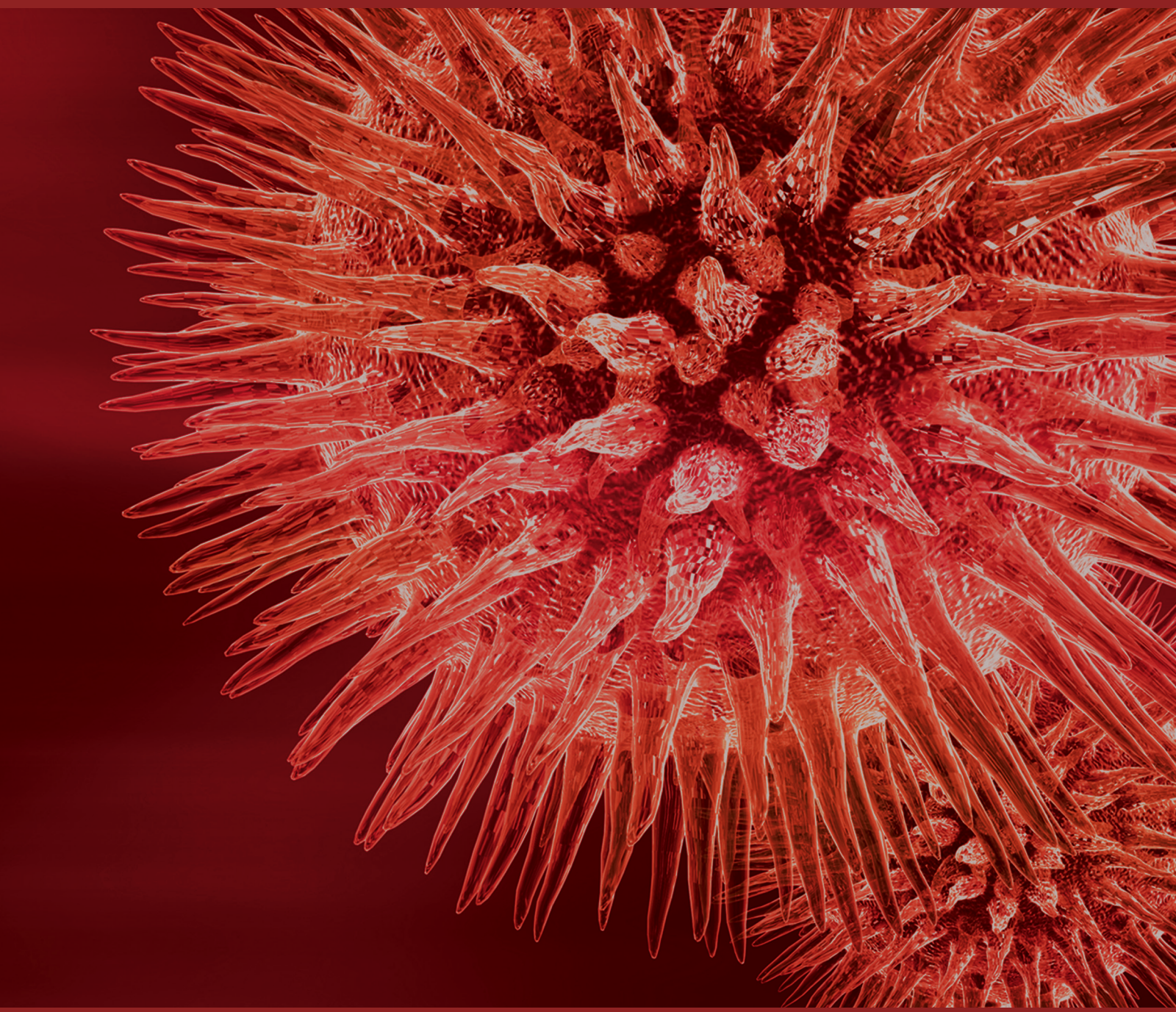
BioMed Research International

Intelligent Informatics in Translational Medicine 2016

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



Intelligent Informatics in Translational Medicine 2016

BioMed Research International

Intelligent Informatics in Translational Medicine 2016

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

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2017, Article ID 1572730, 2 pages
http://dx.doi.org/10.1155/2017/1572730

Editorial

Intelligent Informatics in Translational Medicine 2016

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

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

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

Received 27 December 2016; Accepted 27 December 2016; Published 28 February 2017

This special issue is the second annual special issue of Intelligent Informatics in Translational Medicine (IITM). Only seven papers were published in this special issue after rigorous reviewing processes, of which four papers were selected from the 10th International Conference on Systems Biology (ISB 2016) organized by Computational Systems Biology Society of ORSC (China) in August 2016.

In 2013, we have launched the first special issue with high-quality original research articles, and the title of the special issue was named as Intelligent Informatics in Biomedicine. Most of the published papers were recommended by the International Workshop on Intelligent Informatics in Biology and Medicine (IIBM). In order to extend the scope of research areas and continue this fruitful discussion, we renamed the special issue as the new title of Intelligent Informatics in Translational Medicine (IITM) in 2014 and organized this special issue as a series edition for discovering more biological insights from genomics data or any medical big data. This special issue focuses on the challenges and solutions for information processes with an emphasis on forthcoming high throughput technology and systems biomedicine.

Advances in information technologies are facilitating and accelerating research on molecular biology and medicine. Especially in the precision-medicine era, the increasing complexity and volume of biological data mean that more sophisticated computational techniques are urgently required. Such methods must be able to support new sensing techniques that are being developed to improve the quality of healthcare and medicine. The use of artificial intelligence, machine learning, and data mining can potentially play a significant role in addressing these important challenges. We hope the published series special issue will provide an opportunity for academic and industry professionals to discuss the latest issues and progress in the area of biomedicine.

The current special issue aims to combine intelligent informatics and bioinformatics on diseases and translation medicine, and these collected papers address the data-analytical method design, algorithm development, mathematical modeling, and computational simulation techniques to the translational medical applications.

X. Jiang et al. in their article entitled "Differentially Coexpressed Disease Gene Identification Based on Gene Coexpression Network" designed a novel framework to identify disease-related genes and developed a differentially coexpressed disease-related gene identification method based on gene coexpression network to screen differentially coexpressed genes. They constructed phase-specific gene coexpression network using time-series gene expression data and defined the conception of differential coexpression of genes in coexpression network. They also designed two metrics to measure the value of gene differential coexpression according to the change of local topological structures between different

phase-specific networks. At last, they performed meta-analysis of gene differential coexpression based on a rank-product approach. Their experimental results have shown the feasibility and effectiveness of such a gene coexpression network and the superior performance over other popular disease-related gene selection approaches.

Y. Su et al. described that oncolytic virus is a kind of tumor killer virus which can infect and lyse cancer cells and spread through the tumor, while leaving normal cells largely unharmed. We know that appropriate mathematical models could help biologists to understand the tumor-virus dynamics and find better treatment strategies. The authors proposed a new mathematical model of tumor therapy with oncolytic virus and MEK inhibitor. Due to mitogen-activated protein kinase providing greater oncolytic virus infection into cancer cells and limiting the replication of the virus, in order to provide the best dosage of MEK inhibitors and balance the positive and negative effects of the inhibitors, authors proposed an optimal control strategy regarding the inhibitor. Simulations have shown that the optimal control indeed possessed better control effects than constant control under the same initial conditions.

C. Quan et al. proposed a multichannel convolutional neural network for automated biomedical relation extraction. The proposed model provides two contributions: it enables the fusion of multiple versions in word embedding and the need for manual feature engineering can be obviated by automated feature learning with convolutional neural network. The authors have evaluated the proposed model on two biomedical relation extraction tasks including drug-drug interaction extraction and protein-protein interaction extraction. For conducting several experimental trials on benchmark testing datasets, the proposed system outperformed the standard SVM based systems within a range from 2.7% to 5.6% on F-score measurement.

Y. X. Goay et al. in their article entitled "Identification of Five Novel *Salmonella* Typhi-Specific Genes as Markers for Diagnosis of Typhoid Fever Using Single-Gene Target PCR Assays" identified new markers to detect the *Salmonella* Typhi pathogen and developed sensitive and specific diagnostic tests in this study. Based on genomic comparison of *Salmonella* Typhi with other enteric pathogens, there are six *Salmonella* Typhi genes found to be specific, and corresponding PCR assays for each target gene were developed to verify their specificity and sensitivity in vitro. The experimental results showed that 5 genes selected from the 6 candidate genes could be demonstrated with perfect verification results (100% sensitivity and 100% specificity), and these genes could be applied as important biomarkers for diagnosis of typhoid fever through single-gene target PCR-assays.

