 2003.

[3] S. E. Calvano, W. Xiao, D. R. Richards et al., "A network-based analysis of systemic inflammation in humans," *Nature*, vol. 437, no. 7061, pp. 1032–1037, 2005.

[4] D. C. Angus, W. T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, and M. R. Pinsky, "Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care," *Critical Care Medicine*, vol. 29, no. 7, pp. 1303–1310, 2001.

[5] A. Kumar, D. Roberts, K. E. Wood et al., "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock," *Critical Care Medicine*, vol. 34, no. 6, pp. 1589–1596, 2006.

[6] D. C. Angus, "The lingering consequences of sepsis: a hidden public health disaster?" *The Journal of the American Medical Association*, vol. 304, no. 16, pp. 1833–1834, 2010.

[7] C. Pierrakos and J.-L. Vincent, "Sepsis biomarkers: a review," *Critical Care*, vol. 14, no. 1, article R15, 2010.

[8] J. D. Faix, "Biomarkers of sepsis," *Critical Reviews in Clinical Laboratory Sciences*, vol. 50, no. 1, pp. 23–36, 2013.

[9] T. Chan and F. Gu, "Early diagnosis of sepsis using serum biomarkers," *Expert Review of Molecular Diagnostics*, vol. 11, no. 5, pp. 487–496, 2011.

[10] V. Ambros, "The functions of animal microRNAs," *Nature*, vol. 431, no. 7006, pp. 350–355, 2004.

[11] D. P. Bartel, "MicroRNAs: genomics, Biogenesis, Mechanism, and Function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[12] M. A. Cortez and G. A. Calin, "MicroRNA identification in plasma and serum: a new tool to diagnose and monitor diseases," *Expert Opinion on Biological Therapy*, vol. 9, no. 6, pp. 703–711, 2009.

[13] Y. Li, Z. Zhang, F. Liu, W. Vongsangnak, Q. Jing, and B. Shen, "Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis,," *Nucleic Acids Research*, vol. 40, no. 10, pp. 4298–4305, 2012.

[14] R. Ranjha and J. Paul, "Micro-RNAs in inflammatory diseases and as a link between inflammation and cancer," *Inflammation Research*, vol. 62, no. 4, pp. 343–355, 2013.

[15] S. Akkina and B. N. Becker, "MicroRNAs in kidney function and disease," *Translational Research*, vol. 157, no. 4, pp. 236–240, 2011.

[16] M. V. Iorio, M. Ferracin, C.-G. Liu et al., "MicroRNA gene expression deregulation in human breast cancer," *Cancer Research*, vol. 65, no. 16, pp. 7065–7070, 2005.

[17] M. Zhao, J. Sun, and Z. Zhao, "Synergetic regulatory networks mediated by oncogene-driven microRNAs and transcription factors in serous ovarian cancer," *Molecular BioSystems*, vol. 9, no. 12, pp. 3187–3198, 2013.

[18] H. J. Wang, P. J. Zhang, W. J. Chen et al., "Four serum microRNAs identified as diagnostic biomarkers of sepsis," *Journal of Trauma and Acute Care Surgery*, vol. 73, no. 4, pp. 850–854, 2012.

[19] H. Wang, P. Zhang, W. Chen, D. Feng, Y. Jia, and L. X. Xie, "Evidence for serum miR-15a and miR-16 levels as biomarkers that distinguish sepsis from systemic inflammatory response syndrome in human subjects," *Clinical Chemistry and Laboratory Medicine*, vol. 50, no. 8, pp. 1423–1428, 2012.

[20] H. Wang, P. Zhang, W. Chen, D. Feng, Y. Jia, and L. Xie, "Serum microRNA signatures identified by Solexa sequencing predict

sepsis patients' mortality: a prospective observational study," *PLoS ONE*, vol. 7, no. 6, Article ID e38885, 2012.

[21] Y. P. Chen, X. Jin, Z. Xiang, S. H. Chen, and Y. M. Li, "Circulating MicroRNAs as potential biomarkers for alcoholic steatohepatitis," *Liver International*, vol. 33, no. 8, pp. 1257–1265, 2013.

[22] S. Rahmann, M. Martina, J. H. Schultec, J. Kösterb, T. Marschalle, and A. Schrammb, "Identifying transcriptional miRNA biomarkers by integrating high-throughput sequencing and real-time PCR data," *Methods*, vol. 59, no. 1, pp. 154–163, 2013.

[23] H. Si, X. Sun, Y. Chen et al., "Circulating microRNA-2a and microRNA-21 as novel minimally invasive biomarkers for primary breast cancer," *Journal of Cancer Research and Clinical Oncology*, vol. 139, no. 2, pp. 223–229, 2013.

[24] H. Wang, W. Peng, X. Ouyang, W. Li, and Y. Da, "Circulating microRNAs as candidate biomarkers in patients with systemic lupus erythematosus," *Translational Research*, vol. 160, no. 3, pp. 198–206, 2012.

[25] Y. Wang, M. Chen, Z. Tao, Q. Hua, S. Chen, and B. Xiao, "Identification of predictive biomarkers for early diagnosis of larynx carcinoma based on microRNA expression data," *Cancer Genetics*, vol. 206, no. 9-10, pp. 340–346, 2013.

[26] H. Zhao, J. Shen, L. Medico, D. Wang, C. B. Ambrosone, and S. Liu, "A pilot study of circulating miRNAs as potential biomarkers of early stage breast cancer," *PLoS ONE*, vol. 5, no. 10, Article ID e13735, 2010.

[27] G. Zheng, Y. Xiong, and W. Xu al, "A two-microRNA signature as a potential biomarker for early gastric cancer," *Oncology Letters*, vol. 7, no. 3, pp. 679–684, 2014.

[28] W. Zhang, J. Zang, and X. Jinge al, "Identification of candidate miRNA biomarkers from miRNA regulatory network with application to prostate cancer," *Journal of Translational Medicine*, vol. 12, article 66, 2014.

[29] J. Chen, D. Zhang, W. Zhang et al., "Clear cell renal cell carcinoma associated microRNA expression signatures identified by an integrated bioinformatics analysis,," *Journal of Translational Medicine*, vol. 11, article169, 2013.

[30] S. Griffiths-Jones, H. K. Saini, S. Van Dongen, and A. J. Enright, "miRBase: tools for microRNA genomics," *Nucleic Acids Research*, vol. 36, no. 1, pp. D154–D158, 2008.

[31] C. Vasilescu, S. Rossi, M. Shimizu et al., "MicroRNA fingerprints identify miR-150 as a plasma prognostic marker in patients with sepsis," *PLoS ONE*, vol. 4, no. 10, Article ID e7405, 2009.

[32] V. Chongsuvivatwong, "Epicalc: epidemiological calculator," R package version 2. 15. 1. 0, 2012, http://CRAN.R-project.org/package=epicalc.

[33] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li, "miRecords: an integrated resource for microRNA-target interactions," *Nucleic Acids Research*, vol. 37, no. 1, pp. D105–D110, 2009.

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

[35] Q. Jiang, Y. Wang, Y. Hao et al., "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, no. 1, pp. D98–D104, 2009.

[36] S.-D. Hsu, F.-M. Lin, W.-Y. Wu et al., "miRTarBase: a database curates experimentally validated microRNA-target interactions," *Nucleic Acids Research*, vol. 39, no. 1, pp. D163–D169, 2011.

[37] V. A. Gennarino, M. Sardiello, R. Avellino et al., "MicroRNA target prediction by expression analysis of host genes," *Genome Research*, vol. 19, no. 3, pp. 481–490, 2009.

[38] E. R. Gamazon, H.-K. Im, S. Duan et al., "ExprTarget: an integrative approach to predicting human microRNA targets," *PLoS ONE*, vol. 5, no. 10, Article ID e13534, 2010.

[39] J.-H. Yang, J.-H. Li, P. Shao, H. Zhou, Y.-Q. Chen, and L.-H. Qu, "StarBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data," *Nucleic Acids Research*, vol. 39, no. 1, pp. D202–D209, 2011.

[40] M. Ding, H. Wang, J. Chen, B. Shen, and Z. Xu, "Identification and functional annotation of genome-wide ER-regulated genes in breast cancer based on ChIP-Seq data," *Computational and Mathematical Methods in Medicine*, vol. 2012, Article ID 568950, p. 10, 2012.

[41] G. Liu, M. Ding, J. Chen et al., "Computational analysis of microRNA function in heart development," *Acta Biochimica et Biophysica Sinica*, vol. 42, no. 9, pp. 662–670, 2010.

[42] Y. Tang, W. Yan, J. Chen, C. Luo, A. Kaipia, and B. Shen, "Identification of novel microRNA regulatory pathways associated with heterogeneous prostate cancer," *BMC Systems Biology*, vol. 7, supplement 6, 2013.

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

[44] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, and R. Apweiler, "QuickGO: a web-based tool for gene ontology searching," *Bioinformatics*, vol. 25, no. 22, pp. 3045–3046, 2009.

[45] J.-F. Wang, M.-L. Yu, G. Yu et al., "Serum miR-146a and miR-223 as potential new biomarkers for sepsis," *Biochemical and Biophysical Research Communications*, vol. 394, no. 1, pp. 184–188, 2010.

[46] H. Wang, K. Meng, W. J. Chen, D. Feng, Y. Jia, and L. Xie, "Serum miR-574-5p: a prognostic predictor of sepsis patients," *Shock*, vol. 37, no. 3, pp. 263–267, 2012.

[47] A. E. Pasquinelli, "MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship," *Nature Reviews Genetics*, vol. 13, no. 4, pp. 271–282, 2012.

[48] X. Lai, A. Bhattacharya, U. Schmitz, M. Kunz, J. Vera, and O. Wolkenhauer, "A Systems' Biology Approach to Study microRNA-mediated gene regulatory networks," *BioMed Research International:*, vol. 2013, Article ID 703849, 15 pages, 2013.

[49] B. Liu, J. Chen, and B. Shen, "Genome-wide analysis of the transcription factor binding preference of human bi-directional promoters and functional annotation of related gene pairs," *BMC Systems Biology*, vol. 5, no. 1, article S2, 2011.

[50] 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.

[51] G. Ramachandran, "Gram-positive and gram-negative bacterial toxins in sepsis: a brief review," *Virulence*, vol. 5, no. 1, 2013.

[52] R. S. Hotchkiss, P. E. Swanson, B. D. Freeman et al., "Apoptotic cell death in patients with sepsis, shock, and multiple organ dysfunction," *Critical Care Medicine*, vol. 27, no. 7, pp. 1230–1251, 1999.

[53] U. Fiedler and H. G. Augustin, "Angiopoietins: a link between angiogenesis and inflammation," *Trends in Immunology*, vol. 27, no. 12, pp. 552–558, 2006.

[54] P. M. Roger, H. Hyvernat, M. Ticchioni, G. Kumar, J. Dellamonica, and G. Bernardin, "The early phase of human sepsis is

characterized by a combination of apoptosis and proliferation of T cells," *Journal of Critical Care*, vol. 27, no. 4, pp. 384–393, 2012.

[55] J.-L. Vincent, "Clinical sepsis and septic shock—definition, diagnosis and management principles," *Langenbeck's Archives of Surgery*, vol. 393, no. 6, pp. 817–824, 2008.

[56] H. Liang and W.-H. Li, "MicroRNA regulation of human protein-protein interaction network," *RNA*, vol. 13, no. 9, pp. 1402–1408, 2007.

[57] C.-W. Tseng, C.-C. Lin, C.-N. Chen, H.-C. Huang, and H.-F. Juan, "Integrative network analysis reveals active microRNAs and their functions in gastric cancer," *BMC Systems Biology*, vol. 5, article 99, 2011.

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

[59] P. A. Ward, "Sepsis, apoptosis and complement," *Biochemical Pharmacology*, vol. 76, no. 11, pp. 1383–1388, 2008.

[60] J. Rusiecka-ZióŁkowska, M. Walszewska, J. Stekla, and B. Szponar, "Role of endotoxin in pathomechanism of sepsis," *Polski Merkuriusz Lekarski*, vol. 25, no. 147, pp. 260–265, 2008.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 309650, 16 pages
http://dx.doi.org/10.1155/2014/309650

Research Article

Evaluation and Comparison of Multiple Aligners for Next-Generation Sequencing Data Analysis

Jing Shang,[1,2] Fei Zhu,[1,3] Wanwipa Vongsangnak,[1] Yifei Tang,[1] Wenyu Zhang,[1] and Bairong Shen[1]

[1] *Center for Systems Biology, Soochow University, 1st Shizi Street, Suzhou, Jiangsu 215006, China*
[2] *Suzhou Institute of Nano-Tech and Nano-Bionics, Chinese Academy of Sciences, Suzhou 215123, China*
[3] *School of Computer Science and Technology, Soochow University, Suzhou 215006, China*

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

Received 17 December 2013; Accepted 4 February 2014; Published 23 March 2014

Academic Editor: Junfeng Xia

Next-generation sequencing (NGS) technology has rapidly advanced and generated the massive data volumes. To align and map the NGS data, biologists often randomly select a number of aligners without concerning their suitable feature, high performance, and high accuracy as well as sequence variations and polymorphisms existing on reference genome. This study aims to systematically evaluate and compare the capability of multiple aligners for NGS data analysis. To explore this capability, we firstly performed alignment algorithms comparison and classification. We further used long-read and short-read datasets from both real-life and *in silico* NGS data for comparative analysis and evaluation of these aligners focusing on three criteria, namely, application-specific alignment feature, computational performance, and alignment accuracy. Our study demonstrated the overall evaluation and comparison of multiple aligners for NGS data analysis. This serves as an important guiding resource for biologists to gain further insight into suitable selection of aligners for specific and broad applications.

1. Introduction

With a very high speed, large-scale sequencing reads, and drastically reduced costs available, next-generation sequencing (NGS) technology has appeared to be very fashionable [1]. There are a large number of studies that have successfully used NGS technology for their investigations under biological contexts of interests. For instance, in the nucleotide level, NGS technology is effectively used for genome evolution and genetic variation studies [2, 3]. In the transcription level, it is often applied for microRNA discovery and genomewide expression analysis [4, 5]. For the protein level, ChIP-sequencing technology is efficiently used for the identification of transcription factor binding sites [6] and histone modification patterns [7, 8]. Through a number of studies mentioned, undoubtedly, NGS represents a great powerful technology today which allows the massive number of sequencing reads to become available for only a short period and routinely be used for various genomewide association studies by aligning and mapping on the reference genome [9]. In recent years, there are several different aligners developed and further used for aligning and mapping for NGS data analysis. For examples, there are Mapping and Assembly with Qualities (MAQ) developed by Li et al. [10], Basic Oligonucleotide Alignment Software (BOAT) developed by Zhao et al. [11], Periodic Seed Mapping (PerM) developed by Chen et al. [12], Short Oligonucleotide Analysis Package (SOAPv2) developed by Li et al. [13, 14], and Global Alignment Short Sequence Search Software (GASSST) developed by Rizk and Lavenier [15].

In order to align and map NGS data using aligners, biologists often randomly select aligner without concerning to its feature, performance, and accuracy. Sequence variations and sequencing errors usually exist in the reference genome (e.g., repetitive regions and polymorphisms); hence, NGS reads frequently showed poor aligning and mapping [16]. In this case, if an unsuitable aligner is selected with existing repetitive regions and polymorphisms, the results may

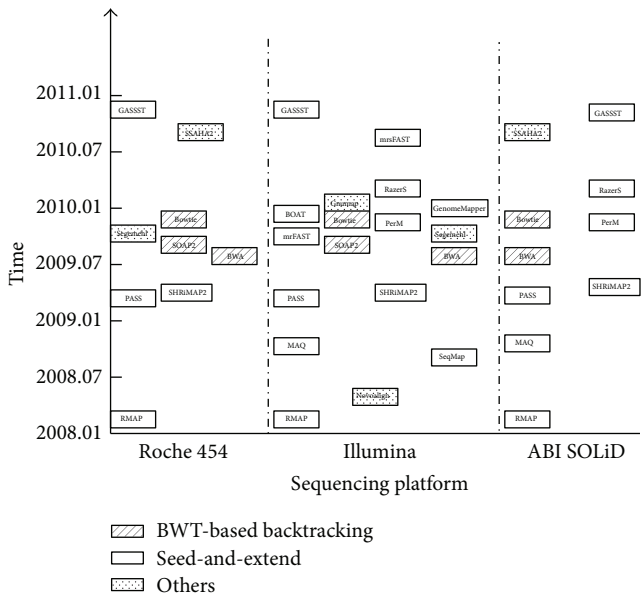


Figure 1: Aligners based on algorithms classification across different NGS platforms. Rectangles with different gray scales represent hash table-based algorithm, BWT-based backtracking algorithm, and other algorithms, individually. Aligners for specific types of data generated by different sequencing platforms are separately shown in three columns, namely, Roche 454, Illumina, and ABI SOLiD.

then convey error messages and mislead interpretation of biological outcome. It is therefore valuable for the biologists to consider the capability of individual software tool in terms of its feature, performance, and accuracy [5, 17]. This study is aimed to systematically evaluate and compare the capability of multiple aligners for NGS data analysis. Initially, we classified multiple aligners based on their developed algorithms. Here, hash table-based algorithm and Burrows-Wheeler Transform- (BWT-) based backtracking algorithm were considered. Under these two algorithms, we then selected favorable aligners for comparative analysis and further evaluation focused on three criteria (i.e., application-specific alignment feature, computational performance, and alignment accuracy). Literature searching and our own programming implementation were performed in order to evaluate different application-specific alignment features. Real-life datasets sampled from different organisms, including long-read datasets from Roche 454 sequencing platform and short-read datasets from Illumina sequencing platform, were used for comparative analysis of multiple aligners for computational performance evaluation. To further evaluate alignment accuracy, our generated *in silico* short-read and long-read datasets based on varying sequencing characteristics were used for comparison of multiple aligners. Through the end, the overall evaluation and comparison of multiple aligners with respect to the three criteria could guide the biologists for suitable selection of aligners for NGS data analysis for proper interpretation through different biological questions.

2. Results and Discussion

2.1. Algorithm-Based Classification of Multiple Aligners. Currently, three NGS platforms, namely, Roche 454, Illumina, and ABI SOLiD, are employed at large extent, of biomedical researches. SOLiD platform generated two-base encoding data to discriminate between sequencing errors and SNPs [18], while Roche 454 platform has the ability to generate reads with length up to 500 nt or even longer, which is especially specific for de novo sequencing and resequencing [16]. Illumina platform is capable of producing hundreds of millions of much shorter reads at faster speed and lower cost than others. In addition, Roche 454 platform is more likely to have higher sequencing error rate of insertions and deletions, while Illumina platform typically possesses higher sequencing error rate of mismatches [19]. To adapt to high-throughput data from three NGS platforms, multiple aligners were designed with various algorithms. According to two main strategies employed behind the multiple algorithms, multiple aligners for NGS data were classified as the hash table-based algorithm and the BWT-based backtracking algorithm. As presented in Figure 1, we show 19 aligners based on these two algorithms for the three NGS platforms. According to the popularity of multiple aligners (see Supplementary File 3 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/309650), the aligners, like RMAP, SeqMap, MAQ, SHRiMP2, BWA, SOAP2, and Bowtie, are popular for Illumina platform. RMAP, SHRiMP2, BWA, SOAP2, Bowtie, and SSAHA2 are widely applied for Roche 454, while RMAP, MAQ, SHRiMP2, BWA, Bowtie, and SSAHA2 are favorable for SOLiD platform.

To describe the hash table-based algorithm, initially, this algorithm accurately aligns massive data volumes produced by the present sequencing machines following an essential multistep strategy, called seed-and-extend [20]. To quickly identify limited subset of possible read mapping locations in the reference genome, the first step in the hash table-based algorithm is an attempt to localize the common k-mer substrings shared by both reads and genome sequences through the hash tables, called seeds detection. This step is specifically designed for accelerating high-throughput short reads. To determine the exact locations of the reads in the reference genome, the second step is subsequently to perform an extended alignment of seeds with slower and more accurate dynamic programming algorithm, such as Smith-Waterman [21] or Needleman-Wunsch algorithm. The aligners for NGS data analysis which were classified together in the hash table-based algorithm include SeqMap [22], PASS [23], MAQ [10], GASSST [15], RMAP [16], PerM [12], RazerS [24, 25], microread Fast Alignment Search Tool (mrFAST) [26], microread (substitutions only) Fast Alignment and Search Tool (mrsFAST) [27], GenomeMapper [28], and BOAT [11].

However, diverse strategies for seeds detection cause a distinction among multiple alignment algorithms. To handle the reads alignment with errors (e.g., mismatches and indels), RMAP, MAQ, SeqMap, and SOAP2 are based on the pigeonhole principle to chop the reads into small pieces to be perfectly matched to the reference genome for noncandidate

filtration during seeds detection process [10, 16, 22, 29]. Meanwhile, SHRiMP2 [30, 31] and RazerS are implemented from another similar strategy, called q-gram filter. This is an extension of the pigeonhole principle to chop the reads into overlapping pieces to be matched for noncandidate filtration [24, 30].

Furthermore, the capability to align reads with many errors existing is also important bottleneck because pieces of reads are chopped so small with increased errors that lead to multiple match locations in the reference genome [32]. Thus, the algorithm based on the idea of spaced seeds, which is utilizing seeds with nonconsecutive matches in seed detection phase [12, 15, 30, 33, 34], has been used, for instance, in PerM, SHRiMP2, RazerS, BOAT, and GASSST.

In contrast, the BWT-based backtracking algorithm aligns the entire reads instead of the seeds of reads against the substrings sampled from the reference genome. To enable rapid read searching, this algorithm stores all the suffixes of reference genome sequence based on a certain representation of data structure, including prefix/suffix tree, suffix array, and Ferragina-Manziniuse algorithm-based index (FM index) [35]. This strategy is also used to solve alignment to multiple identical copies in the reference genome sequence efficiently, which is superior to the hash table-based algorithm. To reduce the memory occupation of the data structures as mentioned above, BWT [36–38], a reversible data compression algorithm, has been used to reorder the reference genome sequence for data structure compression. Thus, BWT-based backtracking algorithm retrieves the whole BWT-based suffix array for reads aligning and mapping with rapid searching and few memory requirements. Currently, SOAP2, BWA [39, 40], and Bowtie [37] were classified together in the BWT-based backtracking algorithm. For example, Bowtie employs BWT algorithm to compress FM index, while BWA constructs BWT-based suffix array for rapid subsequence search. In conclusions, the hash table-based algorithm and BWT-based backtracking algorithm showed contradiction of the alignment algorithms. To further compare individual aligner with these two alignment algorithms mentioned above, we performed evaluation and comparative analysis of these aligners in terms of computational performance, alignment accuracy, and application-specific features. The results are described as follows.

2.2. Application-Specific Features of the Multiple Aligners. Application-specific features were mined and collected through literature searching and our own programming implementation (see Section 4). Interestingly, we found that most of the aligners could support paired-end alignment for repetitive regions mapping excluding BOAT, GASSST, Gnumap [41], GenomeMapper, and SeqMap. With regard to gapped alignment, it was clearly shown that only 5 aligners lacked the function for SNPs and structural variation discovery, namely, Bowtie, mrsFAST, MAQ, RMAP, and SSAHA2 [42]. For bisulfite alignment used in ChIP-Seq data analysis, only Gnumap, mrsFAST, Novoalign (http://www.novocraft.com/), RMAP, and Segemehl [19] were demonstrated to support this function. To summarize, it was clear that Novoalign and Segemehl beneficially supported wide applications of multiple alignment features analysis, namely, gapped alignment, paired-end alignment, and bisulfite alignment. Table 1 described different application-specific features among multiple aligners.

2.3. Computational Performance Evaluation Using Real-Life Datasets. To evaluate computational performance of individual aligner, we considered three factors that were computation time, maximum memory usage, and mapped read counts as follows.

2.3.1. Computation Time Comparison. As the results shown in Figure 2(a), computation time is plotted against the favorable multiple aligners. The short-read datasets sampled from various organisms, namely, virus *PhiX174*, bacteria *Escherichia coli,* yeast *Saccharomyces cerevisiae*, fruit fly *Drosophila melanogaster*, plant *Oryza sativa*, and human *Homo sapiens*, were used to assess the impact of reference genome size on computation time. Clearly, most of aligners showed a linear relationship between the computation time and the size of reference genome. Besides the genome size, the count of reads had impact on computation time as clearly seen from 2 short-read datasets of *Homo sapiens* with different read counts. Noticeably, it should be stressed that computation time of Novoalign showed more dependence on the count of reads than reference genome size. The detailed information for real-life short-read datasets and reference genomes was listed in Table 2. In such a case of comparison between plant genome (i.e., *O. sativa*) and human genome (i.e., *H. sapiens*), we observed that the computation time of plant genome (>5 hours) was slower than human genome (1.5 hours).

From overall results with short-read datasets produced by Illumina sequencing platform as shown in Figures 2(a) and 2(b), we observe that the computation speed for Bowtie, SOAP2, BWA, and PerM was significantly faster than the other aligners regardless of different reference genome sizes and read counts. These results may be explained by BWT-based backtracking algorithm behind Bowtie, SOAP2, and BWA which probably impacted on reduction of computation time. In particular, PerM obviously showed an outstanding computation speed due to simultaneous utility of available multiple threads. On the other hand, BOAT and RazerS required significant amounts of computation time. Their computation speed was extremely slower than the others under the same computational conditions (see Section 4). Once multiple threads are utilized, computation speed was dramatically increased, such as BOAT (see Figure 2(b)). For the other aligners, apart from Segemehl, Gnumap, and SHRiMP2 [30], the major of aligners obtained ideal computation speed during small reference genome analysis process (e.g., virus, bacteria, etc.). With multiple threads utilized, computation time of the aligners was significantly reduced, such as PASS, GASSST, SHRiMP2, and Segemehl. The results are shown in Figure 2(b). In addition, Figure 2(c) shows a plot of computation time against multiple aligners, regarding long-read datasets generated by Roche 454 sequencing

Table 1: Application-specific alignment features distribution among multiple aligners.

Aligners	Operate system	Programming language	Input Format[1]? (Fasta and Fastq)	Output format	Multithread?	Gapped alignment?	Paired-end alignment?	Trimming alignment?	Bisulfite alignment?	Note
Bowtie	★	C++	√	SAM	√		√	√		Maximum allowed mismatches ≤3
BWA	◎	C++	√	SAM	√	√	√			BWA-short: 200 bp; BWA-SW: 100 kbp
BOAT	◎	C	√	*	√	√				Maximum allowed mismatches ≤3
GASSST	◎	C++	Fasta	SAM	√	√				Merely Fasta format required for reads
Gnumap	◎	C	√ (prb)	SAM	√	√		√	√	Maximum read length <1000 bp
GenomeMapper	◎	C	√	BED	√	√				Maximum read length < 2000 bp
mrFAST	★	C	√	SAM		√	√			Maximum read length <300 bp
mrsFAST	★	C	√	SAM			√		√	Maximum read length <200 bp
MAQ	◎	C++	Fastq	map			√			Maximum read length ≤128 bp
NovoAlign	●	C++	√	SAM	√	√	√	√	√	Restrictions for academic version
PASS	※	C++	√ (sff)	GFF3	√	√	√			Maximum read length <1000 bp
PerM	※	C++	√	SAM	√		√	√		Maximum read length ≤128 bp
RazerS	★	C++	√ (prb)	Eland, GFF		√	√	√		Arbitrary read length
RMAP	◎	C++	√	BED			√		√	Fixed-length reads required
SeqMap	★	C++	Fasta	Eland		√				Maximum allowed mismatches ≤5
SOAPv2	◎	C++	√	*	√	√	√			Maximum read length <1000 bp
SHRiMAP2	◎	Python	Fasta	SAM	√	√	√			Parallel computing supported
Segemehl	◎	C	Fasta	*	√	√	√	√	√	Large memory usage required
SSAHA2	●	NA	√	GFF, SAM			√			For long reads mapping

[1]We here only consider short-reads input format.
※Windows, Linux, or Unix operating system.
★Windows, Linux,Unix, or Mac X operating system.
●Linux,Unix, or Mac X operating system.
◎Linux or Unix operating system.
*The short-read aligning algorithms' own output format.

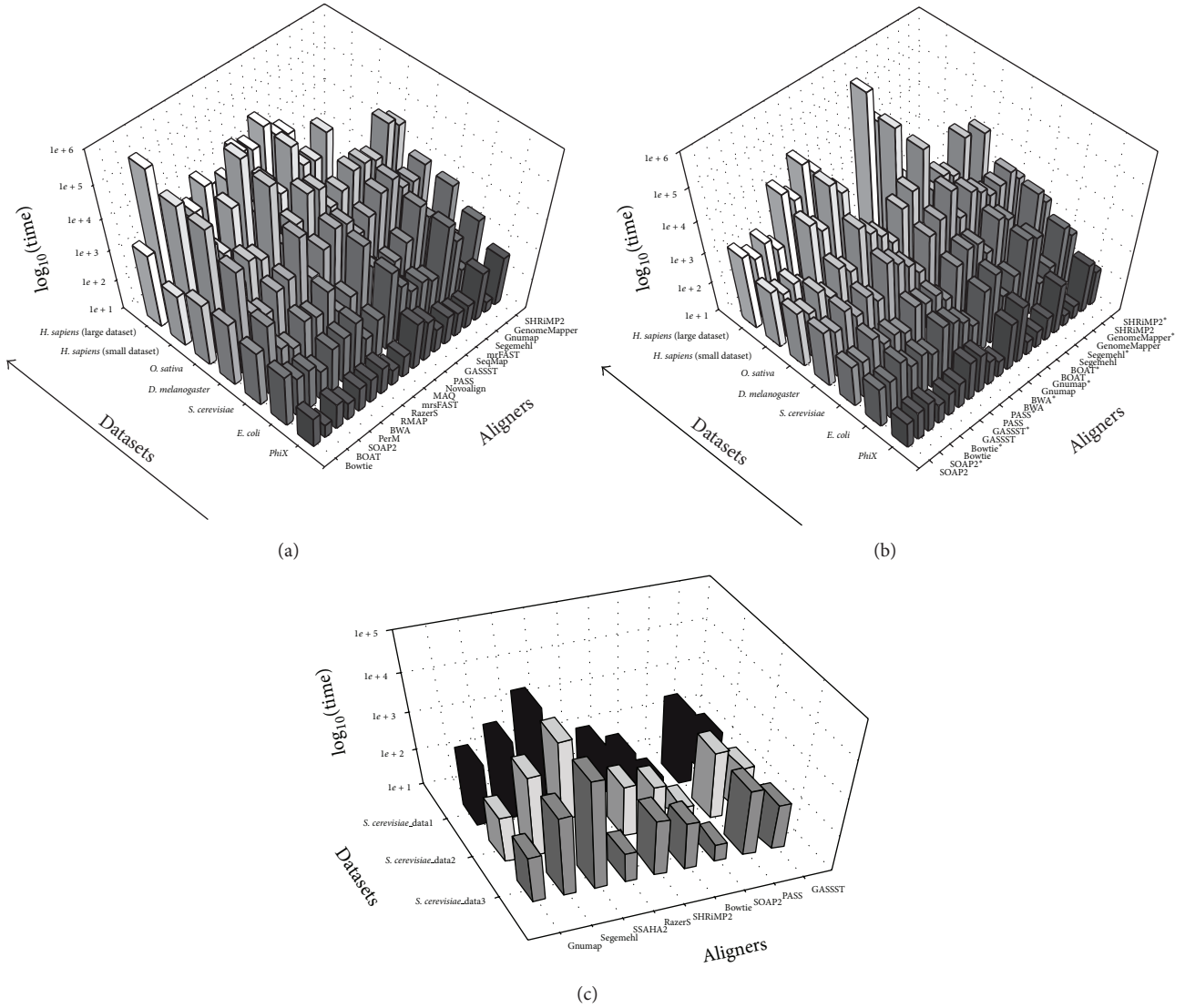


Figure 2: Bar graph illustrates a comparison of different computation time plots against multiple aligners. In this Figure, z-axis is log value of the computation time, y-axis represents real-life datasets, and x-axis represents multiple aligners under this comparison. Based on real-life short-read datasets sampled from various organisms by Illumina sequencing platform, (a) displays computation time comparison in single-thread mode, (b) displays computation time comparison for both in single-thread mode and in three-thread mode, and (c) displays computation time comparison in single-thread mode based on real-life long-read datasets by Roche 454 sequencing platform. (∗) represents the results for aligners supported multiple threads function evaluated in three-thread mode.

Table 2: Detailed information for reference genomes and real-life short-read datasets from Illumina sequencing platform.