N. Yu et al. proposed a single-trial sparse representation-based approach for visual evoked potential (VEP) extraction. Sparse representation is a powerful tool in signal denoising, and visual evoked potentials have been proven to have strong sparsity over an appropriate dictionary. The authors presented such a novel sparse representation-based approach to solving the extraction problem. Three stages were proposed in their article: utilizing the previous electroencephalogram (EEG) data to identify the parameters of the EEG autoregressive (AR) model; applying sparse representation to model the VEPs in the autoregressive-moving average model; calculating the sparse coefficients and deriving VEPs by using the AR model. The proposed method was verified and compared by employing synthetic and real data against different existing methods, and the evaluation demonstrates that their proposed method can well preserve the details of the VEPs for latency estimation, even in low SNR environments.

B. Tang developed a toolkit META2 for DNA methylation annotation and analysis. The tool performs integrative analysis on differentially methylated loci and regions through deep mining and statistical comparison methods. The author examined the association within differentially methylated CpG and differentially methylated region candidates regarding counts and region lengths and identified major transition zones as clues for inferring statistically significant regions. The developed tool can provide a comprehensive analysis approach for epigenetic research and clinical study.

Y. Li et al. constructed a multilevel evolutionary structure for avian influenza virus system based on considering both hemagglutinin and neuraminidase protein fragments. An optimization model was established to determine the rational granularity of the virus system for exploring the intrinsic relationship among the subtypes based on the fuzzy hierarchical evaluation index. To reduce the systematic and computational complexity of the proposed algorithm, the granular signatures of virus system were identified based on the coarse-grained idea. The proposed system can effectively identify the virus signatures and reflect the whole avian influenza virus system, which indicates that the proposed method provides an alternative mechanism to perform new virus subtyping comparison and functional prediction.

Acknowledgments

We wish to express our appreciation to all the authors for their excellent contribution. We also want to thank all reviewers and editors for their hard work on this issue.

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

Hindawi Publishing Corporation
BioMed Research International
Volume 2017, Article ID 5404180, 7 pages
http://dx.doi.org/10.1155/2017/5404180

Research Article

Construction of Multilevel Structure for Avian Influenza Virus System Based on Granular Computing

Yang Li, Qi-Hao Liang, Meng-Meng Sun, Xu-Qing Tang, and Ping Zhu

School of Science, Jiangnan University, Wuxi 214122, China

Correspondence should be addressed to Ping Zhu; zhuping@jiangnan.edu.cn

Received 11 September 2016; Revised 1 December 2016; Accepted 14 December 2016; Published 16 January 2017

Academic Editor: Hao-Teng Chang

Exploring the genetic structure of influenza viruses attracts the attention in the field of molecular ecology and medical genetics, whose epidemics cause morbidity and mortality worldwide. The rapid variations in RNA strand and changes of protein structure of the virus result in low-accuracy subtyping identification and make it difficult to develop effective drugs and vaccine. This paper constructs the evolutionary structure of avian influenza virus system considering both hemagglutinin and neuraminidase protein fragments. An optimization model was established to determine the rational granularity of the virus system for exploring the intrinsic relationship among the subtypes based on the fuzzy hierarchical evaluation index. Thus, an algorithm was presented to extract the rational structure. Furthermore, to reduce the systematic and computational complexity, the granular signatures of virus system were identified based on the coarse-grained idea and then its performance was evaluated through a designed classifier. The results showed that the obtained virus signatures could approximate and reflect the whole avian influenza virus system, indicating that the proposed method could identify the effective virus signatures. Once a new molecular virus is detected, it is efficient to identify the homologous virus hierarchically.

1. Introduction

Exploring the genetic structure of biological population attracts the focus in the field of population biology, molecular ecology, and medical genetics [1]. Influenza A virus is a negative-strand RNA virus, which encodes the 8 structural proteins and 2 nonstructural proteins. In the past several decades, some subtypes of influenza viruses have been identified to infect humans, whose epidemics cause morbidity and mortality worldwide [2, 3]. Subtyping identification of a virus is typically based on viral hemagglutinin (HA) and neuraminidase (NA) fragments among the 10 encoded proteins [4, 5]. So far, dozens of subtypes, combination of the 16 HA and 9 NA types, make up the whole viral system and it was verified that different labeled viruses descend from the same ancestor according to microscopic structural features and genome organization analysis [6]. Evolutionary forces, treated as the most important molecular mechanisms, such as natural selection acting upon rapidly mutating viral populations could shape the genetic structure of influenza viruses in different hosts, geographic regions, and periods of time with genetic mutation [7]. In addition, influenza viruses are equipped with antigenic changes, known as antigenic shifts among different subtypes of influenza viruses, which results in structural changes to escape the immunity [8]. It is of crucial importance to identify the subtypes and analyze the evolutionary relationships for developing antiviral drugs and vaccines. Thus, accessing the viral genomes in a timely fashion and developing effective analyzing methods are urgently needed.

The dramatic progress in sequencing technologies provides unprecedented prospects for the exploration of virus homologous and mutation trajectory in space and time. Understanding the evolution of influenza viruses has benefited from phylogenetic reconstructions of the hemagglutinin protein [9]. In an alternative approach, Lapedes and Farber [10] applied a technique called multidimensional scaling to study antigenic evolution of influenza. Plotkin et al. [8] clustered hemagglutinin protein sequences using the single-linkage clustering algorithm and found that influenza viruses group into several clusters. Upon the dimensional projection technique to characterize hemagglutination inhibition (HI)

data, a low-dimensional clustering method that can detect the clusters containing an incipient dominant strain was presented by He and Deem [11]. However, those works just focused on the one fragment, especially HA protein, to explore the evolutional relationships. And large volume of data poses some daunting challenges for exploring the structure of the complex system and the intrinsic relationship. Therefore, there is a need for less computationally intensive methods.

In recent years, the granular computing (GrC) theory has become a hotspot in the field of artificial intelligence and machine learning, which comes from the idea that people solve the problems from different levels and views [12]. Clustering technique is an effective way to generate granules of complex system. Y. Y. Yao and J. T. Yao accomplished a series of research work for applying the theory to data mining and some other fields [13]. Hartmann et al. [14] proposed supervised hierarchical clustering in fuzzy model identification by using hierarchical tree construction. Tang et al. [15, 16] introduced the granular space to describe the hierarchical structural information by using the algebraic topology based on the fuzzy quotient space theory [12]. He also studied the hierarchical clustering structure and analyzed the fuzzy equivalence (or proximity) relation based on the fuzzy granular space. Constructing the hierarchical structure of complex system and extracting the essential information among the granules on different granularities are the goals.

In this paper, our aim is to explore the evolutional relationships of the avian influenza viruses in the same subtype and among the subtypes considering both HA and NA fragments in the virus system. Moreover, the complex virus system should be reduced for further exploration, faced with thousands of samples in the dataset. Jointing the two protein sequences, the feature vectors are extracted from HA and NA proteins, respectively, for labeling the specific virus. Furthermore, the granular signatures in the viral granules are identified based on the obtained features to reduce the systematic and computational complexity and then its performance will be evaluated. This will provide the supports for the rationality of subtype identification. Once a new molecular virus is detected, it could be analyzed with obtained viral signatures and then the prevention and treatment measures can follow what were applied in the viral signature.

2. Materials and Methods

2.1. Materials. The influenza virus dataset was downloaded from the NCBI Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/) [17]. The influenza virus contains eight linear negative-strand RNA fragments, which encode 10 viral proteins, that is, PB1, PB2, PA, HA, NP, NA, M1, M2, NS1, and NS2, among which most are structural proteins except NS1 and NS2. Notably, HA and NA fragments play the direct and important roles in the viral subtyping identification and the functions [18]. It has been verified that 8 subgroups of avian influenza virus (H5N1, H5N2, H7N2, H7N3, H7N7, H9N2, H10N7, and H7N9) could infect people, which occurred from 1902 to 2015 around the world. The avian influenza viruses are labeled with unambiguous symbols such as the host, outbreak time, and detection sites. Removing some vague and uncompleted viruses, there are 8274 influenza viruses which reserve HA and NA protein fragments simultaneously (13143 HA protein fragments and 9401 NA protein fragments), compositing the whole avian virus protein system, denoted as Ω. According to the physicochemical property [19], amino acids are divided into four types, namely, the polar and hydrophilic (*pq*), polar and hydrophobic (*pr*), nonpolar and hydrophilic (*sq*), and nonpolar and hydrophobic (*sr*). Considering the adjacency statistical information, the 16-dimension feature vector is extracted by calculating the frequency from one protein sequence. Therefore, 32-dimension feature vector is extracted to represent a virus molecule.

2.2. The Optimization Model for Extracting the Hierarchical Structure. A relation R on a universe X is a fuzzy proximity (FP) relation if it satisfies the reflexivity and symmetry [16, 20]. Furthermore, if R is an FP relation on the universe X and satisfies the separable condition ($\forall x, y \in X$, $R(x, y) = 1 \leftrightarrow x = y$), then R is called a separable FP relation (or SFP relation).