Genome	Reads ID	Reads length (bp)	Read count	Genome size	Genome version (ID)
PhiX	ERR007488	36	4516934	<1 Mbp	NC_001422.1 (NCBI)
E. coli	SRR023978	51	9575373	5 Mbp	NC_000913.2 (NCBI)
S. cerevisiae	SRX011891	36	10995605	12 Mbp	sacCer2 (UCSC)
D. melanogaster	SRR001815	36	10760364	172 Mbp	dm3 (UCSC)
Oryza sativa	DRR000023	32	18443432	388 Mbp	NCBI
Homo sapiens	SRR037152	35	4761769	3263 Mbp	hg18 (UCSC)
Homo sapiens	SRX003935	32	18424533	3263 Mbp	hg18 (UCSC)

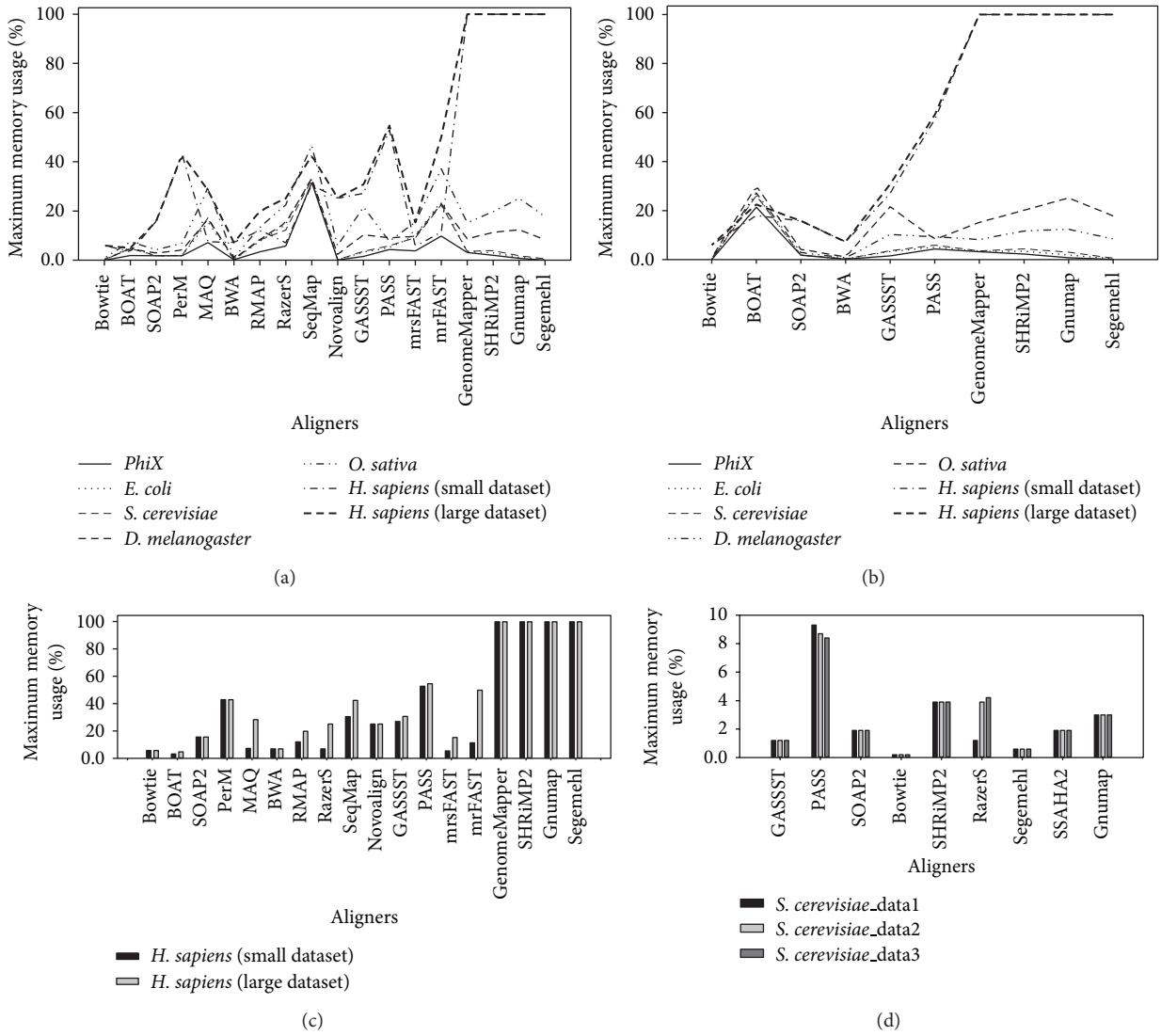


Figure 3: Graphical representation shows a comparison of various memory usage plots against multiple aligners. With real-life short-read datasets sampled from various organisms by Illumina sequencing platform, (a) shows the memory usage requirements of multiple aligners in single-thread mode, (b) shows the memory usage requirements of multiple aligners in three-thread mode, and (c) shows correlations among read count, genome size, and memory usage. Two short-read datasets (e.g., 5 million reads and 18 million reads) from *H. sapiens* were chosen to perform comparative analysis. In addition, (d) shows the memory usage requirements of multiple aligners with real-life long-read datasets produced by Roche 454 platform.

Table 3: Information for reference genomes and real-life long-read datasets from Roche 454 platform.

Genome	Reads ID	Read length (bp)	Read count	Genome size	Genome version (ID)
S. cerevisiae	SRR001091	100–200	323986	12 Mbp	sacCer2 (UCSC)
S. cerevisiae	SRR001092	100–200	409212	12 Mbp	sacCer2 (UCSC)
S. cerevisiae	SRR001093	100–200	430794	12 Mbp	sacCer2 (UCSC)

platform sampled from yeast *S. cerevisiae*. The detailed information for real-life long-read datasets was listed in Table 3. We observed that SSAHA2, Segemehl, and PASS required significant amounts of computation time; in contrast to Bowtie, SOAP2, RazerS, and GASSST relatively showed high computation speed.

2.3.2. Maximum Memory Usage Comparison. For memory usage comparison, we quantified variation of maximum memory usage by cross-comparisons among multiple aligners against maximum memory usage percentage (%) of the server. As illustrated in Figure 3(a), several bottom spots in the plot are clearly pointed out to represent the aligners with

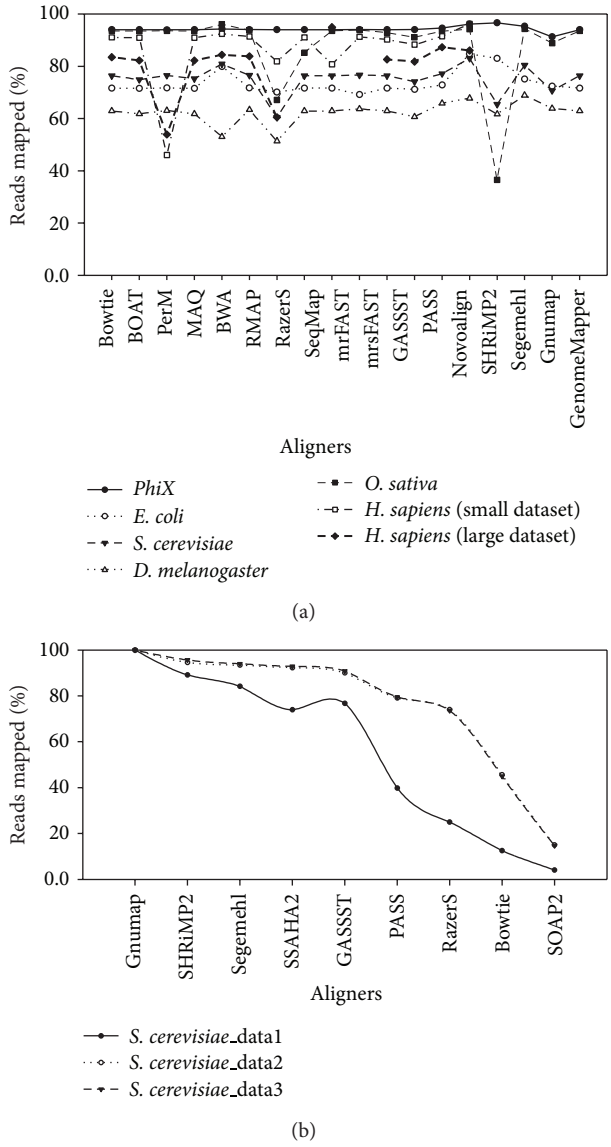


Figure 4: Graphical representation shows a comparison for different mapped reads count plots against multiple aligners with real-life short-read datasets and long-read datasets, respectively.

relatively minor memory usage during short-read datasets aligning process, which were Bowtie, BOAT, SOAP2, BWA, and mrsFAST. The maximum memory usage occupations of these aligners were relatively low and not dependent on the genome size analyzed. It was clearly seen in analysis of human genome as a reference that the maximum memory usage percentages of these aligners were 6.0%, 4.9%, 15.8%, 7.2%, and 15.4%, respectively. Thus, if even low hardware capacity was used, these aligners could not be any problem and could run with full usage on the PC computers. Yet, BOAT had dramatically increased in memory usage when multiple threads were applied (see Figure 3(b)). These results may be explained from the root of data structure constructed, such as bitmap index and prefix tree data structure in BOAT. For PerM, Novoalign, GASSST, and PASS, low memory usage was occupied with small reference genome analyzed, but a sharp increase in memory usage appeared with human genome analyzed in comparison with others, namely, 43.0%, 25.3%, 30.8%, and 54.7%, respectively. Moreover, in case of human genome analyzed, memory usages of GenomeMapper, SHRiMP2, Gnumap, and Segemehl were out of the limitation of the servers.

In addition, we found that maximum memory usage of majority of aligners was kept stable with multiple threads function employed excluding BOAT, PASS, and SHRiMP2. In particular in BOAT, it was slightly shown to be increased in memory usage (Figure 3(b)). Because of the differences in alignment algorithms constructing the index of reads, these greatly made influences on memory usage occupation. This is shown in Figure 3(c). Hence, it is apparently illustrated that the aligners, such as BOAT, MAQ, RMAP, RazerS, SeqMap, mrFAST, and mrsFAST, showed variable memory requirements mainly depending on the count of the reads instead of size of genome, while the aligners, including Bowtie, SOAP2, BWA, PerM, Novoalign, PASS, and GASSST, showed constant memory requirements regardless of the count of reads. Besides, Figure 3(d) shows comparison of the maximum memory usage of different aligners under the long-read datasets from Roche 454 sequencing platform. It was further confirmed that SOAP2, Bowtie, SHRiMP2, and Segemehl showed constant memory requirements regardless of the count of reads and the type of reads as well. Moreover, PASS seemed to show relatively higher requirement for memory usage when it deal with long-read datasets.

2.3.3. Mapped Read Counts Comparison. For mapped read counts, it is considered to be another key factor for computational performance evaluation, since it can quantify relative read density. We calculated the mapped read counts across different aligners. As shown in Figure 4(a), we observe that most aligners showed very similar results of mapped read counts excepting SOAP2, RMAP, and SHRiMP2 which represented low percentage of mapped read counts with the short-read datasets used. On the other hand, we compared the results of mapped read counts with long-read datasets as well (Figure 4(b)). It was clearly shown that SHRiMP2, Segemehl, GASSST, SSAHA2, and Gnumap had relatively better results compared with the rest of aligners. However, we could not make a judgment for capability and sensitivity of mapping aligners, since real-life data could not be employed to evaluate alignment accuracy. Further comparative analysis with *in silico* data is described in following.

2.3.4. Alignment Accuracy Evaluation Using In Silico Datasets. In order to evaluate alignment accuracy of individual aligner, we calculated sensitivity, precision, and % of multimapped reads as indicator values for evaluation. Moreover, we took mismatches, indels, and read lengths into consideration during aligning and mapping process.

To indicate alignment accuracy evaluation for short-read datasets with varying error rate existing, the results

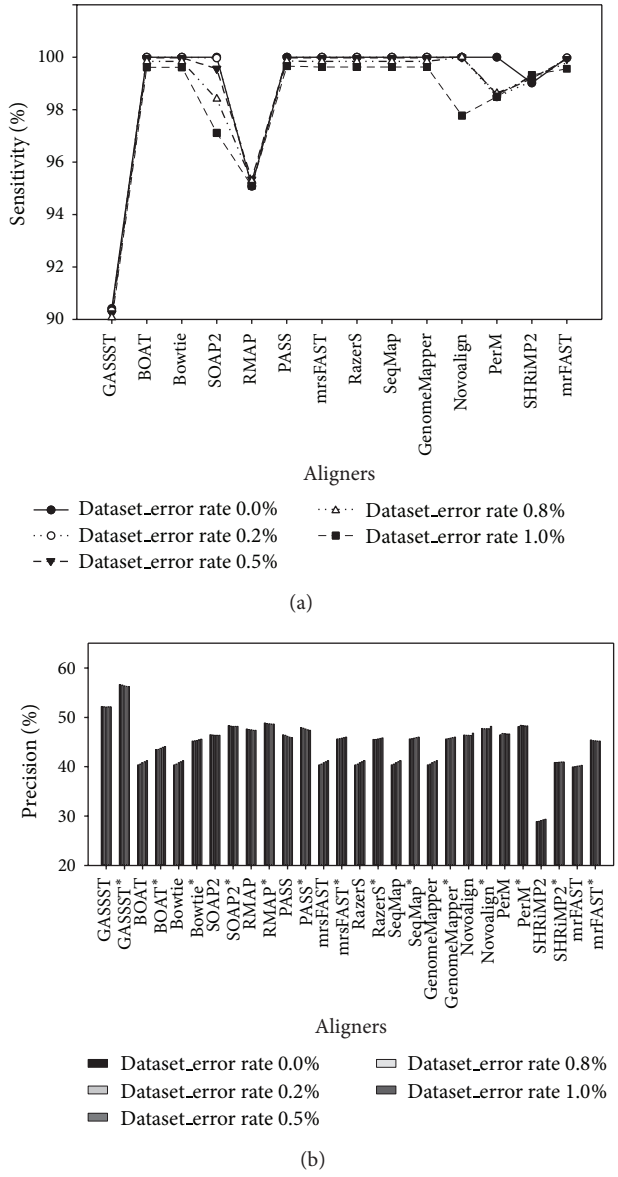


Figure 5: Graphical representation shows alignment accuracy results using *in silico* short-read datasets with varying error rates. Based on *in silico* short-read datasets sampled from chromosome X of *H. sapiens* with varying error rates (e.g., 0%, 0.2%, 0.5%, 0.8%, and 1.0%, resp.), (a) and (b) show accuracy evaluation by sensitivity and precision, respectively. Aligners with (∗) in (b) are used to show alignment accuracy evaluation by precision with consideration of multimapped reads.

are shown in Figure 5(a) for sensitivity and Figure 5(b) for precision comparison. We could see that most aligners showed relatively high sensitivity over 98%, excluding RMAP and GASSST. For Bowtie, Novoalign, and PerM, their sensitivity significantly decreased as the error rate increased (Figure 5(a)). Furthermore, Figure 5(b) also shows that GASSST possessed outstanding performance for precision comparison and PerM, Novoalign, PASS, RMAP, and SOAP2 presented the same level of precision followed behind GASSST, without consideration of multimapped reads. It was also noticed that SHRiMP2 had weak performance in terms of precision. With consideration of multimapped reads, most aligners, excluding PerM, Novoalign, PASS, RMAP, and SOAP, were slightly increased in precision, especially SHRiMP2.