In [16], the granular space of FP (or SFP) relations on the universe X was introduced, and then their properties were explored. Let R be an FP (or SFP) relation on a finite universe $X = \{x_1, x_2, \ldots, x_n\}$, where X is a dataset of K-dimension space. For any $\lambda \in [0, 1]$, we define a relation $R_\lambda : (x, y) \in R_\lambda \leftrightarrow R(x, y) \geq \lambda$, where R_λ is a crisp proximity relation that satisfies the reflexivity and symmetry. Then, the equivalent classes of the transitive closure $\mathrm{tr}(R_\lambda)$ can be marked by $[x]_\lambda$, which is derived by R_λ, and then $X(\lambda) = \{[x]_\lambda \mid x \in X\}$ is a granularity corresponding to λ. The set $\{X(\lambda) \mid \lambda \in [0, 1]\}$ represents a fuzzy granular space on X, which is an ordered set, and satisfies that the bigger the threshold λ is, the finer the granularity is, denoted by $\aleph_{\mathrm{TR}}(X)$ [16].

The granularity derived by λ is marked as $X(\lambda) = \{a_1, a_2, \ldots, a_{c_\lambda}\}$, where $a_i = \{x_{i1}, x_{i2}, \ldots, x_{iJ_i}\}$ satisfying the conditions that $|a_i| = J_i$ ($|\cdot|$ stands for the number of the elements in a set) and $\sum_{i=1}^{c_\lambda} J_i = n$. Some properties are explored, such as $\overline{a_i} = \sum_{k=1}^{J_i} x_{ik}/J_i$ ($i = 1, 2, \ldots, c_\lambda$) is the center of granule a_i and the center of X is $\overline{a} = \sum_{i=1}^{c_\lambda} \sum_{k=1}^{J_i} x_{ik}/n$. From the perspective of statistical theory, two indexes are introduced to measure the deviations within the classes and among the classes on the granulation $X(\lambda)$ [18, 20], defined, respectively, as follows:

$$S_{\mathrm{among}}(X(\lambda)) = \frac{\sum_{i=1}^{c_\lambda} J_i \left\| \overline{a_i} - \overline{a} \right\|_2^2}{n},$$
$$S_{\mathrm{within}}(X(\lambda)) = \frac{\sum_{i=1}^{c_\lambda} \sum_{k=1}^{J_i} \left\| x_{ik} - \overline{a} \right\|_2^2}{n}, \quad (1)$$

where $\|\cdot\|_2$ stand for the 2-norm number in K-dimension space.

By analyzing the variance within and among the classes in statistics [21], $S_{\mathrm{among}}(X(\lambda))$ is monotone increasing, with the granularity changing from the coarse to the fine, while

$S_{\text{within}}(X(\lambda))$ is gradually decreasing. Notably, the total deviation ($S(X(\lambda)) = S_{\text{among}}(X(\lambda)) + S_{\text{within}}(X(\lambda))$) is always constant $S(X(\lambda)) = \sum_{i=1}^{n} \|x_i - \overline{a}\|_2^2/n$. Additionally, $S_{\text{among}}(X(0)) = S_{\text{within}}(X(1)) = 0$ and $S_{\text{among}}(X(1)) = S_{\text{within}}(X(0)) = \sum_{i=1}^{n} \|x_i - \overline{a}\|_2^2/n$. Therefore, a fuzzy hierarchical evaluation index (FHEI) based on the fuzzy granular space is proposed as follows:

$$\text{FHEI}(X(\lambda)) = \left| S_{\text{among}}(X(\lambda)) - S_{\text{within}}(X(\lambda)) \right|. \quad (2)$$

We establish an optimization model to determine the reasonable granulation in the granular space with the minimal objective; that is, FHEI($X(\lambda)$) reaches the minimum. There exists only one $\lambda = \lambda_0$ to meet the optimization model, marked as Model (2):

$$X(\lambda_0) = \arg \min_{X(\lambda) \in \aleph_{\text{TR}}(X)} \{\text{FHEI}(X(\lambda))\}. \quad (3)$$

Remark 1. Model (2) is a global optimization model without constraints on the hierarchical structure of the finite universe X. Compared with [18], their model for determining the optimal hierarchical clustering has the restriction $S_{\text{among}}(X(\lambda)) > S_{\text{within}}(X(\lambda))$.

Given an FP relation (or SFP relation) R on the finite set $X = \{x_1, x_2, \ldots, x_n\}$ and $D = \{R(x, y) \mid x, y \in X\} = \{r_0, r_1, \ldots, r_N\}$, satisfying $1 = r_0 > r_1 > \cdots > r_N$, an algorithm is presented to detect the optimized hierarchical clustering and construct the hierarchy of complex system based on the fuzzy granular space [16].

Algorithm A.

Input: an FP relation (or SFP relation).

Output: the optimized hierarchical structure and the corresponding threshold.

Step 1

$$X(r_i) = C = \{a_1, a_2, \ldots, a_{c_1}\},$$
$$S_0 \Longleftarrow \left| S_{\text{among}}(X(r_i)) - S_{\text{within}}(X(r_i)) \right| \quad (4)$$
$$i = 0.$$

Step 2

$$i \Longleftarrow i + 1. \quad (5)$$

Step 3

$$A \Longleftarrow C. \quad (6)$$

Step 4

$$B \Longleftarrow \varnothing,$$
$$C \Longleftarrow \varnothing. \quad (7)$$

Step 5. For any $a_j \in A$, $B \Leftarrow B \cup a_j$, $A \Leftarrow A \setminus a_j$.

Step 6. For $\forall a_k \in A$, if $\exists x_j \in a_j$, $y_k \in a_k$ satisfying $R(x_j, x_k) \geq r_i$, $B \Leftarrow B \cup a_k$, $A \Leftarrow A \setminus a_k$.

Step 7

$$C \Longleftarrow \{B\} \cup C. \quad (8)$$

Step 8. If $A = \varnothing$, $X(r_i) = C$; otherwise, go to Step 5.

Step 9. If $X(r_i) \neq X(r_{i-1})$, $S_1 \Leftarrow |S_{\text{among}}(X(r_i)) - S_{\text{within}}(X(r_i))|$; otherwise, go to Step 2.

Step 10. If $S_0 > S_1$, $S_0 \Leftarrow S_1$, go to Step 2.

Step 11. Output r_{i-1}, $X(r_{i-1})$ and S_0.

The computational complexity of Algorithm A is $O(n^2)$. The concrete problems are decomposed hierarchically, which is consistent with the core idea of GrC. Given an FP (or SFP) relation on the finite set X, the optimization clustering structure constructed by Algorithm A is its first level structure. Furthermore, its second level structure is obtained if Algorithm A is repeatedly applied to all the equivalent classes in its first level structure. Therefore, Algorithm A can be used to construct multilevel structure in practical application.

2.3. Identification of Granular Signature. Once the optimal granularity of the complex system is determined, it is of crucial importance to construct information granules for abstracting original samples. Generally, the granules are obtained according to the principle: the samples with the same features assemble in one granule. And the average of all samples in one class or the center of the class is efficacious to represent the core information. Suppose that a multilevel structure (or granularity) $X^* = \{a_1, a_2, \ldots, a_J\}$ is constructed, where $J = |X^*|$. To reduce the complexity of the system, feature viruses (or signature viruses) could be extracted to approximately represent the equivalent class. According to the nearest-to-center principle, an objective function to select the signature is established, and it is formulated as follows:

$$p_i = \arg \max_{1 \leq k \leq J_i} \{R(x_{ik}, \overline{a_i})\}, \quad (9)$$

where p_i is the signature item of the granule a_i and $P = \{p_1, p_2, \ldots, p_J\}$ is a signature set of the granularity X^*. In some way, the signature set P can be used to represent approximately the complex system X.

2.4. Validation of Granular Signature Set. To evaluate the performance of selected signature set P, a classifier is designed for classifying the rest of the samples of the corresponding classes according to the principle of maximum similarity, marked as Model (3). Given a virus q_j ($\in X \setminus P$), the classifier is designed:

$$L_j = \arg \max_i \{R(q_j, p_i)\}, \quad (10)$$

Table 1: The 8 subtypes of avian influenza virus.

Subtype	Number	Subtype	Number	Subtype	Number	Subtype	Number
H5N1	306	H5N2	127	H7N3	70	H9N2	199
H7N9	24	H7N2	40	H7N7	68	H10N7	75

where $p_i \in P$, $i = 1, 2, \ldots, J$, and L_j is the class the virus q_j belongs to.