As expected, Figure 6 shows alignment accuracy evaluation for short-read datasets with fixed indel frequency (0.1%) as the average indel sizes vary. Apparently, we found that GASSST and PerM were confirmed to have weak performance in sensitivity (<80%), but SHRiMP2, GenomeMapper, and Novoalign had relatively high sensitivity from overall results (Figure 6(a)). In addition, it can be seen

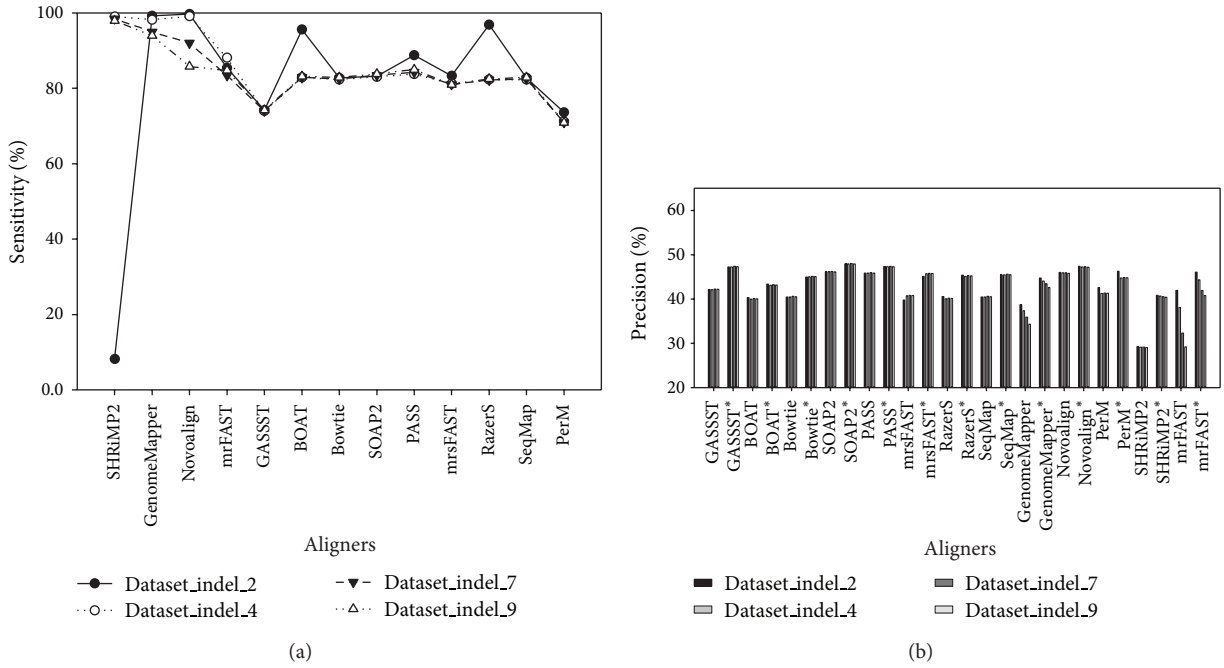


Figure 6: Graphical representation shows alignment accuracy results using *in silico* short-read datasets with varying indel sizes. Based on *in silico* short-read datasets sampled from chromosome X of *H. sapiens* with varying indel sizes (e.g., 2 bp, 4 bp, 7 bp, and 9 bp, resp.), (a) shows alignment accuracy evaluation by sensitivity and (b) shows alignment accuracy evaluation by precision. Aligners with (∗) as shown in (b) are used to show alignment accuracy evaluation by precision with consideration of multimapped reads.

in Figure 6(b) that Novoalign, PASS, SOAP2, and GASSST showed very favorable precision values, while SHRiMP2 provided the unsatisfactory precision value. However, it can also been seen that precision was improved by almost 5% among GASSST, mrsFAST, mrFAST, RazerS, SeqMap, GenomeMapper, and SHRiMP2, when multimapped reads were considered. Meanwhile, it was emphasized that GenomeMapper and mrFAST might not be better suited for indel calling due to their weak accuracy in terms of both sensitivity and precision, as indel sizes significantlyincreased.

The alignment accuracy evaluation provided by multiple aligners supported long-read alignment with varying read length on *E. coli* genome was primarily highlighted in Figure 7. As seen in this figure, PASS, SHRiMP2, Segemehl, and SSAHA2 had the highest sensitivity, while SOAP2, GenomeMapper, and Bowtie presented relatively low sensitivity and their sensitivity depended strictly on read length (Figure 7(a)). Moreover, it is also clearly seen in Figure 7(b) that GASSST showed the highest sensitivity and a significant increase in sensitivity with increasing read lengths. It seems that GASSST was the most robust to longer reads and particularly useful as reads get longer.

For datasets with varying error rates, indel sizes and read lengths existed; the results are shown in Figure 8. We evaluated % of total multimapped reads and % of corrected multimapped reads. As presented in Figure 8, the results were used to confirm influence of multimapped reads on alignment accuracy. GASSST, SHRiMP2, GenomeMapper, SeqMap, RazerS, mrFAST, mrsFAST, Bowtie, and BOAT could provide relatively high percentage of total multimapped reads (>20%) and high percentage of corrected multimapped reads as well when dealing with short-read datasets with varying error rates and indel sizes, especially SHRiMP2 (Figures 8(a) and 8(b)). However, it was indicated that these aligners could provide more information within multimapped reads, and this might result in missing important biological information without consideration of multimapped reads. In contrast, when dealing with long-read datasets with varying read lengths, the situation showed a tremendous difference in percentage of total multimapped reads and correctly mapped multimapped reads. Less information was provided by all the aligners within multimapped reads for long-read aligning and mapping. The results are shown in Figure 8(c).

3. Conclusions

Currently, optimal aligners have been called for the variety of applications and specific types of data-based NGS technology. This study aims to systematically evaluate and compare the capability of multiple aligners to provide guiding resource for choosing suitable aligners dependent on the user's specific research aims with NGS data. We evaluated multiple aligners based on criteria, including application-specific alignment feature, computational efficiency, and alignment accuracy. To assess the multiple aligners, real-life short-read datasets and long-read datasets sampled from various organisms and *in silico* datasets with varying error rates, indel sizes, and read lengths were considered as standard datasets for different applications and sequencing technologies. Table 4

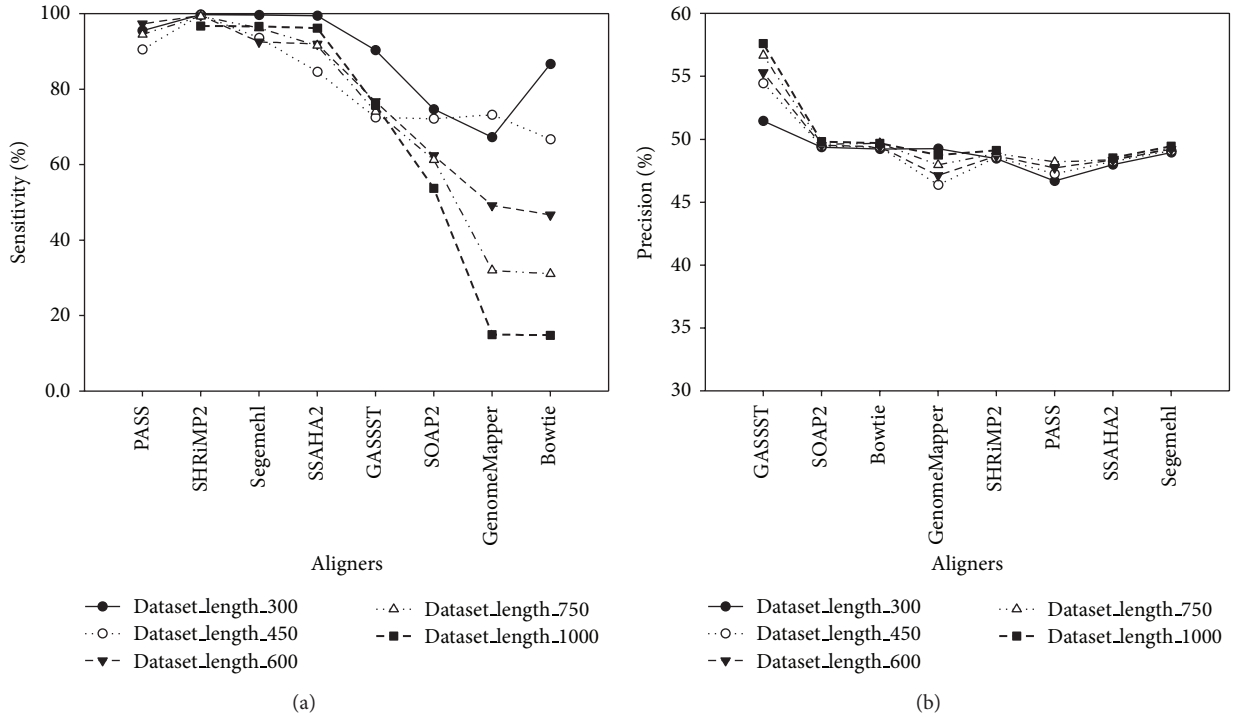


Figure 7: Graphical representation shows alignment accuracy results using *in silico* long-read datasets with varying read lengths. Based on *in silico* long-read datasets sampled from *E. coli* at different lengths of 300 bp, 450 bp, 600 bp, 750 bp, and 1000 bp, evaluated by 8 aligners (e.g., GASSST, Bowtie, SOAP2, PASS, SSAHA, SHRiMP2, GenomeMapper, and Segemehl), (a) shows alignment accuracy evaluation by sensitivity and (b) shows alignment accuracy evaluation by precision.

provided the overall summary on aligning and mapping evaluations in terms of computation speed, memory usage, and accuracy as well. It is concluded that Bowtie, BWA, and SOAP2 clearly show high computational efficiency in single-thread mode and increasing trend of computation efficiency in multi-thread mode on real-life datasets. However, PerM and Novoalign show outstanding performance on improving computation efficiency by adjusting thread mode automatically and indexing read datasets, respectively. Indeed we conclude that they can be suitable and efficient aligners for short-read aligning and mapping. It is also shown that memory usage requirements of Bowtie, BWA, BOAT, mrsFAST, and SOAP2 are relatively low both in single-thread mode and multithread mode and their memory usage requirements are kept low regardless of the number of reads and the size of genomes. Moreover, it could be seen that GenomeMapper, Novoalign, and SHRiMP2 show high sensitivity, while GASSST, Novoalign, PASS, and SOAP2 show high precision when dealing with mismatch and indel errors existed in simulated datasets. With high alignment accuracy evaluation obtained from *in silico* datasets, we conclude that GASSST, PerM, Novoalign, PASS, RMAP, and SOAP2 can be better choices, since they possess high accuracy without indels for ungapped alignment, while Novoalign, PASS, and SOAP2 have high accuracy with indels for gapped alignment. In particular, GASSST can be a candidate aligner for long reads aligning and mapping. In addition, it is implied that Novoalign and Segemehl can be representative aligners to apply for wide applications, such as gapped alignment for SNPs and structural variation discovery, paired-end alignment for mapping of repetitive region, bisulfite alignment for ChIP sequencing data analysis, and SNPs calling. Finally, we believe that our evaluation will be a benefit for biologists engaged in variety of genomics researches. The overall evaluation and comparison of multiple aligners for NGS data analysis might serve as an essential recommendation for suitable selection of aligners.

4. Methods

The pipeline of the whole procedure in this study is illustrated in Figure 9. We collected 25 unspliced read aligners developed for NGS data from different websites and published articles (Supplementary File 1). Notably, spliced read aligners were not taken in this evaluation and comparison because they were primarily used to map the reads from exon-exon junctions, which were specific algorithm for RNA-Seq [43]. However, the aligners with any extra mandatory, which made them unavailable for most of biologists, were not taken into

Table 4: Overall evaluation and comparison of multiple aligners.

Aligners	Computational speed				Memory usage		Accuracy			
	Speed with single thread	Speed with multithread	Key factor impacting speed (genome size or read count)	Overall evaluation	Key factor impacting memory (Genome size or read count)	Memory usage with multithread	Sensitivity	Precision	% of multimapped	%Corrected Multi-Mapped
Bowtie1	Fast	↑	Genome size	Low	Genome size	≡	High	—	—	
BWA	Fast	↑	Both	Low	Genome size	≡				
BOAT	Slow	↑↑	Genome size	Low	Read count	↑↑	High	—	—	Low
GASSST	—	↑	Genome size	High**	Genome size	≡	Low	High	—	
Gnumap	Slow	↓	Genome size	High**	Genome size	≡				
GenomeMapper	Slow	≡	Genome size	Low$^{\blacktriangle}$	Genome size	≡	High	—	—	
mrFAST	Slow	×	Genome size	High**	Read count	×	High	—	—	
mrsFAST	—	×	Genome size	Low	Read count	×	High	—	—	
MAQ	—	×	Genome size	High**	Read count	×				
NovoAlign$^{\#}$	—	/	Read count	Low$^{\blacktriangle}$	Genome size	/	High	High	Low	Low
PASS	—	↑	Genome size	Low$^{\blacktriangle}$	Genome size	↑	High	High	Low	Low
PerM$^{※}$	Fast		Genome size	Low$^{\blacktriangle}$	Genome size	/	Ind: low	—	Low	
RazerS	Slow	×	Genome size	High**	Read count	×	High	—	—	
RMAP	—	×	Genome size	High*	Genome size	×	Mis: low	High	Low	
SeqMap	—	×	Genome size	High***	Read count	×	High	—	—	
SOAPv2	Fast	↑	Genome size	Low	Genome size	≡	High	High	Low	
SHRiMAP2	Slow	↑	Genome size	High**	Genome size	↑	High	Low	High	
Segemehl	—	↑	Both	High***	Genome size	≡	High	—		

PerM$^{※}$ could adjust the threads automatically during running process.
Novoalign$^{\#}$ could support multithread only for commercial version.
For computational speed, we defined the aligners which are extremely faster than others as fast, while we defined the ones which are extremely slower as slow.
For memory usage, we evaluated the aligners as follow: among the s even datasets, the maximum memory usage ≤4 G, low; the maximum memory usage ≥32 G, high***.
Low$^{\blacktriangle}$ represents that the maximum memory usage will have an extreme increase with *H. sapiens* datasets (≥4 G).
×: without multithread function.
— represents medium level remark.
≡ means there is no obvious change.

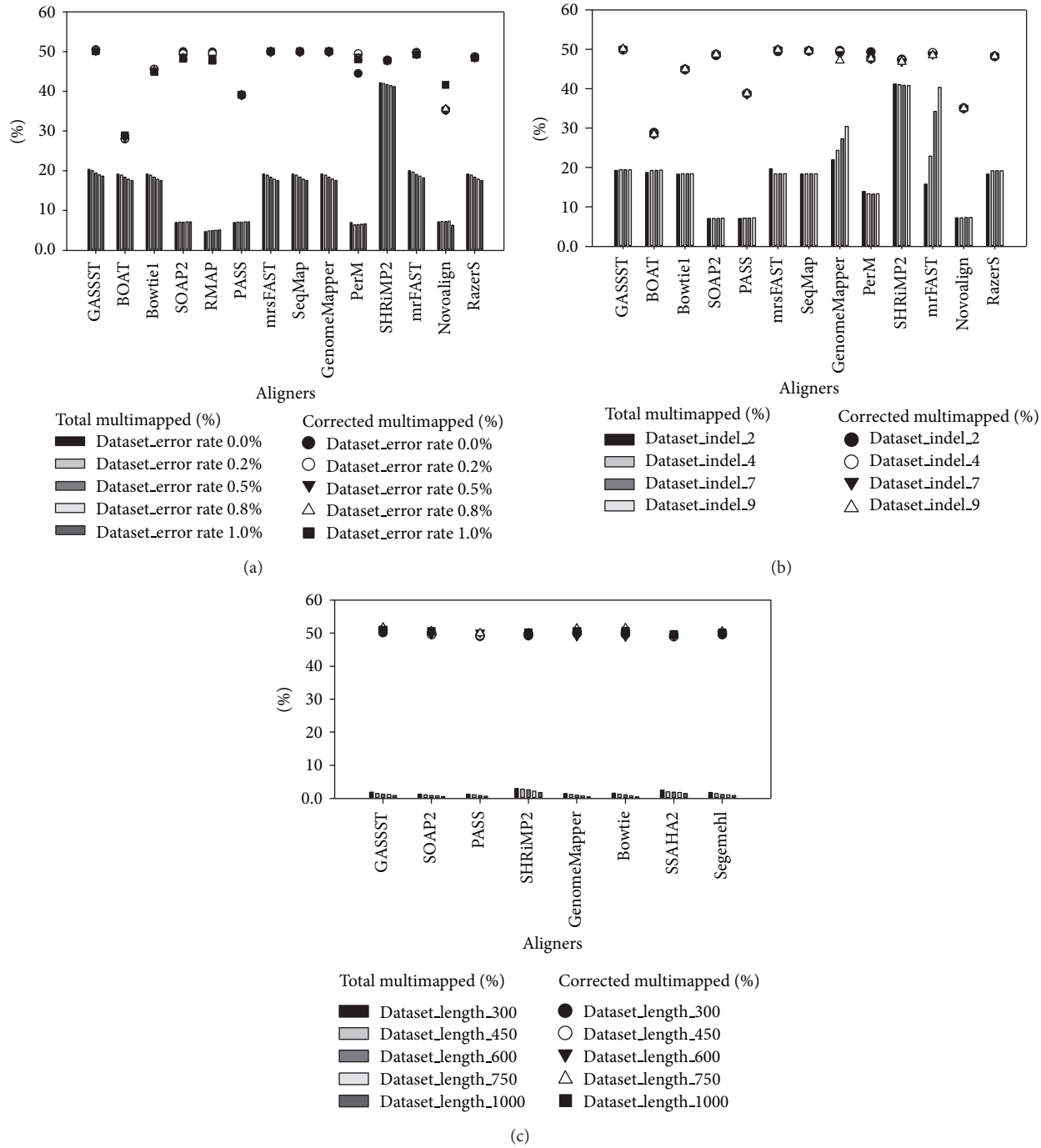


Figure 8: Graphical representation shows impact of total multimapped reads and corrected multimapped reads on alignment accuracy results using *in silico* datasets. (a), (b), and (c) show % of total multimapped reads and % of corrected multimapped reads for *in silico* datasets with varying error rates, indel sizes, and read lengths, respectively.

account. For example, SOAP3 [44] depended on a CUDA-enabled GPU, CloudBurst [45] required cloud computing, and ZOOM [33] was commercial version. Therefore, 19 favorable aligners were eventually considered for further evaluation and comparison process. Details of the selected aligners are shown in Supplementary File 2. Supplementary File 3 shows the number of citation papers associated with each aligner in order to provide the information of the popularity. In the following, we describe evaluation and comparison of the multiple aligners.