Model (3) states that the signature viruses are treated as the classifying targets and the other samples in $X \setminus P$ are assigned to $|P|$ classes. All samples in $X \setminus P$ are divided into $|P|$ classes according to Model (3), marked as b_k, $k = 1, 2, \ldots, |P|$. The accuracy ratio r is introduced to measure the efficiency of signature set for constructing the multilevel structure X^*. It is defined as

$$r = \frac{\sum_{k=1}^{|P|} |a_k \cap b_k|}{|X \setminus P|}. \tag{11}$$

In formula (11), the overlapped ratio r is proposed, which measures the rationality of the obtained signature to represent the whole virus system. And the bigger the value r is, the better the result is.

3. Results and Analysis

In this section, we apply the proposed model to the avian influenza virus system for constructing the evolutionary structure, which contains 8274 viral HA and NA protein fragments simultaneously within 8 subtypes, listed in Table 1.

Based on the feature vectors extracted from the viral HA and NA proteins, the 32-dimension vector $x_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_{32}})$ labels the specific virus x_i. Furthermore, the similarity between viruses x_i and x_j is measured:

$$R(x_i, x_j) = \frac{(x_i, x_j)}{\sqrt{(x_i, x_i) \cdot (x_j, x_j)}}, \tag{12}$$

where $(x_i, x_j) = \sum_{i=1}^{32} x_{ik} \cdot x_{jk}$ stands for the inner product in 32-dimension space. Obviously, R is an SFP relation.

The virus dataset has redundant information as many viruses are labeled with the same host, the same occurrence time, and the same outbreak sites, which could pose the obstacle to explore the intrinsic relationship and difference among the subtypes. Thus, those with the same host, the same occurrence time, and the same location combine as one new point (a representative virus), which is the preliminary system simplification, and then a unique virus database Ω^* is obtained. The FHEI is applied to virus system Ω^* containing 909 avian influenza viruses, to obtain the reasonable partition and evolutionary structure.

On the basis of the virus database Ω^*, the viral granular space (evolutionary structure) is constructed by using Algorithm A. On the first level, 3 equivalent classes were finally determined to partition the whole system, and the corresponding signature viruses are obtained, shown in Table 2. For the virus granules on the first level, class A1 contains the most viruses (about 93.5%), and granule A3 arises, containing an isolated virus (A/American green-winged teal/Washington/1595750/2014(H5N1)). Therefore, it is necessary to construct the second level of virus system. For each virus granule on the first level, Algorithm A is used repeatedly, which is to refine the granules to get the detailed evolutional structure. 14 equivalent subclasses are identified, denoted as b_k^* ($k = 1, 2, \ldots, 14$), and the virus signatures are extracted, shown in Table 3. From Tables 2 and 3, we construct the two-level feature structure of the whole virus system by using the signature viruses on first level and second level structure.

Table 2: Three signature viruses of the first level structure.

Number	Virus number	Virus signature
A1	850	A/Pekin duck/Singapore/F59/04/98(H5N2)
A2	58	A/chicken/Tunisia/145/2012(H9N2)
A3	1	A/American green-winged teal/Washington/1595750/2014(H5N1)
Sum	909	

The virus signature could be used to approximate the whole system for they are selected from the classes as the granule information. Moreover, the classifier, designed based on the principle of maximum similarity, is applied to validate the performance of virus signature. The accuracy rate of the signature virus set P^* on the second level structure is 76.57% by comparison, indicating that the second level structure of viruses system constructed by our model is effective.

Remark 2. Evaluating the performance of virus signature, the error rate is still 23.43%, which might be caused by the approximation process since all signature viruses are selected according to the nearest-to-center principle and they are not just on the center of each subclass, respectively. From the perspective of approximation, the signature set contains the most information of virus system according to the accuracy rate 76.57%. Therefore, the signature virus set P^* containing 14 viruses can be used to approximate the whole system containing 909 viruses.

The phylogenetic tree of the signature virus set P^* can be constructed by applying the hierarchical clustering algorithm [16], shown in Figure 1. According to Remark 2, it can also be treated as the core structure of whole influenza viruses system, which helps us understand the evolutionary history and the mechanism of evolution [22].

Table 3: The virus classes on the second level structure.

Number	Virus number	First level	Virus signature
B1	1	A1	A/chicken/Cambodia/LC/2006(H5N1)
B2	1	A1	A/dog/Shandong/JT01/2009(H5N2)
B3	177	A1	A/chicken/Israel/184/2009(H9N2)
B4	665	A1	A/duck/Taiwan/DV1236/2009(H5N2)
B5	1	A1	A/blue-winged teal/LA/AI13-1225/2013(H7N7)
B6	2	A1	A/duck/Korea/A349/2009(H7N2)
B7	2	A1	A/chicken/Abbottabad/NARC-2419/2005(H7N3)
B8	1	A1	A/mallard/Netherlands/22/2010(H10N7)
B9	12	A2	A/swine/Hong Kong/2106/98(H9N2)
B10	35	A2	A/chicken/Italy/330/1997(H5N2)
B11	1	A2	A/chicken/Queensland/1995(H7N3)
B12	9	A2	A/chicken/Iran/261/01(H9N2)
B13	1	A2	A/oystercatcher/Peru/MM152/2008(H10N7)
B14	1	A3	A/American green-winged teal/Washington/195750/2014(H5N1)

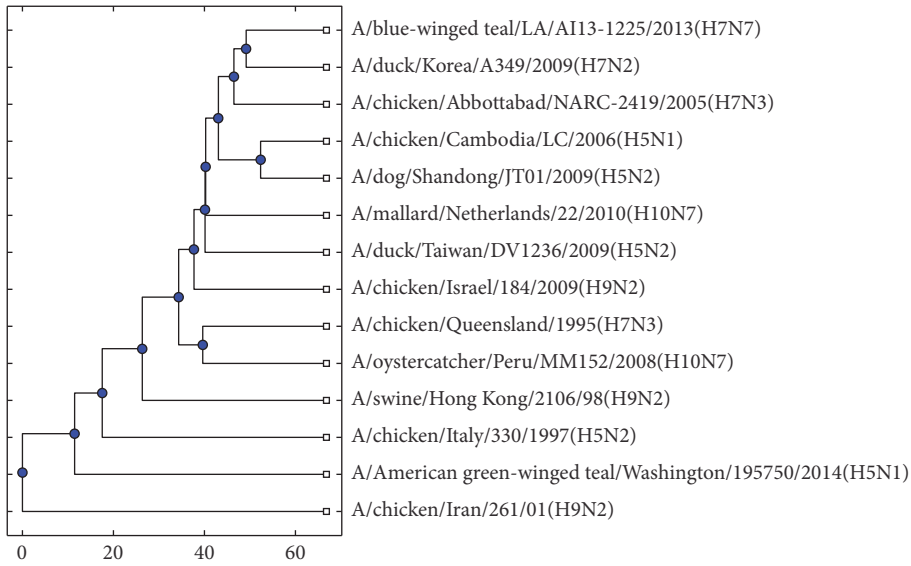


Figure 1: The phylogenetic tree of signature viruses on second level structure.

Among the 8 virus subtypes, 7 viruses are identified as the signature viruses except H7N9, for H7N9 viruses account for the minority of whole system (Figure 1). Exploring the intrinsic relation, it is obvious that H7N9 belongs to class B4, elucidating that the variation of H7N9 is not significant [23] and can be viewed as a new member in the family of viruses. Based on the coarse-grained idea, one signature virus represents the corresponding class. However, some isolated points are detected, such as A/chicken/Cambodia/LC/2006(H5N1), A/dog/Shandong/JT01/2009(H5N2), and A/chicken/Queensland/1995(H7N3), which might be caused by the big change to virus RNA strain.

From the hierarchical structure of the feature viruses, A/blue-winged teal/LA/AI13-1225/2013(H7N7), A/duck/Korea/A349/2009(H7N2), and A/chicken/Abbottabad/NARC-2419/2005(H7N3) have similar evolution relationship (connect closely) for they equip the same HA type (H7). Besides, A/chicken/Cambodia/LC/2006(H5N1) and A/dog/Shandong/JT01/2009(H5N2) have the consistent conclusions. However, A/chicken/Italy/330/1997(H5N2) is far from them, which could be due to the fact that the outbreak time plays an important role in sequence mutation. If just considering the HA and NA proteins, the subtypes, such as H9N2 [24], should be redefined. Comparing the two-level

structure and hierarchical structure of virus signature, the intrinsic relationship among A1 on the first level structure is consistent with that in the hierarchical structure, while class A2 has the dispersed structure where the feature viruses in different subtypes scatter in chaos, which indicates that constructing the second level structure is meaningful.

4. Conclusions

The rapid variation of influenza viruses results in low-accuracy subtyping identification and makes it difficult to develop effective drugs. This article explored the homology of avian influenza virus system and identified the subtypes according to HA and NA protein fragments, which might provide the support for developing antiviral drugs and vaccines according to different subtypes. Phylogenetic reconstructions serve understanding the evolution of influenza viruses. However, the large amounts of virus dataset pose an obstacle for analyzing the evolutionary relationship and identifying the correct subtypes to predict the biological functions. Granular computing theory was applied to determine the partition of virus system based on the constructed granular space. A method and the corresponding algorithm were proposed for detecting the rational granularity. With the proposed algorithm applied repeatedly, a multilevel structure of whole system was constructed. To reduce the computational complexity, some key viruses were selected to approximate the whole system based on the coarse-grained idea. According to the nearest-center principle, virus signatures were identified and constructed the granular signature set of a multilevel structure of complex system. By designing a classifier, the performance of virus signatures was evaluated and the result showed that the virus signatures could reflect the most properties of virus system. Furthermore, hierarchical structure of virus signature was constructed by using hierarchical clustering algorithm. Both of the two structures have some consistent intrinsic relationship among the virus systems and between the different subtypes. Some viruses were detected as isolated points in the structure thought equipped with the same labels, which might be caused by the rapid variations in the RNA strands. The virus signatures have the potential use in new virus subtyping comparison and functional prediction.

Disclosure

The work was previously presented in "The 10th International Conference on Systems Biology (ISB 2016, held in Weihai, China, August 19–22, 2016)."

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The work was supported by National Natural Science Foundation of China (Grant nos. 11371174, 11271163) and Colleges and Universities in Jiangsu Province Plans to Graduates Research and Innovation (Grant no. KYLX15_1188).

References

[1] T. Jombart, S. Devillard, and F. Balloux, "Discriminant analysis of principal components: a new method for the analysis of genetically structured populations," *BMC Genetics*, vol. 11, no. 1, article 94, 2010.

[2] F. G. Hayden, "Prevention and treatment of influenza in immunocompromised patients," *The American Journal of Medicine*, vol. 102, no. 3, pp. 55–60, 1997.

[3] N. J. Cox and K. Subbarao, "Global epidemiology of influenza: past and present," *Annual Review of Medicine*, vol. 51, pp. 407–421, 2000.

[4] R. M. Bush, C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch, "Predicting the evolution of human influenza A," *Science*, vol. 286, no. 5446, pp. 1921–1925, 1999.

[5] J. Xu, C. T. Davis, M. C. Christman et al., "Evolutionary history and phylodynamics of influenza A and B neuraminidase (NA) genes inferred from large-scale sequence analyses," *PLoS ONE*, vol. 7, no. 7, Article ID e38665, 2012.

[6] R. G. Webster, W. J. Bean, O. T. Gorman, T. M. Chambers, and Y. Kawaoka, "Evolution and ecology of influenza A viruses," *Microbiological Reviews*, vol. 56, no. 1, pp. 152–179, 1992.

[7] E. Ghedin, N. A. Sengamalay, M. Shumway et al., "Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution," *Nature*, vol. 437, no. 7062, pp. 1162–1166, 2005.

[8] J. B. Plotkin, J. Dushoff, and S. A. Levin, "Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 9, pp. 6263–6268, 2002.

[9] C. A. Russell, T. C. Jones, I. G. Barr et al., "The global circulation of seasonal influenza A (H3N2) viruses," *Science*, vol. 320, no. 5874, pp. 340–346, 2008.

[10] A. Lapedes and R. Farber, "The geometry of shape space: application to influenza," *Journal of Theoretical Biology*, vol. 212, no. 1, pp. 57–69, 2001.

[11] J. He and M. W. Deem, "Low-dimensional clustering detects incipient dominant influenza strain clusters," *Protein Engineering, Design and Selection*, vol. 23, no. 12, pp. 935–946, 2010.

[12] B. Zhang and L. Zhang, *Theory and Applications of Problem Solving*, Elsevier Science Inc, Amsterdam, Netherlands, 1992.

[13] Y. Y. Yao and J. T. Yao, "Granular computing as a basis for consistent classification problems," in *Proceedings of the Workshop on Foundations of Data Mining (PAKDD '02)*, pp. 101–102, 2002.

[14] B. Hartmann, O. Bänfer, O. Nelles, A. Sodja, L. Teslić, and I. Škrjanc, "Supervised hierarchical clustering in fuzzy model identification," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 6, pp. 1163–1176, 2011.

[15] X.-Q. Tang, P. Zhu, and J.-X. Cheng, "The structural clustering and analysis of metric based on granular space," *Pattern Recognition*, vol. 43, no. 11, pp. 3768–3786, 2010.

[16] X.-Q. Tang and P. Zhu, "Hierarchical clustering problems and analysis of fuzzy proximity relation on granular space," *IEEE Transactions on Fuzzy Systems*, vol. 21, no. 5, pp. 814–824, 2013.

[17] Y. Bao, P. Bolotov, D. Dernovoy et al., "The influenza virus resource at the National Center for Biotechnology Information," *Journal of Virology*, vol. 82, no. 2, pp. 596–601, 2008.

[18] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, Amsterdam, Netherlands, 2011.

[19] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *Journal of Molecular Biology*, vol. 238, no. 1, pp. 54–61, 1994.

[20] B. De Baets and E. Kerre, "Fuzzy relations and applications," *Advances in Electronics and Electron Physics*, vol. 89, pp. 255–324, 1994.

[21] M. R. Anderberg, *Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks*, Academic Press, Cambridge, Mass, USA, 2014.

[22] Z.-W. Chen and X.-Q. Li, "Whole-genome phylogeny based on protein domain information," *China Journal of Bioinformatics*, vol. 1, article 10, 2012.

[23] R. Gao, B. Cao, Y. Hu et al., "Human infection with a novel avian-origin influenza A (H7N9) virus," *The New England Journal of Medicine* , vol. 368, no. 20, pp. 1888–1897, 2013.

[24] M. Peiris, K. Y. Yuen, C. W. Leung et al., "Human infection with influenza H9N2," *The Lancet*, vol. 354, no. 9182, pp. 916–917, 1999.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 1597489, 10 pages
http://dx.doi.org/10.1155/2016/1597489

Research Article

META2: Intercellular DNA Methylation Pairwise Annotation and Integrative Analysis

Binhua Tang[1,2]

[1]*Epigenetics & Function Group, School of Internet of Things, Hohai University, Jiangsu 213022, China*
[2]*School of Public Health, Shanghai Jiao Tong University, Shanghai 200025, China*

Correspondence should be addressed to Binhua Tang; bh.tang@outlook.com

Received 17 September 2016; Accepted 12 December 2016

Academic Editor: Hao-Teng Chang

Genome-wide deciphering intercellular differential DNA methylation as well as its roles in transcriptional regulation remains elusive in cancer epigenetics. Here we developed a toolkit META2 for DNA methylation annotation and analysis, which aims to perform integrative analysis on differentially methylated loci and regions through deep mining and statistical comparison methods. META2 contains multiple versatile functions for investigating and annotating DNA methylation profiles. Benchmarked with T-47D cell, we interrogated the association within differentially methylated CpG (DMC) and region (DMR) candidate count and region length and identified major transition zones as clues for inferring statistically significant DMRs; together we validated those DMRs with the functional annotation. Thus META2 can provide a comprehensive analysis approach for epigenetic research and clinical study.

1. Introduction

Genome-wide DNA methylation analysis and annotation across multiple samples are essential in interrogating pairwise base-pair differences, while it still remains elusive in recent pancancer studies [1–7]. Pancancer DNA methylation study can retrieve cell- and tissue-specific properties by detecting differentially methylated loci and regions.

Heyn et al. adopted the Illumina Infinium 450 K technique to identify DOK7 as novel biomarker in breast cancer [8]; and the genome-wide composition, patterning, cell specificity, and dynamics of DNA methylation at single-base resolution in human and mouse frontal cortex throughout their lifespan were reported recently [9]; Bell et al. applied whole-blood DNA methylation to investigate molecular clues in chromic pain [10].

However, till now, our knowledge about the genome-wide distribution of DNA methylation, how to decipher the genome-wide difference, and how it relates to other epigenetic modifications in mammals remains limited. And there still lacks comprehensive analysis toolkits for biochemical experiment design and postexperiment validation.

Herein we developed an analysis toolkit, META2, for intercellular DNA methylation annotation and analysis. META2 is mainly designed for analyzing the reduced representation bisulfite sequencing (RRBS) profiling data [11–13]; together it can analyze data with the right formats from other platforms, such as HumanMethylation 450 K beadchip assay [14–16]. META2 can implement intercellular interrogation of DNA methylation status among multiple samples, perform statistical analysis on methylated CpG loci and regions, and yield integrative visualization for the analysis results.