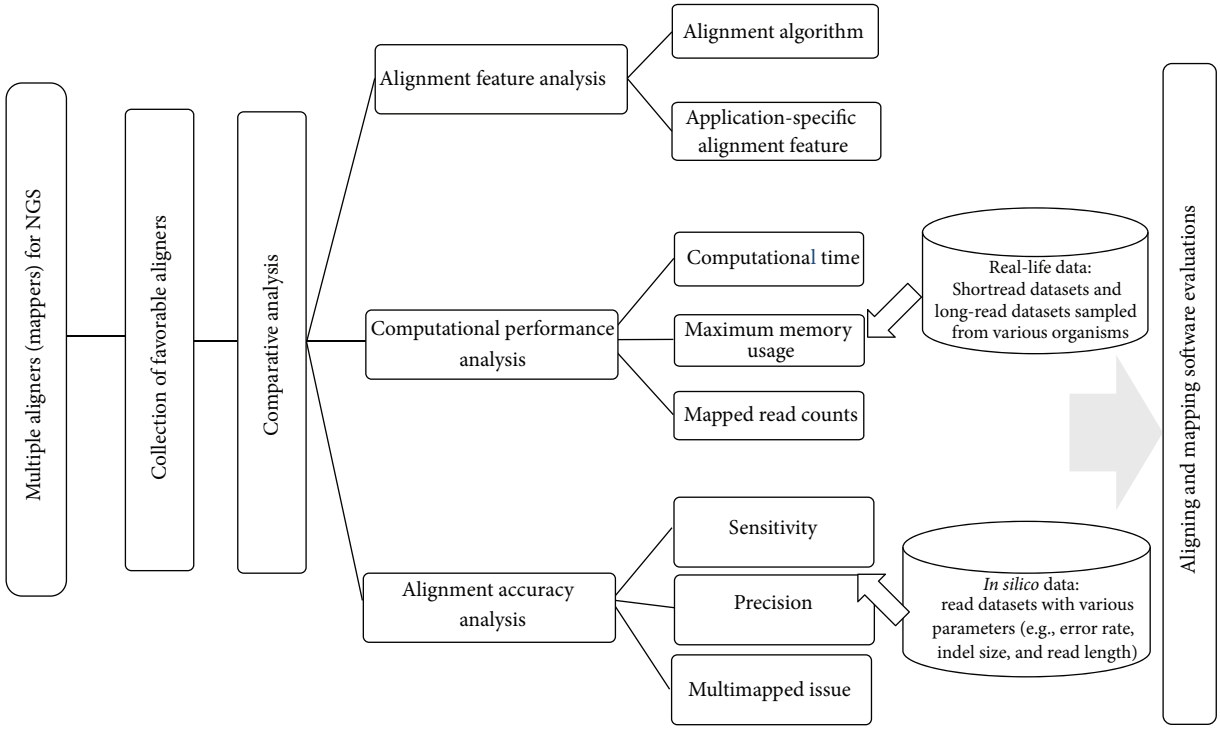


Figure 9: Flow chart for evaluation and comparison process of multiple aligners. The process contains three main steps, namely, alignment feature comparison, computational performance comparison, and alignment accuracy comparison for NGS data analysis.

4.1. Evaluation and Comparison of the Multiple Aligners

4.1.1. Literature Searching and Programming Implementation for Application-Specific Alignment Features Evaluation. To evaluate application-specific alignment features, at the beginning, we performed literature searching to grasp and compare the alignment algorithms of 19 favorable aligners. Based on principal common characteristics of alignment algorithms sharing by multiple aligners, we then classified these aligners into two different algorithms applied, namely, hash table-based algorithm and BWT-based backtracking algorithm. However, information about important alignment features or characteristics of the multiple aligners is essential for various genomewide association studies. To collect and evaluate application-specific alignment features, we manually mined literature and other documentation and inspected the source code for individual aligner. Moreover, we implemented our own programming for individual aligner according to its alignment features as well. The application-specific alignment features were considered as follows: multithread, gapped alignment analysis, paired-end alignment analysis, trimming alignment analysis, and bisulfite alignment analysis.

4.1.2. Using Real-Life Data for Accessing Computational Performance. To evaluate computational performance for different practical applications, we used 3 real-life long-read datasets from Roche 454 sequencing platform and 7 real-life short-read datasets from Illumina sequencing platform as representative input. They were sampled from various organisms, namely, virus *PhiX174* (1 dataset), bacteria *Escherichia coli* (1 dataset), yeast *Saccharomyces cerevisiae* (4 datasets), *fruit fly Drosophila melanogaster* (1 dataset), plant *Oryza sativa* (1 dataset), and human *Homo sapiens* (2 datasets). They were downloaded from National Center for Biotechnology Information (NCBI) Short-Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/). In addition, the reference genome sequences were obtained from NCBI (http://www.ncbi.nlm.nih.gov/) and UCSC Genome Browser Home (http://genome.ucsc.edu/). The description of real-life datasets from different sequencing platform was detailed in Tables 2 and 3.

Besides input data used for evaluation, computer hardware requirements and determined parameters setting were also concerned. For computer hardware platform, we used a large-memory server with a four-core 2.4 GHzAMD Opteron processor and a maximum of 32 GB of RAM. For parameters setting, two mismatches were allowed within a full read length without considering any insertions and deletions (indels) during the mapping process of Illumina short-read datasets, while gapped alignment was allowed considering indels during the mapping process of Roche 454 long-read datasets, since indel frequency is extremely low within short-read datasets produced by Illumina sequencing platform instead of long-read datasets produced by Roche 454 sequencing platform. In addition to the default parameter values, the other parameters for each aligner were applied in an attempt to achieve parameter optimization. To account for threading when assessing computational efficiency, we employed all the aligners to perform aligning process in single-thread mode without any competition and we also

Table 5: Parameters setting for *in silico* data: read lengths, error rates, indel sizes, indel freq.

Accuracy Evaluation	Read length (bp)	Read number	Error rate (%)	Indel size (bp)	Indel freq. (%)
Mismatch factor	50	5000000	0, 0.2, 0.5, 0.8, 1.0	0	0
Indel factor	50	5000000	0	2, 4, 7, 9	0.1
Read-length factor	300, 450, 600, 750, 1000	1000000	0.5	4	0.1

were careful about some aligners supported multiple threads function to accelerate computation speed; thus these aligners were evaluated and compared in three-thread mode without any competition.

Computational performance was evaluated by consideration of three factors: computation time, maximum memory usage, and mapped read counts. These three factors mainly used to measure computational efficiency, hardware availability, and qualified read density. To obtain computation time, wall-clock time was computed for each computational process with excluding index time. Since computation time was slightly affected by computational condition of the hardware, minor discrepancy appeared definitely during each computational process. Thus, we chose the set of results under relatively stable computational process as representative results across multiple runs.

To record maximum memory usage, we developed a tool written by Python (Supplementary File 4) to monitor each programming process and then reported maximum memory usage percentage of our server's memory (32G). For mapped read counts, not only we considered uniquely mapped reads but also multimapped reads were included in the mapped reads to provide a rough perspective of alignment sensitivity for each aligner.

4.1.3. Using In Silico Data for Accessing Alignment Accuracy. To access capability of individual aligner, we evaluated not only computational performance but also alignment accuracy. It has limitations to use real-life data for accessing alignment accuracy, since true alignment locations are unknown. Hereby, we therefore wrote a Perl script to generate *in silico* data by computational simulation (Supplementary File 5). Concerning influence of mismatches, indels, and read lengths, *in silico* datasets were therefore generated according to the characteristics as listed in Table 5. The characteristics included read lengths, read counts, sequencing error rates, indel sizes, and indel frequency. Once the simulating completed, 9 *in silico* short-read datasets from chromosome *X* of *H. sapiens* were achieved. In addition to short-read datasets, we also simulated 5 long-read datasets from *E. coli* with different lengths. Besides *in silico* data, computer hardware requirements were similarly determined as previously described for accessing computational performance section. Exceptionally during the mapping process, parameters (e.g., maximum allowed mismatches and indels) were set upon own datasets feature. Finally, we measured the alignment accuracy of different aligners in terms of sensitivity and precision. The formula is shown as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FP} + \text{FN}} \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (1)$$

In addition, we further took multimapped reads into consideration, which were ambiguously mapped. Multimapped reads existing in alignment results frequently cause difficulty for the biologists to choose their real locations. This may result in missing some biological information. Thus, % of multimapped and % corrected of multimapped are thus applied as new criteria to access the capability of these aligners as follows:

$$\begin{aligned} &\%\ \text{Total multi-mapped reads} \\ &\quad = \frac{\text{multimapped reads}}{\text{multimapped reads} + \text{unique mapped reads}}, \\ &\%\ \text{Corrected multimapped reads} \\ &\quad = \frac{\text{Corrected multimapped reads}}{\text{multimapped reads}}. \end{aligned} \quad (2)$$

Conflict of Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (31170795, 31200989, 91230117, and 61303108), the Specialized Research Fund for the Doctoral Program of Higher Education of China (20113201110015), and the National High Technology Research and Development Program of China (863 Program, Grant no. 2012AA02A601).

References

[1] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nature Biotechnology*, vol. 26, no. 10, pp. 1135–1145, 2008.

[2] J. M. Otero, W. Vongsangnak, M. A. Asadollahi et al., "Whole genome sequencing of *Saccharomyces cerevisiae*: from genotype to phenotype for improved metabolic engineering applications," *BMC Genomics*, vol. 11, no. 1, article 723, 2010.

[3] A. V. Dalca and M. Brudno, "Genome variation discovery with high-throughput sequencing data," *Briefings in Bioinformatics*, vol. 11, no. 1, pp. 3–14, 2010.

[4] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, vol. 18, no. 9, pp. 1509–1517, 2008.

[5] Y. Li, Z. Zhang, F. Liu, W. Vongsangnak, Q. Jing, and B. Shen, "Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis," *Nucleic Acids Research*, vol. 40, no. 10, pp. 4298–4305, 2012.

[6] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," *Science*, vol. 316, no. 5830, pp. 1497–1502, 2007.

[7] T. A. Down, V. K. Rakyan, D. J. Turner et al., "A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis," *Nature Biotechnology*, vol. 26, no. 7, pp. 779–785, 2008.

[8] S. J. Cokus, S. Feng, X. Zhang et al., "Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning," *Nature*, vol. 452, no. 7184, pp. 215–219, 2008.

[9] S. Marguerat, B. T. Wilhelm, and J. Bähler, "Next-generation sequencing: applications beyond genomes," *Biochemical Society Transactions*, vol. 36, part 5, pp. 1091–1096, 2008.

[10] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome Research*, vol. 18, no. 11, pp. 1851–1858, 2008.

[11] S. Q. Zhao, J. Wang, L. Zhang et al., "BOAT: basic oligonucleotide alignment tool," *BMC Genomics*, vol. 10, 3, article S2, 2009.

[12] Y. Chen, T. Souaiaia, and T. Chen, "PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds," *Bioinformatics*, vol. 25, no. 19, pp. 2514–2521, 2009.

[13] R. Li, C. Yu, Y. Li et al., "SOAP2: an improved ultrafast tool for short read alignment," *Bioinformatics*, vol. 25, no. 15, pp. 1966–1967, 2009.

[14] R. Li, Y. Li, K. Kristiansen, and J. Wang, "SOAP: short oligonucleotide alignment program," *Bioinformatics*, vol. 24, no. 5, pp. 713–714, 2008.

[15] G. Rizk and D. Lavenier, "GASSST: global alignment short sequence search tool," *Bioinformatics*, vol. 26, no. 20, pp. 2534–2540, 2010.

[16] A. D. Smith, Z. Xuan, and M. Q. Zhang, "Using quality scores and longer reads improves accuracy of Solexa read mapping," *BMC Bioinformatics*, vol. 9, article 128, 2008.

[17] W. Zhang, J. Chen, Y. Yang, Y. Tang, J. Shang, and B. Shen, "A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies," *PLoS ONE*, vol. 6, no. 3, Article ID e17915, 2011.

[18] B. D. Ondov, A. Varadarajan, K. D. Passalacqua, and N. H. Bergman, "Efficient mapping of applied biosystems SOLiD sequence data to a reference genome for functional genomic applications," *Bioinformatics*, vol. 24, no. 23, pp. 2776–2777, 2008.

[19] S. Hoffmann, C. Otto, S. Kurtz et al., "Fast mapping of short sequences with mismatches, insertions and deletions using index structures," *PLoS Computational Biology*, vol. 5, no. 9, Article ID e1000502, 2009.

[20] P. Flicek and E. Birney, "Sense from sequence reads: methods for alignment and assembly," *Nature Methods*, vol. 6, no. 11, supplement, pp. S6–S12, 2009.

[21] M. Farrar, "Striped Smith-Waterman speeds database searches six times over other SIMD implementations," *Bioinformatics*, vol. 23, no. 2, pp. 156–161, 2007.

[22] H. Jiang and W. H. Wong, "SeqMap: mapping massive amount of oligonucleotides to the genome," *Bioinformatics*, vol. 24, no. 20, pp. 2395–2396, 2008.

[23] D. Campagna, A. Albiero, A. Bilardi et al., "PASS: a program to align short sequences," *Bioinformatics*, vol. 25, no. 7, pp. 967–968, 2009.

[24] D. Weese, A. Emde, T. Rausch, A. Döring, and K. Reinert, "RazerS—fast read mapping with sensitivity control," *Genome Research*, vol. 19, no. 9, pp. 1646–1654, 2009.

[25] D. Weese, M. Holtgrewe, and K. Reinert, "RazerS 3: faster, fully sensitive read mapping," *Bioinformatics*, vol. 28, no. 20, pp. 2592–2599, 2012.

[26] C. Alkan, J. M. Kidd, T. Marques-Bonet et al., "Personalized copy number and segmental duplication maps using next-generation sequencing," *Nature Genetics*, vol. 41, no. 10, pp. 1061–1067, 2009.

[27] F. Hach, F. Hormozdiari, C. Alkan et al., "MrsFAST: a cache-oblivious algorithm for short-read mapping," *Nature Methods*, vol. 7, no. 8, pp. 576–577, 2010.

[28] K. Schneeberger, J. Hagmann, S. Ossowski et al., "Simultaneous alignment of short reads against multiple genomes," *Genome Biology*, vol. 10, no. 9, article R98, 2009.

[29] C. Trapnell and S. L. Salzberg, "How to map billions of short reads onto genomes," *Nature Biotechnology*, vol. 27, no. 5, pp. 455–457, 2009.

[30] S. M. Rumble, P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, and M. Brudno, "SHRiMP: accurate mapping of short color-space reads," *PLoS Computational Biology*, vol. 5, no. 5, Article ID e1000386, 2009.

[31] M. David, M. Dzamba, D. Lister, L. Ilie, and M. Brudno, "SHRiMP2: sensitive yet practical short read mapping," *Bioinformatics*, vol. 27, no. 7, pp. 1011–1012, 2011.

[32] S. Schbath, V. Martin, M. Zytnicki, J. Fayolle, V. Loux, and J. F. Gibrat, "Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis," *Journal of Computational Biology*, vol. 19, no. 6, pp. 796–813, 2012.

[33] H. Lin, Z. Zhang, M. Q. Zhang, B. Ma, and M. Li, "ZOOM! zillions of oligos mapped," *Bioinformatics*, vol. 24, no. 21, pp. 2431–2437, 2008.

[34] N. Homer, B. Merriman, and S. F. Nelson, "BFAST: an alignment tool for large scale genome resequencing," *PLoS ONE*, vol. 4, no. 11, Article ID e7767, 2009.

[35] S. Descorps-Declère, D. Ziébelin, F. Rechenmann, and A. Viari, "Genepi: a blackboard framework for genome annotation," *BMC Bioinformatics*, vol. 7, article 450, 2006.

[36] K. Daily, P. Rigor, S. Christley, X. Xie, and P. Baldi, "Data structures and compression algorithms for high-throughput sequencing technologies," *BMC Bioinformatics*, vol. 11, article 514, 2010.

[37] 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.

[38] T. W. Lam, W. K. Sung, S. L. Tam, C. K. Wong, and S. M. Yiu, "Compressed indexing and local alignment of DNA," *Bioinformatics*, vol. 24, no. 6, pp. 791–797, 2008.