We also validated the toolkit on the real RRBS data retrieved from ENCODE consortium and demonstrated its integrative analysis on the last section. Our developed toolkit aims to provide a versatile analysis approach to the epigenetic research fields, and we also deposited the toolkit on GitHub for public convenient usage.

2. Structure and Function Composed in META2

The toolkit META2 contains several major functional procedures, namely, (i) DNA methylation raw data acquisition and

preprocess; (ii) statistical analysis and information retrieval; and (iii) integrative analysis and visualization, as depicted in Figure 1.

The first functional procedure of META2 is the acquisition and curation of raw DNA methylation data, for example, sequencing-based RRBS and array-based 450 K platforms [17, 18]. This procedure covers preprocessing the raw DNA methylation information, from integration of sample list (pairwise control versus treatment replicates) to genome-wide identification of differentially methylated loci information (chr1 to chr22, chrX, and chrY).

The second functional procedure is statistical analysis, genomic annotation, and functional information retrieval from the curated DNA methylation profiles, which outputs the differentially methylated CpG (DMC) or region (DMR) for pairwise samples and intercellular interrogation [19], together with the statistical property analysis of those output sources.

The last analysis procedure is the integration and visualization, which provides insightful clues for statistical comparison and further experiment validation; it aims to identify statistically significant DMRs with underlying biological functions of interest and annotate those DMRs with genetic transcript information, together with region-specific reference genome sequence information. Thus, such integrative comparison can shed light on the vital regulatory processes leading to carcinogenesis with a systematic approach.

In the following sections, we will demonstrate the major analysis procedures and corresponding statistical comparisons and integrative visualization on the curated DNA methylation data in RRBS format [17, 18], and we will identify DMR and classify the hyper- and hypo-DMR candidates [19] and implement function annotation for methylated CpG sites and regions.

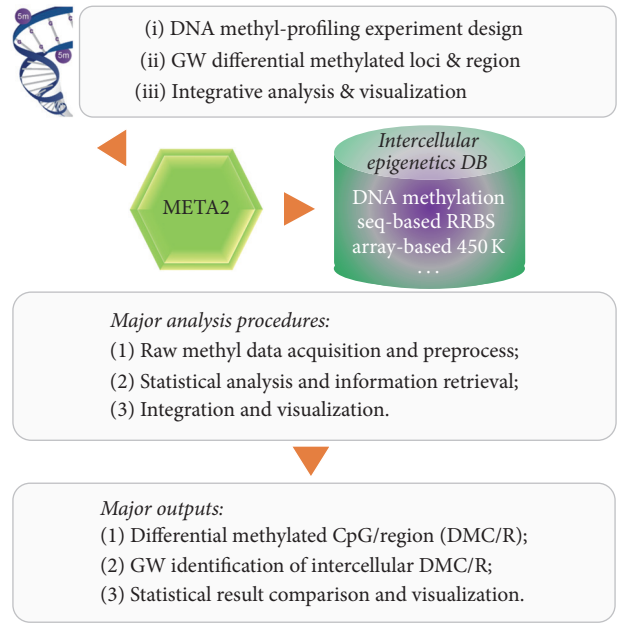


Figure 1: Schematic illustration for META2 structure and functions. META2 performs three major functional procedures, namely, DNA methylation raw data acquisition and preprocess (cell line curation and data format process), statistical analysis and information retrieval (CpG annotation, differential methylated CpG loci, and regions), and integration and results visualization (comparison and validation), together with the corresponding outputs as depicted on the bottom.

3. Comprehensive Analysis and Functional Annotation in META2

Here we propose the functions and analysis procedure in META2. As depicted in Figure 1, it mainly includes three major procedures as DNA methylation data source preprocess, information retrieval, and DNA methylation annotation. Thus, the below analysis results contain the following steps.

3.1. Statistics for Sequencing Read Coverage and Methylation Distribution. Firstly as for a high-throughput Next-Generation Sequencing (NGS) experiment, such as ChIP-seq or RRBS experiment, the necessary preprocess includes data quality check and preliminary statistical interrogation; thus biologists may gather the basic experiment quality information for following interrogation. Thus, we performed statistical calculation for the sequencing reads coverage counts (Cs and Ts) for the 1,135,337 CpG sites across the T-47D cell line.

Figure 2 illustrates the sequencing reads coverage information and DNA methylation distribution of RRBS data format for T-47D cell type. Figure 2 indicates that for both conditions' samples there exists the bimodal density pattern in the genome-wide methylation level with respect to the positive and negative strands, respectively.

And we also perform the genome-wide correlation analysis on the RRBS DNA methylation profile, and we find that there exists high correlation by pairwise comparison on control and treatment samples, with correlation coefficient from 0.94 to 0.96; see Figure 3(a).

Furthermore, we implement the region-specific analysis on those 1,135,337 CpG loci, and we find that genomic promoter and exon regions host more hypermethylated loci (≥25%) than hypomethylated loci (≤25% of methylation difference), which indicates that it is generally with hypermethylated status for most genes in T-47D cell. Together we also find that hypermethylated loci occur in CpG islands (59%) much more than hypomethylated loci (43%), which is basically consistent with the previous results; while CpG shores host 15% hypermethylated and 11% hypomethylated loci, respectively; see Figure 3(b).

Thus, based on the preprocess results, we perform the differential methylation analysis on those 1,135,337 CpG loci, and we get 3,651 statistically significant differentially methylated CpG loci (DMC), namely, absolute methylation difference ≥25% and its adjusted q-value ≤ 0.01. Those statistically significant DMCs provide meaningful clues for underlying genetic regulatory process when they are interrogated with further annotation and in silico deep analysis.

Thus, in the subsequent section, we will carry out differential methylated region (DMR) analysis on those identified

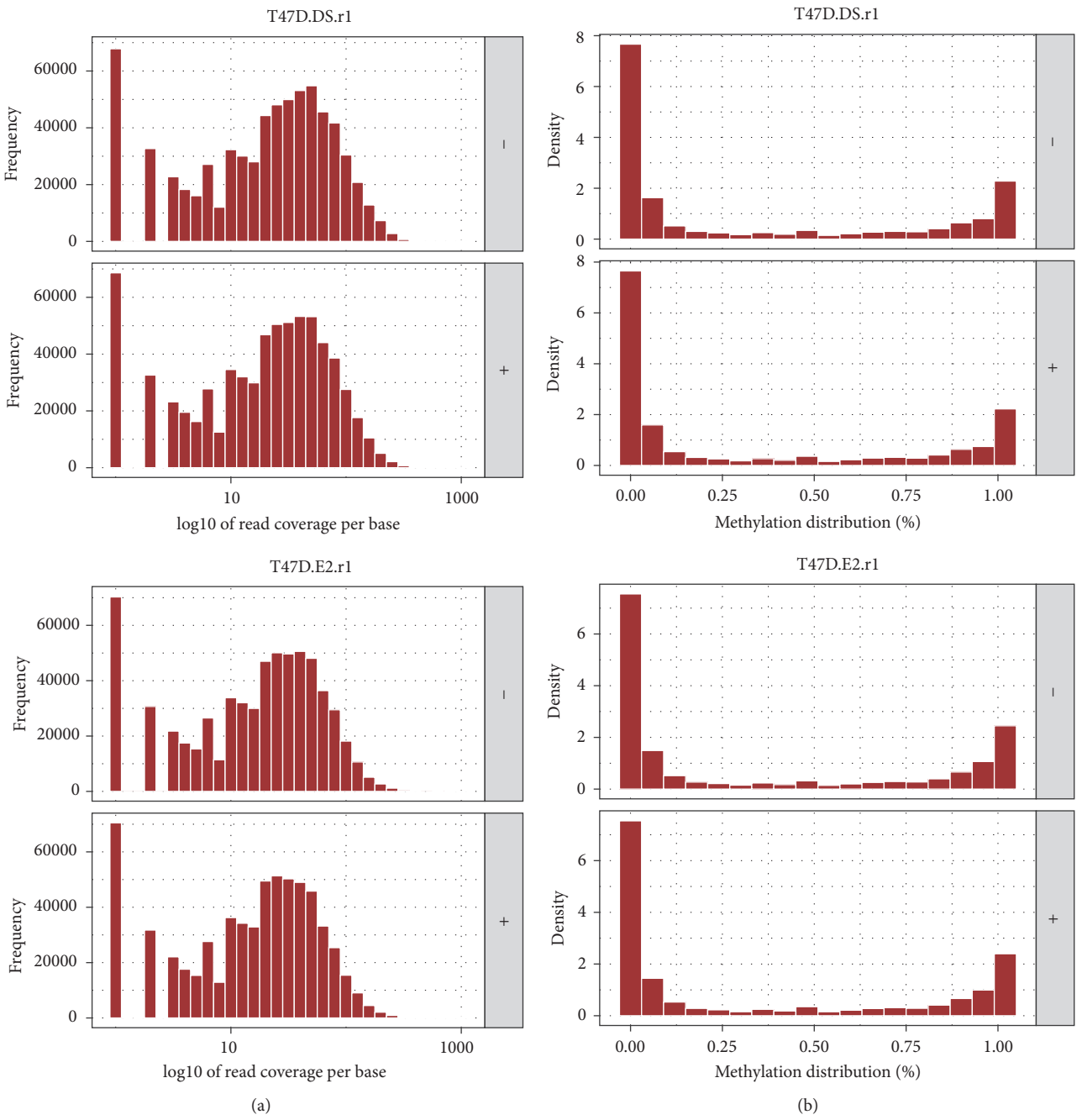


FIGURE 2: Schematic panel of statistical analysis on the raw RRBS data, that is, control (DS) versus treatment (E2) replicates with respect to positive and negative strands. (a) indicates the statistics for RRBS read coverage per base and (b) for the DNA methylation distribution for both control and treatment replicates.

DMCs, together with integrative analysis of genomic annotation.

3.2. Statistical Identification and Analysis of the Length-Specific DMRs. For consistence, we map genome-wide methylated CpG loci on each single chromosome (chr1 to chr22, chrX and chrY); see Figure 4, where each dot represents the differential CpG methylation level (in percentage, %) at the corresponding genomic position, and the line illustrates the general trend of differential methylation level across the whole chromosome.

We can see that chromosomes 1 and 2 host the longest differential methylation ranges, where the differentially methylated loci on chromosome 1 account for the most percentage (8.745%, 99,280 loci); the loci on chromosomes 17 and 19 also account for 6.506% (73,863 loci) and 6.959% (79,005 loci),

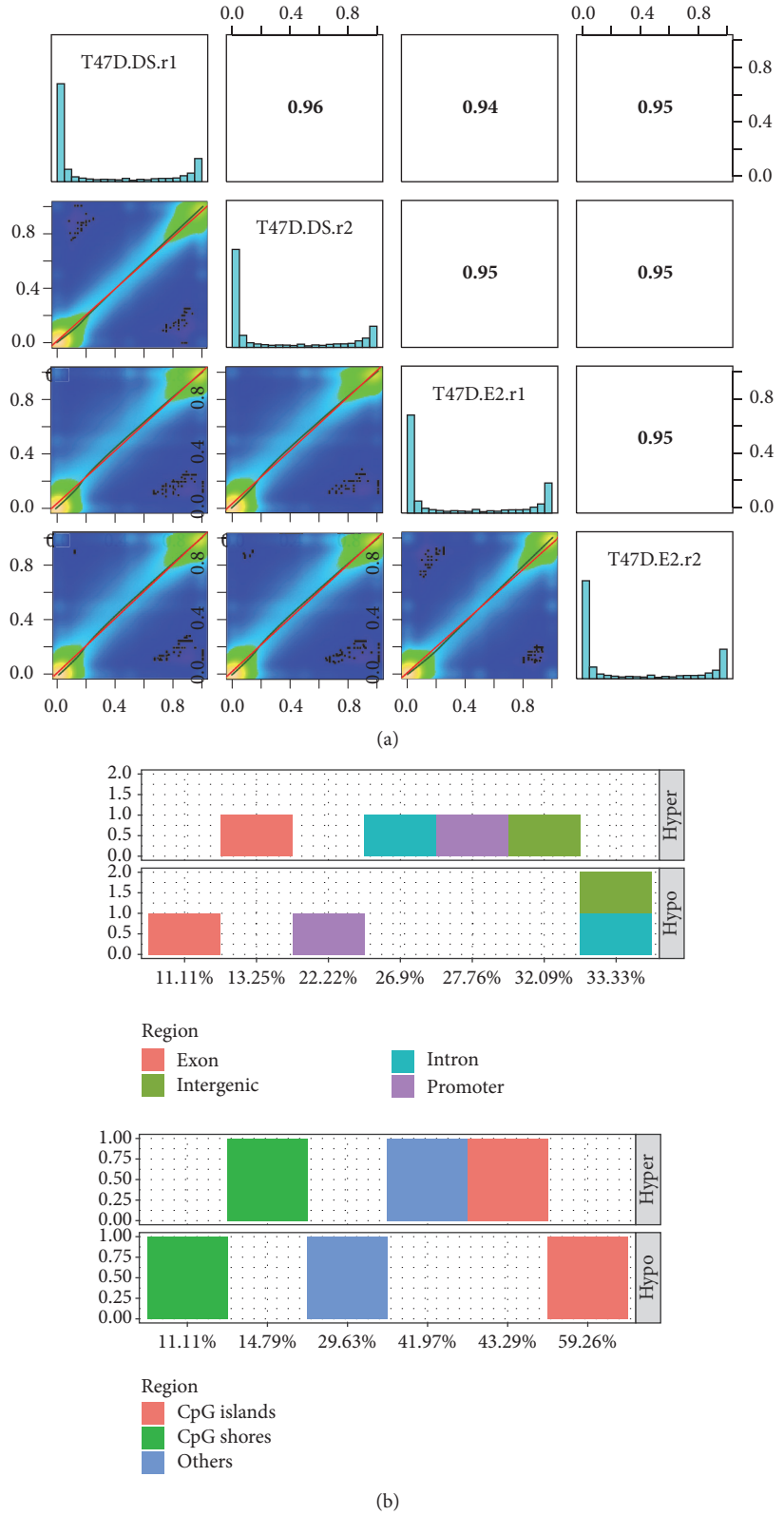


Figure 3: Schematic illustration of statistical correlation and methylation loci/region annotation analysis. (a) Correlation analysis for replicate methylation level (in percentage) from RRBS profiling technology; (b) genomic distribution for the differential methylated loci with respect to hyper- and hypomethylation status.

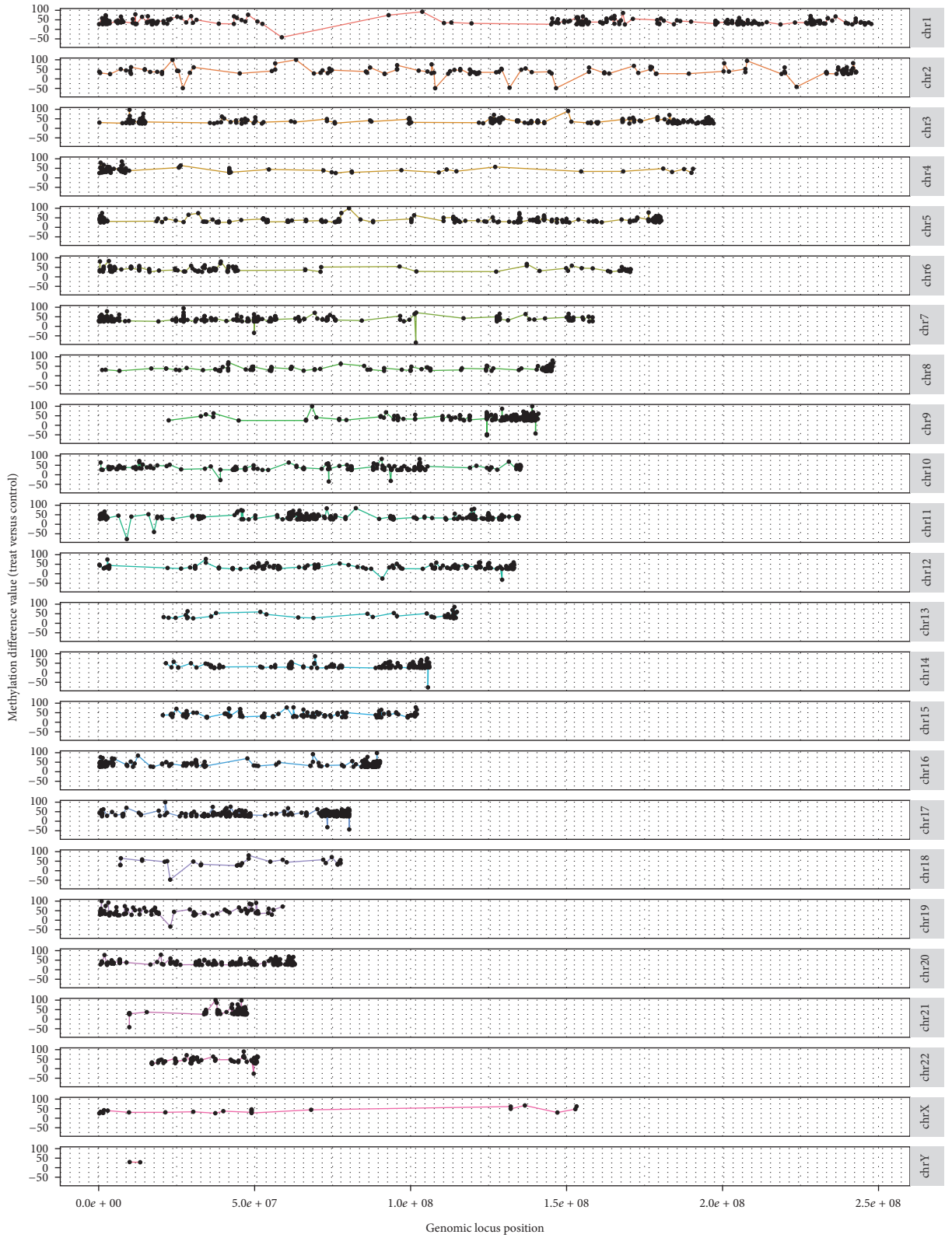


Figure 4: Genome-wide illustration for the identified differentially methylated CpG loci for T-47D cell type. Each black dot denotes the differentially methylated loci in base pair, and each curve depicts the general differential methylation trend for each chromosome.

ranking as the second and third, respectively, and those on chromosome 2 account for 6.351% (72,107 loci) of the total loci count (1,135,337).