[39] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 26, no. 5, pp. 589–595, 2010.

[40] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[41] N. L. Clement, Q. Snell, M. J. Clement et al., "The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing," *Bioinformatics*, vol. 26, no. 1, pp. 38–45, 2009.

[42] Z. Ning, A. J. Cox, and J. C. Mullikin, "SSAHA: a fast search method for large DNA databases," *Genome Research*, vol. 11, no. 10, pp. 1725–1729, 2001.

[43] S. Marguerat and J. Bähler, "RNA-seq: from technology to biology," *Cellular and Molecular Life Sciences*, vol. 67, no. 4, pp. 569–579, 2010.

[44] C. M. Liu, T. Wong, E. Wu et al., "SOAP3: ultra-fast GPU-based parallel alignment tool for short reads," *Bioinformatics*, vol. 28, no. 6, pp. 878–879, 2012.

[45] M. C. Schatz, "CloudBurst: highly sensitive read mapping with MapReduce," *Bioinformatics*, vol. 25, no. 11, pp. 1363–1369, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 762126, 10 pages
http://dx.doi.org/10.1155/2014/762126

Research Article

Data Analysis and Tissue Type Assignment for Glioblastoma Multiforme

Yuqian Li,[1] Yiming Pi,[1] Xin Liu,[1] Yuhan Liu,[1] and Sofie Van Cauter[2]

[1] *School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China*
[2] *Department of Radiology and Department of Imaging and Pathology, University Hospitals of Leuven, 3001 Leuven, Belgium*

Correspondence should be addressed to Yuqian Li; yuqianli@uestc.edu.cn

Received 18 November 2013; Revised 13 January 2014; Accepted 23 January 2014; Published 3 March 2014

Academic Editor: Bairong Shen

Glioblastoma multiforme (GBM) is characterized by high infiltration. The interpretation of MRSI data, especially for GBMs, is still challenging. Unsupervised methods based on NMF by Li et al. (2013, *NMR in Biomedicine*) and Li et al. (2013, *IEEE Transactions on Biomedical Engineering*) have been proposed for glioma recognition, but the tissue types is still not well interpreted. As an extension of the previous work, a tissue type assignment method is proposed for GBMs based on the analysis of MRSI data and tissue distribution information. The tissue type assignment method uses the values from the distribution maps of all three tissue types to interpret all the information in one new map and color encodes each voxel to indicate the tissue type. Experiments carried out on *in vivo* MRSI data show the feasibility of the proposed method. This method provides an efficient way for GBM tissue type assignment and helps to display information of MRSI data in a way that is easy to interpret.

1. Introduction

Glioblastoma multiforme (GBM), which typically consists of three tissue types (i.e., normal, tumor, and necrosis), is a type of extensively heterogeneous tumors. Accurate diagnosis of GBMs is of great importance for guiding therapy and planning operations. Being different from other brain tumors which present similar spectral patterns, GBMs are characterized by high infiltration [1, 2]. Such characteristic brings huge difficulty in tumor typing and diagnosis.

Magnetic resonance spectroscopy imaging (MRSI) [3] is a very useful noninvasive tool for brain tumor diagnosis, especially for highly heterogeneous tumors like GBMs. Unlike magnetic resonance imaging (MRI) which only shows the brain structure, MRSI combines MRI and magnetic resonance spectroscopy (MRS) [4] to provide the localized biochemical information. By investigating the spectra from multivoxels, the clinicians could have a better insight into the pathological change of brain tissues.

However, the interpretation of MRSI data is still challenging which hinders its application in tumor diagnosis. Efforts for exploiting MRSI data have been made using both supervised and unsupervised methods. Nosologic imaging is created using linear discriminant analysis [5, 6], canonical correlation analysis (CCA) [7, 8], Bayesian frameworks [9, 10], and nonnegative matrix factorization (NMF) [11]. NMF [11] is an alternative blind source separation technique with only nonnegative constraint. It has shown great potentials in brain tissue differentiation [2, 12–14]. In our previous work, we proposed an unsupervised method, namely, hierarchical nonnegative matrix factorization (hNMF), to interpret the MRSI data for GBMs without prior knowledge and provided an easy way to interpret MRSI data of GBMs for each tissue type [15].

Unlike the supervised classification methods, which labels each voxel based on large training sets [5–10], tissue typing for NMF tissue differentiation is not usually considered [12, 13, 16]. Recently, a tissue typing method was carried out by simply exploring which tissue contributes most to the voxel [14]. Such an approach ignored the voxels with intensively mixed tissues, that is, the different tissues contributing fairly equal. We tried to integrate the distribution information of each pure tissue in one image by encoding each of them as

a color channel [16]. The obtained images, known as nosologic images, showed the spatial distribution of all tissue types. However, the tissue distribution is only shown in shading colors and the tissue type of each voxel is not indicated clearly.

In this paper, we improved upon [15] by proposing an approach for GBM tissue type recognition. The previous work is extended by analyzing both the pure and mixed data labeled by an expert. The spectral data labeled as each tissue type is analyzed and the relationship of different tissue types is studied. Then, we proposed criteria to assign each voxel to a certain tissue type (i.e., pure tissue normal, tumor, necrosis, mixed tissues normal/tumor, or tumor/necrosis, hereafter noted as "C", "T", "N", "C/T", and "T/N," resp.) using the tissue distribution maps. *In vivo* experiments are performed using short-TE ^{1}H MRSI data from GBM patients. We then evaluate its performance using the expert labeling.

2. Materials

2.1. Data Acquisition Protocol. The MRSI protocol had the same imaging parameters as in our previous work [15, 16]. All the MRSI data were acquired at the University Hospital of Leuven (UZ Leuven, Belgium) on a 3 T MR scanner (Achieva, Philips, Best, The Netherlands). A body coil for transmission and eight-channel head coil for signal reception were used. The MRSI protocol had the following imaging parameters: point-resolved spectroscopy (PRESS) [17] that was used as the volume selection technique; TR/TE = 2000/35 ms; field of view, 16 cm × 16 cm; volume of interest, 8 cm × 8 cm (maximum size); slice thickness, 1 cm; acquisition voxel size, 1 cm × 1 cm; reconstruction voxel size, 0.5 cm × 0.5 cm; receiver bandwidth, 2000 Hz; samples, 2048; number of signal averages, 1; water suppression method, MOIST; shimming, pencil beam shimming; first- and second-order parallel imaging with SENSE factor: left-right, 2; anterior-posterior, 1.8; 10 circular 30 mm outer-volume saturation bands in order to avoid lipid contamination from the skull. Standard anatomical MR images were also acquired.

2.2. Patients and Data. MRSI data sets from 6 GBM patients (typically present three tissue patterns, i.e., normal, tumor, and necrosis) were selected for this study. The MRSI data was acquired prior to any treatment from 6 patients with brain tumors that were subsequently diagnosed as GBM based on histological examination and followed the rules of the World Health Organization (WHO) classification for tumor grading [18]. The institutional review board approved the study. Written informed consent was obtained from all patients before their participation in the study. Data preprocessing was done as in our previous papers [15, 16] using the in-house software SPID [19].

2.3. Expert Labeling. MR spectra were judged by a spectroscopist (a radiologist with five years of experience). The expert spectroscopist was presented with the real spectra in a range from 4.3 to 0 ppm.

Firstly, spectral quality assessment was performed as recommended by Kreis [20]. Spectra were judged acceptable if the following criteria were met: FWHM of metabolites < 0.07–0.1 ppm, no unexplained features in the residuals, no doubled peaks or evidence for movement artifacts, symmetric lineshape, and no outer-volume ghosts or other artifacts present.

Afterwards, the spectra with acceptable quality were assigned to different tissue classes: normal appearing brain parenchyma, tumoral tissue, or necrosis, based on the spectra and the corresponding T1-weighted image after contrast administration.

3. Method

3.1. Spectra Investigation for GBMs Using Biomarkers. N-acetylaspartate (NAA), choline (Cho), and lipids are known to be the three most important biomarkers for investigating brain tumorigenesis. The concentration of these metabolites changes under disease condition. In the context of GBM spectroscopy, necrosis mostly contains lipids. NAA concentration is higher than Cho in normal tissue and gliomas are characterized by decreased NAA and increased Cho and lipids. But these biomarkers are not enough for MRSI spectra differentiation. In a specific frequency region of a spectrum, the peak height of the metabolite can be measured. Here, we use the NAA-to-Cho index (NCI) and NAA-to-Lips index (NLI), which measure the ratios of the peak heights of these components, to investigate the spectra for all GBM patients. We select all voxels containing pure tissues and mixed tissues to observe if the biomarkers are capable of clustering the same tissues. Each point represents a spectrum from a single voxel. Its coordinate values (x, y) correspond to the NCI value and NLI value, respectively. The points are colored using expert labeling to indicate their tissue types, blue for "C," cyan for "C/T," green for "T," yellow for "T/N," and red for "N. "

3.2. Spectra Variation Investigation Using Expert Labeling. In this section, we investigate the relationship of spectral variation and expert labeling, including two aspects: (1) the variation of spectra labeled as the same tissue type and (2) the variation of spectral difference between two tissue types.

The expert labeled spectrum in each voxel as a certain tissue type "C," "T," "N," "C/T," and "T/N." However, because of the voxel size of MRSI data and the infiltration property of GBMs, there is no clear boundary between different tissue types, especially in the area of tumor proliferation. Therefore, the spectra labeled as the same tissue type could have different profiles. In order to investigate the spectra variation, we plot all spectra of each pure tissue type and their mean spectra.

The correlation coefficients between normal and tumor spectra R^{CT} and the correlation coefficients between tumor and necrotic spectra R^{TN} can evaluate the spectral difference of different tissue types. For each spectrum labeled as a pure tissue type, we calculate the correlation coefficient of this spectrum and a spectrum labeled as another pure tissue type. With box plots, the variation of spectral difference between two different tissue types could be observed easily. Combined data of all patients and also that of individual patients are both analyzed to investigate the spectral difference between different tissue types.

3.3. Tissue Differentiation with hNMF. Spectra from a MRSI grid can be approximated as a linear combination of r constituent spectra. We define a data matrix X containing all spectra from the voxels of interest (VOI). Each column of matrix X represents a spectrum from one voxel. With conventional NMF, X can be factorized into a new nonnegative matrix W (each row represents a constituent spectrum of normal tissue or necrosis) and a new nonnegative matrix H,

$$X_{m\times n} \approx W_{m\times r} H_{r\times n} \quad \text{subject to } W, H \geq 0. \tag{1}$$

The reshape of each row of H, hereafter called "h-map," that is, the "tissue distribution maps" we mentioned before, gives the spatial distribution of the corresponding spectrum.

For the low grade gliomas, conventional NMF is able to differentiate normal and abnormal (i.e., tumor) tissues. While there are more than two tissue types (e.g., GBMs), the conventional NMF sometimes fails to recover the biomeaningful constituent spectra robustly. Therefore, a hierarchical approach based on NMF (i.e., hNMF) was proposed to recover the spectra of MRSI data for GBMs which contains 3 constituent spectra [15]. HNMF firstly differentiates the data matrix into normal and abnormal. Then, with an optimized threshold, the abnormal part is further differentiated into tumor and necrosis. As a result, the three constituent spectra of normal, tumor, and necrosis are recovered and their h-maps for different tissue types are obtained simultaneously. Note that, in each voxel, there are 3 values from the 3 h-maps h_i^{C}, h_i^{T}, and h_i^{N} for each tissue type.

3.4. h-Map Investigation Using Expert Labeling. As introduced in the previous section, the h-maps, which are normalized between 0 and 1, can be obtained from the result of hNMF. Then, the h-map of each tissue type represents the tissue distribution using a number for a voxel. However, during expert labeling, each voxel is arbitrarily labeled as a certain tissue type instead of a number. It is obvious that the h-values (hereafter, the value from a single voxel in an h-map referred to as "h-value") from the voxels, which are labeled as the same tissue type, could be different. In this section, we will exploit the h-values of each tissue type. H-values of all patients and each patient are also exploited to reveal the extent of individual difference.

3.5. Tissue Type Assignment. Based on the data analysis of h-maps, each voxel could be assigned to a certain tissue type. In each voxel, there are 3 h-values from 3 h-maps for "C," "T," and "N," respectively. Obviously, the h-values of "C" (i.e., h_i^{C}) should be bigger than the h-values of "T" (i.e., h_i^{T}) and the h-values of "N" (i.e., h_i^{N}) for normal voxels. Analogically, for voxels of tumor and necrosis tissue, h_i^{T} and h_i^{N} should overwhelm h-values of other tissue types, respectively.

However, there are mixed tissues where h-values of each tissue type vary significantly and thus the above criteria cannot be simply used to decide the tissue type. To properly separate the different tissues from the mixed tissues, a parameter ρ should be added; that is, h_i^{C} should be bigger than $h_i^{\mathrm{T}} + \rho^{\mathrm{CT}}$ for the voxel to be assigned to be "C." Similarly, h_i^{N} should be bigger than $h_i^{\mathrm{T}} + \rho^{\mathrm{TN}}$ for the voxel to be assigned to be "N." Therefore, we make the following criteria for tissue type assignment.

The Rules for Tissue Type Assignment

While $h_i^{\mathrm{C}} > h_i^{\mathrm{T}}$, $h_i^{\mathrm{C}} > h_i^{\mathrm{N}}$, and $h_i^{\mathrm{C}} > h_i^{\mathrm{T}} + \rho^{\mathrm{CT}}$, assign the voxel to be "C";

While $h_i^{\mathrm{N}} > h_i^{\mathrm{T}}$, $h_i^{\mathrm{N}} > h_i^{\mathrm{C}}$, and $h_i^{\mathrm{N}} > h_i^{\mathrm{T}} + \rho^{\mathrm{TN}}$, assign the voxel to be "N";

While $h_i^{\mathrm{T}} > h_i^{\mathrm{C}} + \rho^{\mathrm{CT}}$ and $h_i^{\mathrm{T}} > h_i^{\mathrm{N}} + \rho^{\mathrm{TN}}$, assign the voxel to be "T";

Else if $h_i^{\mathrm{N}} < h_i^{\mathrm{C}}$ and $h_i^{\mathrm{N}} < h_i^{\mathrm{T}}$, assign the voxel to be "C/T";

Else if $h_i^{\mathrm{C}} < h_i^{\mathrm{T}}$ and $h_i^{\mathrm{C}} < h_i^{\mathrm{N}}$, assign the voxel to be "T/N."

According to the above criteria, we can have all the voxels assigned to a certain type, including the ones originally labeled as "B" by expert.

3.6. Validation. The efficacy of the proposed tissue type assignment approach is validated using expert labeling information. The computed tissue type of each voxel is compared with the tissue type labeled by expert. We use the correct rate, false alarm rate, and the omission rate to evaluate the performance of the proposed approach.

Correct rate describes correct assignment among all the assignment,

$$\text{Correct rate} = \frac{N_{\text{correct}}}{N_{\text{assigned}}}, \tag{2}$$

where N_{assigned} represents the number of voxels which are assigned to a certain tissue type using the proposed method. And N_{correct} represents the number of voxels assigned to a certain tissue type that our assignment is the same as that of an expert.

False alarm rate describes the wrong assignments which should not be counted,

$$\text{False alarm rate} = \frac{N_{\text{error}}}{N_{\text{assigned}}}, \tag{3}$$

where N_{error} represents the number of voxels which are assigned to be a certain tissue type using the proposed approach but not labeled by an expert as the same tissue type.

Omission rate describes the wrong assignment which is missed,

$$\text{Omission rate} = \frac{N_{\text{omission}}}{N_{\text{assigned}}}, \tag{4}$$

where N_{omission} represents the number of voxels which are labeled by an expert to be a certain tissue type but not assigned as the same tissue type using the proposed approach.

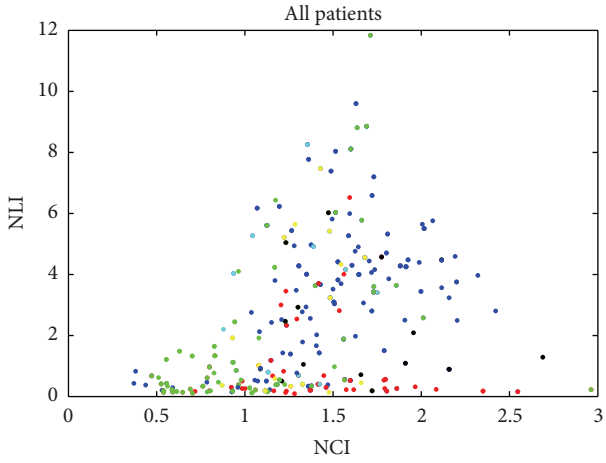


Figure 1: Investigation of *in vivo* ^{1}H MRSI data from GBM patients. x-axis is the NAA-to-Cho index (NCI) and y-axis is the NAA-to-Lips index (NLI). Blue, green, and red points indicate normal, tumor, and necrotic tissue, respectively. Mixed colors represent mixed tissue. Blue for "C," cyan for "C/T," green for "T," yellow for "T/N," and red for "N."

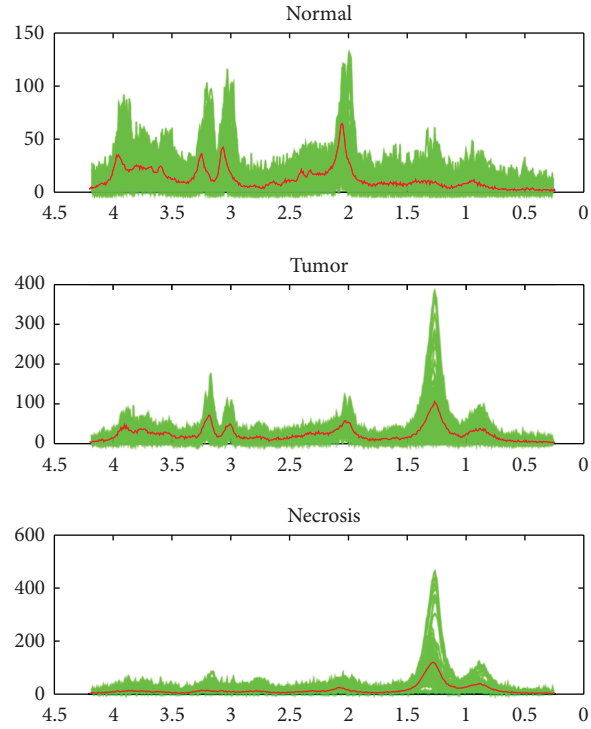


Figure 2: Spectral variation of pure tissues. Each green spectrum is from a voxel. The red bold line represents the mean spectrum.

4. Results and Discussion

4.1. Spectra Investigation for GBMs Using Biomarkers. We investigated all the 6 data sets which were pathologically confirmed to be GBM by clinicians. Figure 1 shows the overall tissue types of all the GBM data sets. Each point represents a spectrum from one voxel among all the data sets. The points are colored using expert labeling, same color for same tissue type. Due to the variation of the spectra, the distribution map shows serious overlap between tissue types. Though there are two vaguely centralized clusters for normal (higher NLI and NCI) and necrosis (very low NLI and lower NCI), there are no clear dividing lines between tissue types. Tumor cannot be separated from normal and necrosis. Mixed tissues cannot be differentiated from other tissue types, either.

4.2. Spectra Variation Investigation Using Expert Labeling. The spectral variation of pure tissues labeled by an expert is investigated by plotting all spectra of the same tissue from all GBM patients in one figure. As shown in Figure 2, the green spectra are from all the voxels labeled as normal, tumor, and necrosis by an expert. Serious spectral variations for the same tissue type can be observed. It demonstrates that spectra for the same tissue type are possibly not identical. The red bold line plots the mean spectrum for normal, tumor, and necrosis, respectively. We can observe that most of the green spectra have great difference with the mean spectra.

The spectral relationships of different tissue types are investigated using correlation coefficients. The correlation coefficients of each spectrum labeled as "C" by expert and the spectrum labeled as "T" by expert, noted as R^{CT}, are calculated to investigate the difference of normal spectra and tumor spectra and its variation. Figure 3(a) shows the R^{CT} for all the GBM patients and each individual patient. As shown, most of the correlation coefficients R^{CT} are between 0.3 and 0.7. However, some values are extremely small or big because of the variation of tumor spectra. Differences between patients are not significant except for two patients, that is, patients 4 and 6. It demonstrates that the serious variation among patients is not common but possible. The lower quartile of R^{CT}, $Q1^{CT} = 0.2167$, could be used to describe the relationship between normal and tumor.

Similarly, the correlation coefficients of each spectrum labeled as "T" by expert and the spectrum labeled as "N" by expert, noted as R^{TN}, are calculated to investigate the difference of tumor spectra and necrotic spectra and its variation, as shown in Figure 3(b). Compared to R^{CT}, the variation of R^{TN} is more serious. However, the R^{TN} values of all patients inside the box are between 0.3 and 0.8. The lower quartile of R^{TN}, $Q1^{TN} = 0.2950$, could be used to describe the relationship between tumor and necrosis.

4.3. h-Maps Variation for Different Tissue Types. The values in h-maps for each labeled specific tissue type are analyzed. Figure 4 gives the h-values from the 6 GBM patients. For the 6 data sets, there are 6 normal h-maps, 6 tumor h-maps, and 6 necrosis h-maps. For h-maps of each tissue type, we analyzed the data distribution of tissue types for all patients.

Figure 4 illustrates the h-values of normal h-map. Each plot contains a box for all patients and 6 boxes for 6 GBM patients. Figure 4(a) depicts the h-values taken from h-maps of normal tissue. The values from voxels labeled as "C," "T," "N," "C/T," and "T /N" are depicted. As shown, the values for "C" are mostly between 0.6 and 1. The values for "N" and

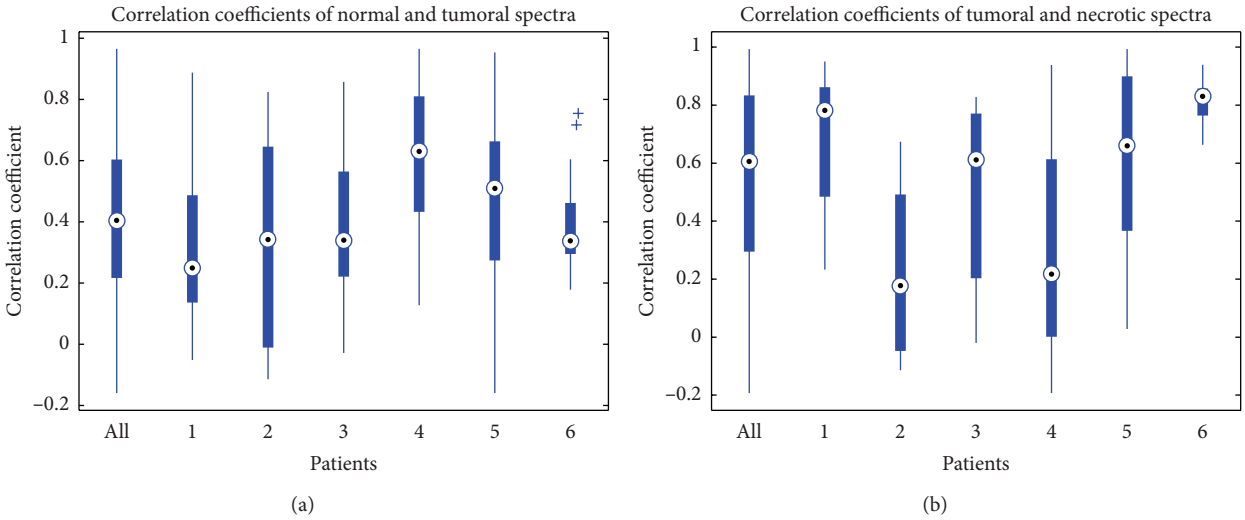


Figure 3: Correlation coefficients of different spectra.

"T/N" are all quite small. It implies that good separation of normal and necrosis is possible using h-maps. For the other two types "T" and "T/N", the values vary greatly. Figure 4(b) depicts the h-values taken from tumor h-maps. As shown, values for all tissue types vary seriously, even for tumor. Figure 4(c) depicts the h-values taken from necrosis h-maps. The values for "N" are mostly between 0.5 and 0.9. The values for "C" and "C/T" are quite small. It also implies that good separation of normal and necrosis is possible using h-maps. The values for "T" and "T/N" vary greatly. In general, the h-values taken from tumor h-maps vary more seriously than the h-values taken from h-maps of normal and necrosis, and the values for "T," "C/T," and "T/N" taken from all three h-maps vary significantly. It implies that the separation of tumor and mixed tissue is more difficult than normal and necrosis.

4.4. Tissue Type Assignment for GBMs. The proposed tissue type assignment method described in Section 3.5 is applied to the h-maps of 6 GBM data sets. $Q1^{CT} = 0.2167$ and $Q1^{TN} = 0.2950$ are used as ρ^{CT} and ρ^{TN}, respectively. The results are compared with expert labeling information. For both the results and the expert labeling, the distribution map is color-coded blue for "C," cyan for "C/T," green for "T," yellow for "T/N," red for "N," and black for "B" which is spectra of low quality of which the tissue type cannot be decided by expert. As shown in Figure 5, the assigned tissue types are approximately in accordance with the expert labeling. The regions of normal and necrosis are more accurate than the regions of tumor and mixed tissues like "C/T" and "T/N." This is mainly because the high infiltration character of gliomas brings higher variation to the spectral profiles of tumor and mixed tissues. The black voxels labeled as "B" by expert can be estimated using the proposed method. After analyzing localization of these voxels and their surrounding voxels, the assignment of these voxels was confirmed to be correct.

4.5. Validation. For each patient, the correct rate, false alarm rate, and the omission rate are calculated for each pure tissue and mixed tissue types by comparing results with expert labeling. The "N/A" in Table 1 represents the situations which do not exist.

As we observe, the assignments of pure tissue "C" and "N" are almost always more accurate than "T." The correct rate of "C" and "N" can be as good as above 0.9 or even 1. The correct rate of mixed tissues (i.e., C/T and T/N) is lower than pure tissues.

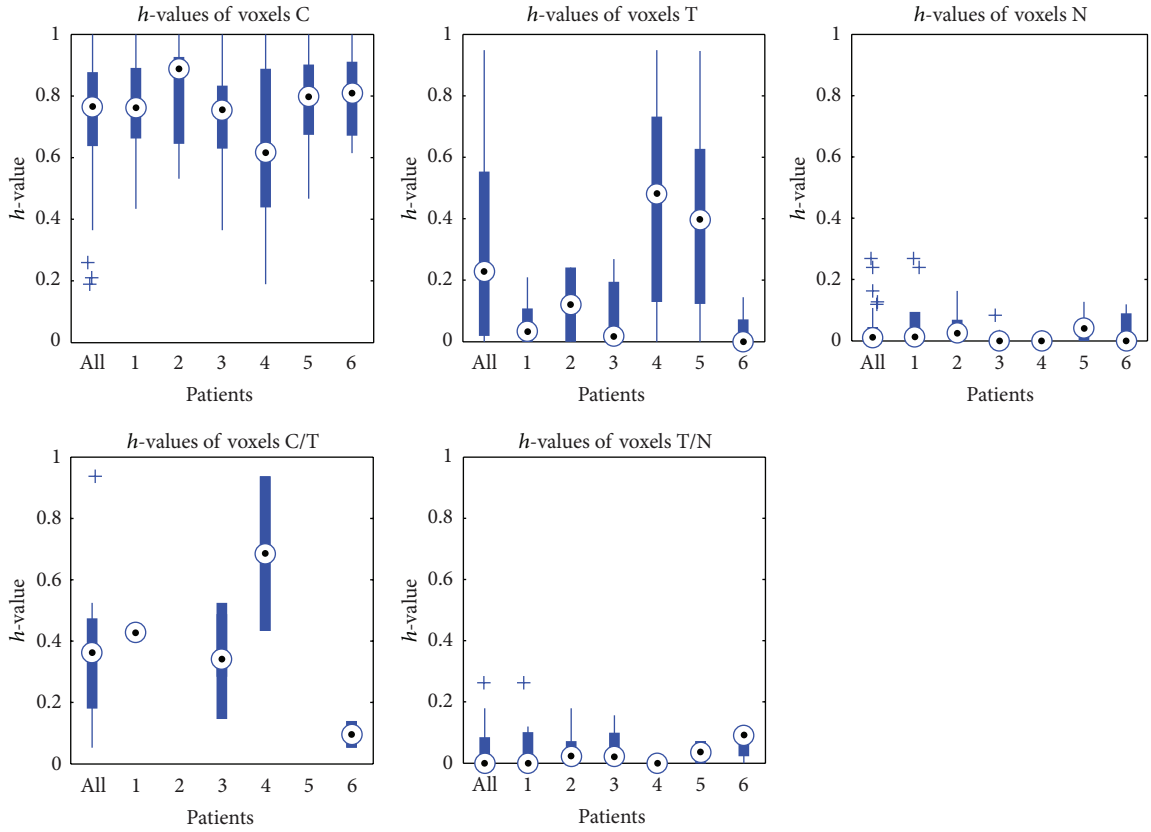
The omission rates of the pure tissue "C" and "N" for all patients are lower than 0.5, mostly lower than 0.4. But for "T" and mixed tissue "C/T," "T/N," the omission rate is higher.

For all results, the pure tissues "C" and "N" perform better than "T" and "T" performs better than the mixed tissues "C/T" and "T/N." Inaccurate assignment of a tissue type influences the assignment of the tissues near it. In other words, the correct rate or error rate of "C," "T," and "N" will be affected by the inaccurate assignment of "C/T" and "T/N."

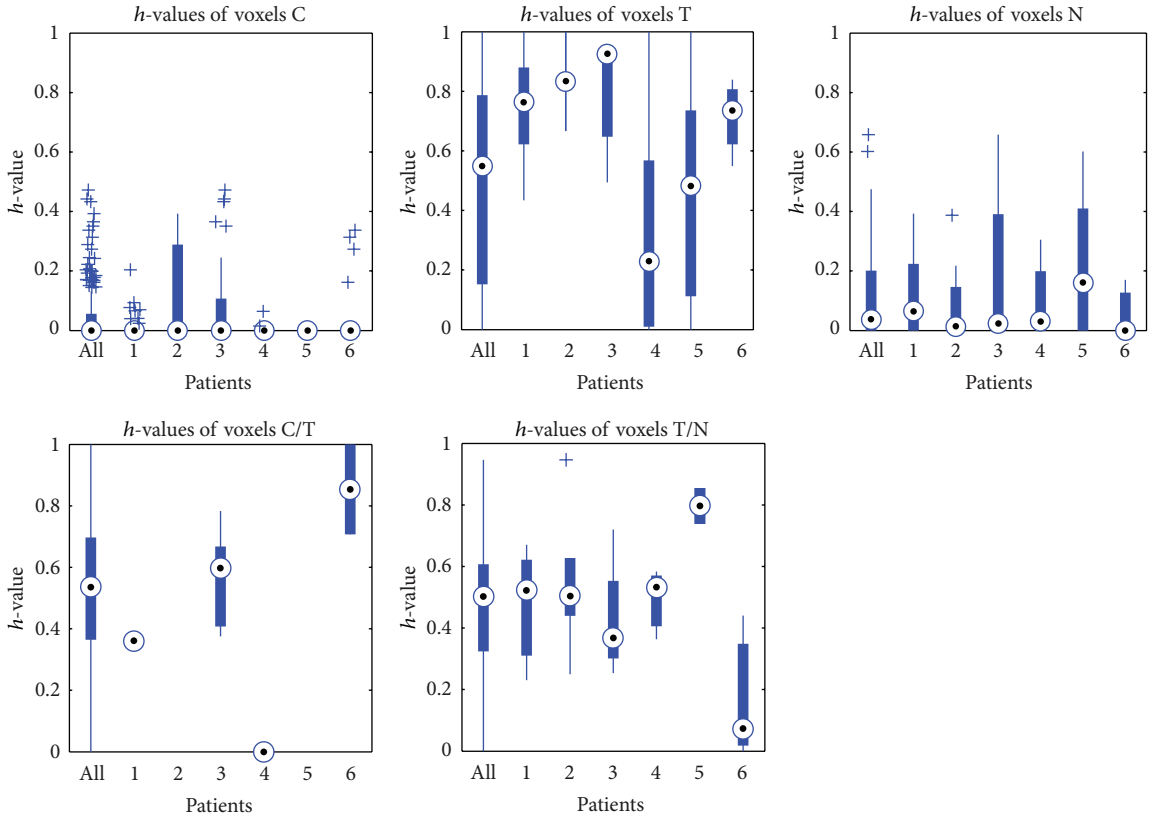
5. Discussions

This study continued with our tissue typing work using hNMF [15]. We explored the possibility of only using several most representative biomarkers for tissue differentiation. The results showed that the different tissue types cannot be well separated, especially for tumor and mixed tissues. Therefore, a new approach for tissue type assignment using hNMF is developed.

Then we evaluated the relationship between spectra of different tissue types. The spectra labeled as a certain tissue type by expert are compared to the spectra labeled as another tissue type. The variation of the different correlation coefficients for both intra- and interpatient indicates the difference of spectra which are labeled as the same tissue type. This implies that the spectra are not identical even if they are labeled as the same tissue type, especially for tumoral spectra. This could be due to the fact that glioblastoma are known to be very heterogeneous lesions. Invasion, regions of increased cellularity, necrosis on a microscopic and a macroscopic scale,

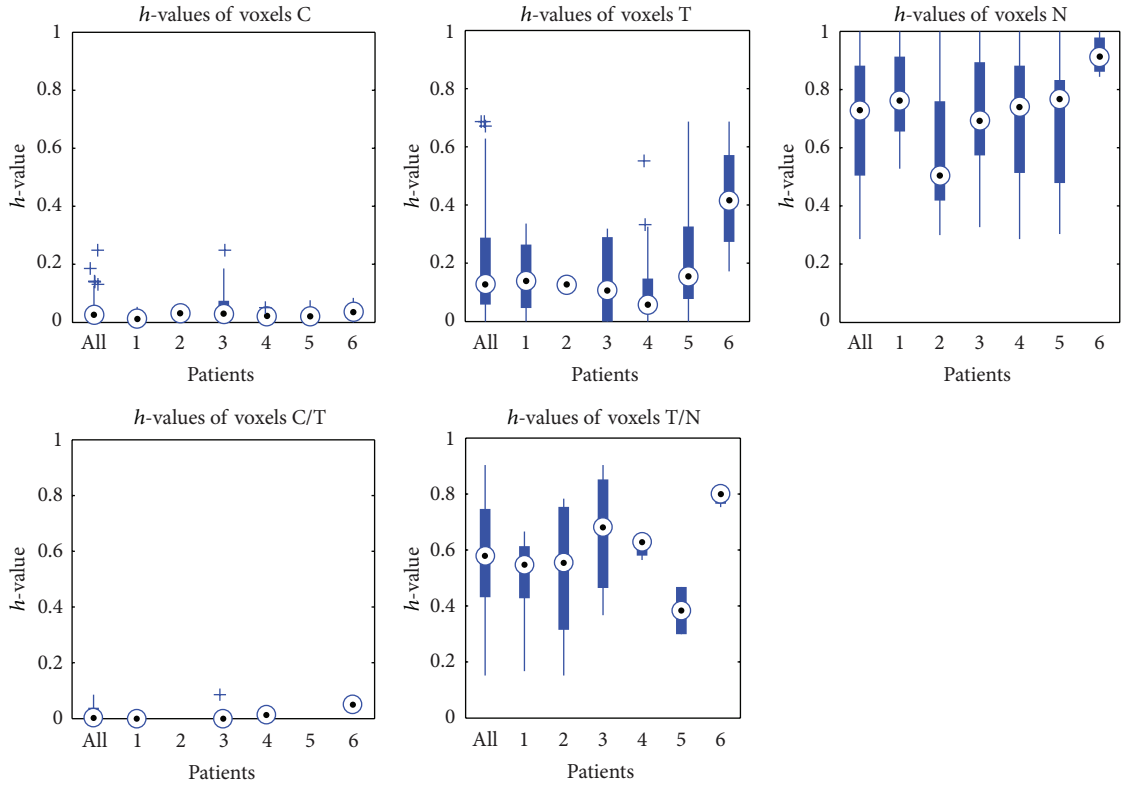


(a) h-value variations of normal h-maps



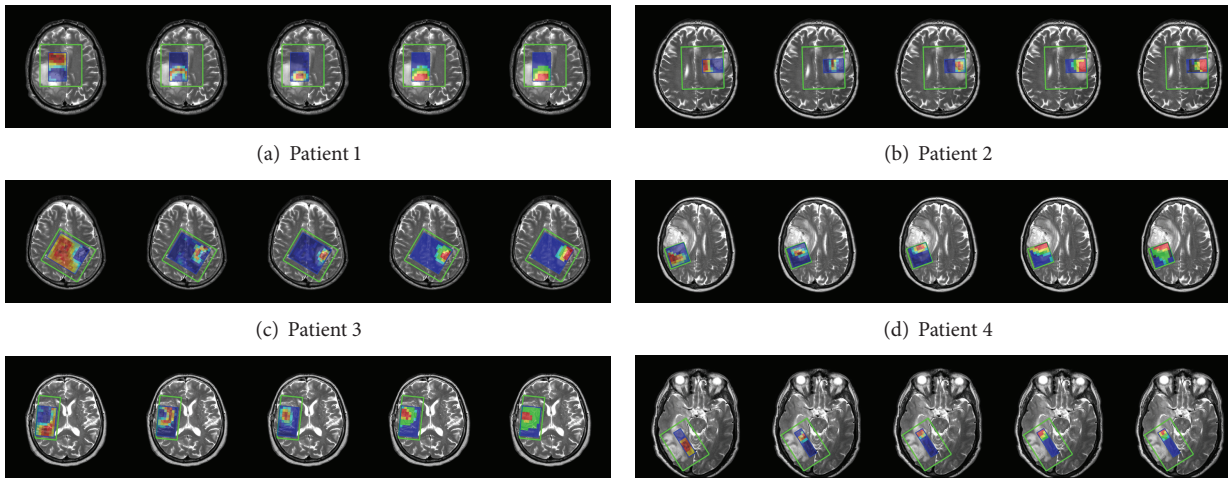
(b) h-values of tumor h-maps

FIGURE 4: Continued.



(c) h-values of necrotic h-maps

FIGURE 4: Variations of h-values of intra- and interpatients.



(a) Patient 1 (b) Patient 2

(c) Patient 3 (d) Patient 4

(e) Patient 5 (f) Patient 6

FIGURE 5: h-maps and tissue type assignment results compared with expert labeling. All the results are overlaid on the T2-weighted MRI. For each patient, the first 3 images are the h-maps for normal, tumor, and necrosis. The fourth image is the assigned tissue types. The last image is the expert labeling.

hemorrhage, and microvascular proliferation are hallmarks of the most malignant of gliomas. This heterogeneity is reflected in the variation of the spectra, related to tumoral tissue. A voxel in the chemical shift imaging (CSI) protocol used in this study is approximately 0.25 cm^3. Thus, thousands of metabolites will contribute to the measured signal. The spectra in MRS are only indirect indicators of metabolism. For example, regions of tumoral tissue are characterized by high cellularity and are perceived as spectra with strongly elevated choline and decreased NAA. Regions with tumoral tissue with necrosis on a microscopic scale will be perceived with moderately elevated lipids and lower values of choline

Table 1: Result validation.

	C	T	N	C/T	T/N
Patient 1					
Detected number	38	12	13	5	9
Correct detected number	38	11	12	0	6
Number of voxels labeled by expert	38	19	12	0	8
Correct rate	1	0.9167	0.9231	0	0.6667
Error rate/false alarm rate	0	0.0833	0.0769	1	0.3333
Omission rate	0	0.4210	0	N/A	0.2500
Patient 2					
Detected number	14	4	13	2	7
Correct detected number	12	3	13	1	5
Number of voxels labeled by expert	13	3	14	0	7
Correct rate	0.8571	0.7500	0.9231	0.5000	0.7142
Error rate/false alarm rate	0.1429	0.2500	0.0769	0.5000	0.2857
Omission rate	0.0769	0	0.1429	N/A	0.2857
Patient 3					
Detected number	108	13	9	9	4
Correct detected number	108	5	7	3	2
Number of voxels labeled by expert	115	5	11	6	6
Correct rate	1	0.3846	0.7778	0.3333	0.5000
Error rate/false alarm rate	0	0.6153	0.2222	0.6667	0.5000
Omission rate	0.0608	0	0.3636	0.5000	0.6667
Patient 4					
Detected number	27	4	9	7	9
Correct detected number	14	4	9	1	4
Number of voxels labeled by expert	16	21	11	4	4
Correct rate	0.5185	1	1	0.1429	0.4444
Error rate/false alarm rate	0.4815	0	0	0.8571	0.5556
Omission rate	0.1250	0.8095	0.1818	0.7500	0
Patient 5					
Detected number	38	19	10	4	6
Correct detected number	20	17	8	0	0
Number of voxels labeled by expert	20	44	11	0	2
Correct rate	0.5263	0.8947	0.8000	0	0
Error rate/false alarm rate	0.4737	0.1053	0.2000	1	1
Omission rate	0	0.6136	0.2727	N/A	1
Patient 6					
Detected number	21	5	6	0	1
Correct detected number	21	3	3	0	0
Number of voxels labeled by expert	21	4	3	2	3
Correct rate	1	0.6000	0.5000	N/A	0
Error rate/false alarm rate	0	0.4000	0.5000	N/A	1
Omission rate	0	0.2500	0	1	1

and NAA. Tumoral regions with moderately elevated cellularity will be perceived as regions with only moderately elevated Cho and moderately lowered NAA. In the end, these spectra represent all tumoral tissue, as designated by the histopathologist as well as by the expert labeling in MR spectroscopy [21–25]. Therefore, spectral variation within the same tissue type, which is introduced by the nature of tumor proliferation and the volume of CSI voxels, could happen and influence the performance of tissue typing method.

As demonstrated, the lower quartiles of correlation coefficients $Q1^{CT}$ and $Q1^{TN}$ could imply the "least spectral similarity" between different tissue types. Additionally, the scale of correlation coefficients R^{CT} and R^{TN} is in the same scale of h-values. Therefore, in the tissue typing assignment experiment, where $Q1^{CT} = 0.2167$ and $Q1^{TN} = 0.2950$, which were calculated using 6 GBM patients, were used as the parameters ρ^{CT} and ρ^{TN}, respectively. Though the patients were few, the value of $Q1^{CT}$ and $Q1^{TN}$ will not change significantly since the voxel number for calculating them is large enough to be stable. Another important point we must stress is that, as long as we have decided the value for parameters ρ^{CT} and ρ^{TN}, we do not need to calculate

$Q1^{CT}$ and $Q1^{TN}$ every time there is a new patient. The tissue assignment method is still automatic since the values used as ρ^{CT} and ρ^{TN} will be fixed numbers. Here, we just proposed a potential way to decide ρ^{CT} and ρ^{TN}.

As the tissue type assignment is based on the h-maps of hNMF, the results could be affected by both the h-maps and the typing criteria. As shown, the results for tumor and mixed tissues are worse than the results for normal and necrosis. On the one hand, there is the serious variation of tumor spectra. On the other hand, the spectral profile of C/T and T/N is highly correlated with the tumor spectra. These facts lower the typing results. However, the assignments of each tissue type shown in Section 4.4 have shown the efficacy of the proposed method.

6. Conclusions

In this paper, we investigate the spectra variation with expert's labeling. Tissue type assignment criteria are proposed to assign each voxel to 5 different tissue types, including 3 pure tissue types "C," "T," and "N" and 2 mixed tissue types "C/T" and "T/N," using the h-maps of normal, tumor, and necrosis obtained by hNMF. Experiments show the feasibility of the proposed method for tissue type assignment.

Conflict of Interests

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

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Project no. 61271287 and Project no. 61371048.

References

[1] P. Kleihues and W. K. Cavenee, *Pathology and Genetics of Tumours of the Nervous System*, International Agency for Research on Cancer Press, Lyon, France, 2000.

[2] A. C. Sava, M. C. Martinez-Bisbal, S. van Huffel, J. M. Cerda, D. M. Sima, and B. Celda, "Ex vivo high resolution magic angle spinning metabolic profiles describe intratumoral histopathological tissue properties in adult human gliomas," *Magnetic Resonance in Medicine*, vol. 65, no. 2, pp. 320–328, 2011.

[3] S. J. Nelson, "Magnetic resonance spectroscopic imaging," *IEEE Engineering in Medicine and Biology Magazine*, vol. 23, no. 5, pp. 30–39, 2004.

[4] D. Gadian, *NMR and Its Applications to Living Systems*, Oxford Science, Oxford, UK, 2nd edition, 1995.

[5] F. S. de Edelenyi, A. W. Simonetti, G. Postma, R. Huo, and L. M. C. Buydens, "Application of independent component analysis to ^{1}H MR spectroscopic imaging exams of brain tumours," *Analytica Chimica Acta*, vol. 544, no. 1-2, pp. 36–46, 2005.

[6] F. S. de Edelenyi, C. Rubin, F. Estève et al., "A new approach for analyzing proton magnetic resonance spectroscopic images of brain tumors: nosologic images," *Nature Medicine*, vol. 6, no. 11, pp. 1287–1289, 2000.

[7] M. de Vos, T. Laudadio, A. W. Simonetti, A. Heerschap, and S. van Huffel, "Fast nosologic imaging of the brain," *Journal of Magnetic Resonance*, vol. 184, no. 2, pp. 292–301, 2007.

[8] T. Laudadio, M. C. Martínez-Bisbal, B. Celda, and S. van Huffel, "Fast nosological imaging using canonical correlation analysis of brain data obtained by two-dimensional turbo spectroscopic imaging," *NMR in Biomedicine*, vol. 21, no. 4, pp. 311–321, 2008.

[9] J. Luts, T. Laudadio, A. J. Idema et al., "Nosologic imaging of the brain: segmentation and classification using MRI and MRSI," *NMR in Biomedicine*, vol. 22, no. 4, pp. 374–390, 2009.

[10] M. F. Ochs, R. S. Stoyanova, F. Arias-Mendoza, and T. R. Brown, "A new method for spectral decomposition using a bilinear Bayesian approach," *Journal of Magnetic Resonance*, vol. 137, no. 1, pp. 161–176, 1999.

[11] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[12] P. Sajda, S. Du, T. R. Brown et al., "Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain," *IEEE Transactions on Medical Imaging*, vol. 23, no. 12, pp. 1453–1465, 2004.

[13] S. Du, X. Mao, P. Sajda, and D. C. Shungu, "Automated tissue segmentation and blind recovery of ^{1}H MRS imaging spectral patterns of normal and diseased human brain," *NMR in Biomedicine*, vol. 21, no. 1, pp. 33–41, 2008.

[14] S. Ortega-Martorell, P. J. G. Lisboa, A. Vellido, M. Julià-Sapé, and C. Arús, "Non-negative matrix factorisation methods for the spectral decomposition of MRS data from human brain tumours," *BMC Bioinformatics*, vol. 13, article 38, 2012.

[15] Y. Li, D. M. Sima, S. van Cauter et al., "Hierarchical non-negative matrix factorization (hNMF): a tissue pattern differentiation method for glioblastoma multiforme diagnosis using MRSI," *NMR in Biomedicine*, vol. 26, no. 3, pp. 307–319, 2013.

[16] Y. Li, D. M. Sima, S. van Cauter et al., "Unsupervised nosologic imaging for glioma diagnosis," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 6, pp. 1760–1763, 2013.

[17] P. A. Bottomley, "Spatial localization in NMR spectroscopy *in vivo*," *Annals of the New York Academy of Sciences*, vol. 508, pp. 333–348, 1987.

[18] D. N. Louis, H. Ohgaki, O. D. Wiestler et al., "The 2007 WHO classification of tumours of the central nervous system," *Acta Neuropathologica*, vol. 114, no. 2, pp. 97–109, 2007.

[19] J. B. Poullet, *Quantification and classification of magnetic resonance spectroscopic data for brain tumor diagnosis [Ph.D. dissertation]*, Department of Electrical Engineering, KU Leuven, Leuven, Belgium, 2008.

[20] R. Kreis, "Issues of spectral quality in clinical ^{1}H-magnetic resonance spectroscopy and a gallery of artifacts," *NMR in Biomedicine*, vol. 17, no. 6, pp. 361–381, 2004.

[21] M. Law, S. Yang, H. Wang et al., "Glioma grading: sensitivity, specificity, and predictive values of perfusion MR imaging and proton MR spectroscopic imaging compared with conventional MR imaging," *American Journal of Neuroradiology*, vol. 24, no. 10, pp. 1989–1998, 2003.

[22] E. J. Delikatny, S. Chawla, D.-J. Leung, and H. Poptani, "MR-visible lipids and the tumor microenvironment," *NMR in Biomedicine*, vol. 24, no. 6, pp. 592–611, 2011.

[23] A. J. Wright, G. Fellows, T. J. Byrnes et al., "Pattern recognition of MRSI data shows regions of glioma growth that agree with DTI markers of brain tumor infiltration," *Magnetic Resonance in Medicine*, vol. 62, no. 6, pp. 1646–1651, 2009.

[24] A. Claes, A. J. Idema, and P. Wesseling, "Diffuse glioma growth: a guerilla war," *Acta Neuropathologica*, vol. 114, no. 5, pp. 443–458, 2007.

[25] D. J. Brat, A. A. Castellano-Sanchez, S. B. Hunter et al., "Pseudopalisades in glioblastoma are hypoxic, express extracellular matrix proteases, and are formed by an actively migrating cell population," *Cancer Research*, vol. 64, no. 3, pp. 920–927, 2004.