Then we may question whether there exists any underlying biological function with those differential methylated loci, especially those with statistical significance and whether or not those loci have any clinical impact, and by which means? With those questions, META2 incorporates our self-compiled functions to interrogate the key points; firstly, META2 aims to uncover the differential methylation regions (DMRs) based on the statistically significant DMCs identified in the previous section.

We utilize a sliding window with a 10 bp length in scanning the whole genome to identify all DMR candidates. We predefine the DMR candidates that should cover more than two distinct but statistically significant DMCs to ensure the underlying biological meanings and also preset the generic DMR length up to 20,000 bp to interrogate the association between DMC count and DMR count with respect to DMR length.

Here we define two statistic indexes for measuring differential methylation level across multiple genomic loci and regions, namely, DMV.Sig for the highest differential methylation level of a significant DMC in a specific DMR and DMV.Avg for the averaged differential methylation level for all DMCs in a specific DMR. The index, DMV.Sig, aims to quantitatively identify the DMR candidates with significant methylation status across a specified range; DMV.Avg is for measuring the averaged methylation level within the DMR length under investigation.

Based on the change trends of DMV.Sig and DMV.Avg, we further utilize the information-theoretic measures, *Pearson* correlation and mutual information, for interrogating the region-specific methylation level; both of the measures intend to capture the statistical properties of dynamic variation in differential methylation profile.

Thus we calculate and illustrate the statistical association between DMR length and DMC/DMR count in Figure 5(a); Figure 5(b) depicts the statistical characteristics between mutual information and correlation analysis on DMV.Sig and DMV.Avg along with the DMR length.

In Figure 5(a), we find that along with DMR length up to 20,000 bp, DMC and DMR counts continue to increase (DMC count with a relatively sharper slope than DMR count), and for the region methylation indexes, DMV.Sig remains comparatively more stable (within the ranges 37.34% and 38.46%) than DMV.Avg, which decreases fleetly from 33.9% to 18.14%.

The analysis results above basically validate the hypothesis that DMR count statistically depends less than DMC count on the preset DMR length; meanwhile DMV.Avg depends more greatly than DMV.Sig on the preset DMR length, which indicates that the index DMV.Sig remains approximately the same for each DMR candidate regardless of the DMR length, while DMV.Avg decreases due to more and more low methylation loci covered by the subsequent DMR candidate.

Furthermore, from the results on the right panel (Figure 5), we find that both statistical curves undergo three critical transitions, that is, the shade zones A, B, and C. Zone A (DMR length at 1,500 bp) manifests the first transition at zero point for both mutual information and correlation coefficient, where DMV.Sig and DMV.Avg begin to take negative correlation; Zone B (DMR length at 5,000 bp) shows the second transition, where both mutual information and correlation for both indexes have evident inflections; Zone C (DMR length at 8,500 bp) indicates the third transition, where the negative correlation of both indexes begins to increase and their mutual information also rises up after a stretch of equilibrium. When the DMR length exceeds 12,500 bp, there is no apparent spinodal where both curves sustain the increase and decrease trends.

Based on those transition zone information, we can further annotate and decipher the underlying regulatory functions and biological meanings hereinafter.

3.3. Genomic Annotation and Identification of Genes Interacting with DMRs. Based on the three identified transition zones, we further implement genomic annotation and statistical analysis on the DMR candidates. For interrogating the inherent biology function, we emphasize the first transition zone; thus hereinafter we consider a specific class of DMRs with the maximum length less than 1,000 bp.

Figure 6 depicts those DMRs with relatively more differentially methylated loci; for illustration, we select nine typical DMRs from chromosomes 1 to 5.

From all the nine DMR distribution curves, we find that those DMRs mostly contain both hypermethylation and hypomethylation loci, while the former's count and differential methylation level are relatively more than the latter's count, which means those DMRs are generally with hypermethylation status. Subplot (g) is a good case in point, with nearly all loci being above the methylation level of 40%.

Meanwhile we further annotate the DMR candidate in subplot (b) with relatively more methylated loci than other DMR candidates. Figure 7 gives genomic annotation and analysis for the DMR in Figure 6(b), which is hosted in chromosome 1 and covers a 747 bp range from 197,743,880 to 197,744,626 bp. We acquire this DMR's methylation information and reference genome sequences from UCSC (hg19), together with protein-coding gene information.

From top to bottom panel, Figure 7 depicts this hyper-/hypomixture DMR genomic location in chromosome 1, as indicated in red line. The second panel depicts the DMR plot with 88 distinct methylation loci converged within the DMR, where those methylation loci constitute a hyper-/hypomixture methylation landscape directly impacting the underlying transcription regulatory processes for the targeted genes.

The third panel gives the reference genome sequence density within the exact DMR range; we can see that C/G content is comparatively higher than A/T in this DMR, which accords with the hypothesis that quite a few CpG sites cover the region. The bottom panel illustrated the five annotated transcripts for DENND1B at 1q31.3 (chr1:197,504,748–197,782,175), where the five transcripts generally maintain the hypermethylation status due to its range covering most hypermethylation loci with its differential methylation level up to 45%, together with a few hypomethylation loci around 10%.

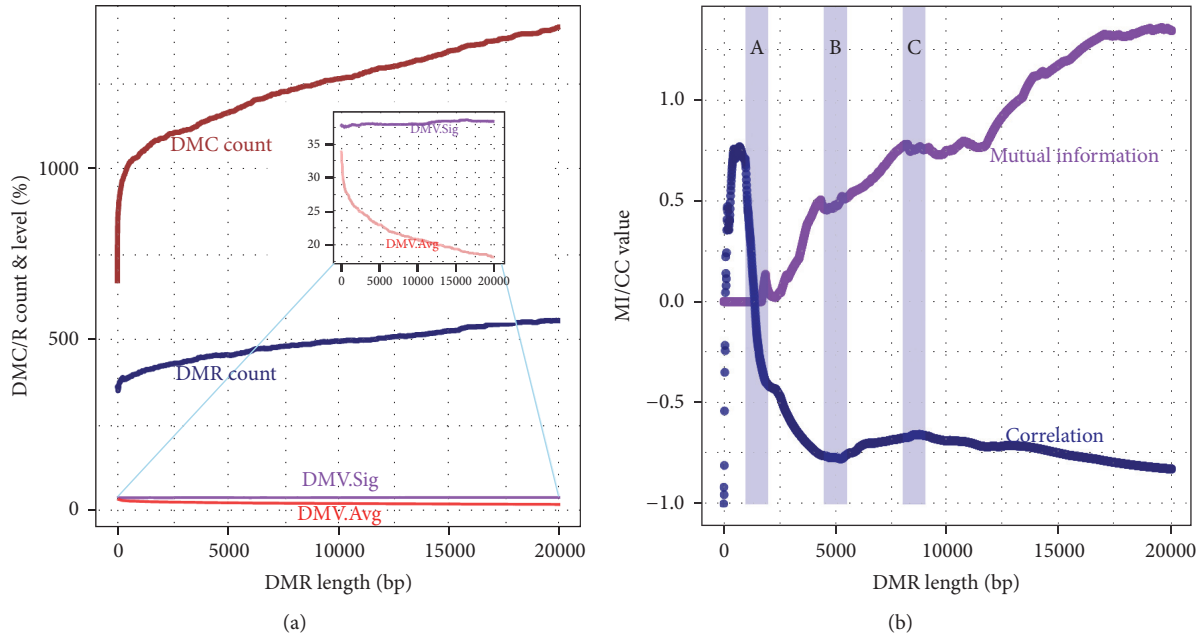


Figure 5: Statistical association analysis for DMC/R count, methylation level with respect to DMR length (a), and dynamic properties of mutual information (MI) and correlation coefficient (CC) with respect to DMR length (b).

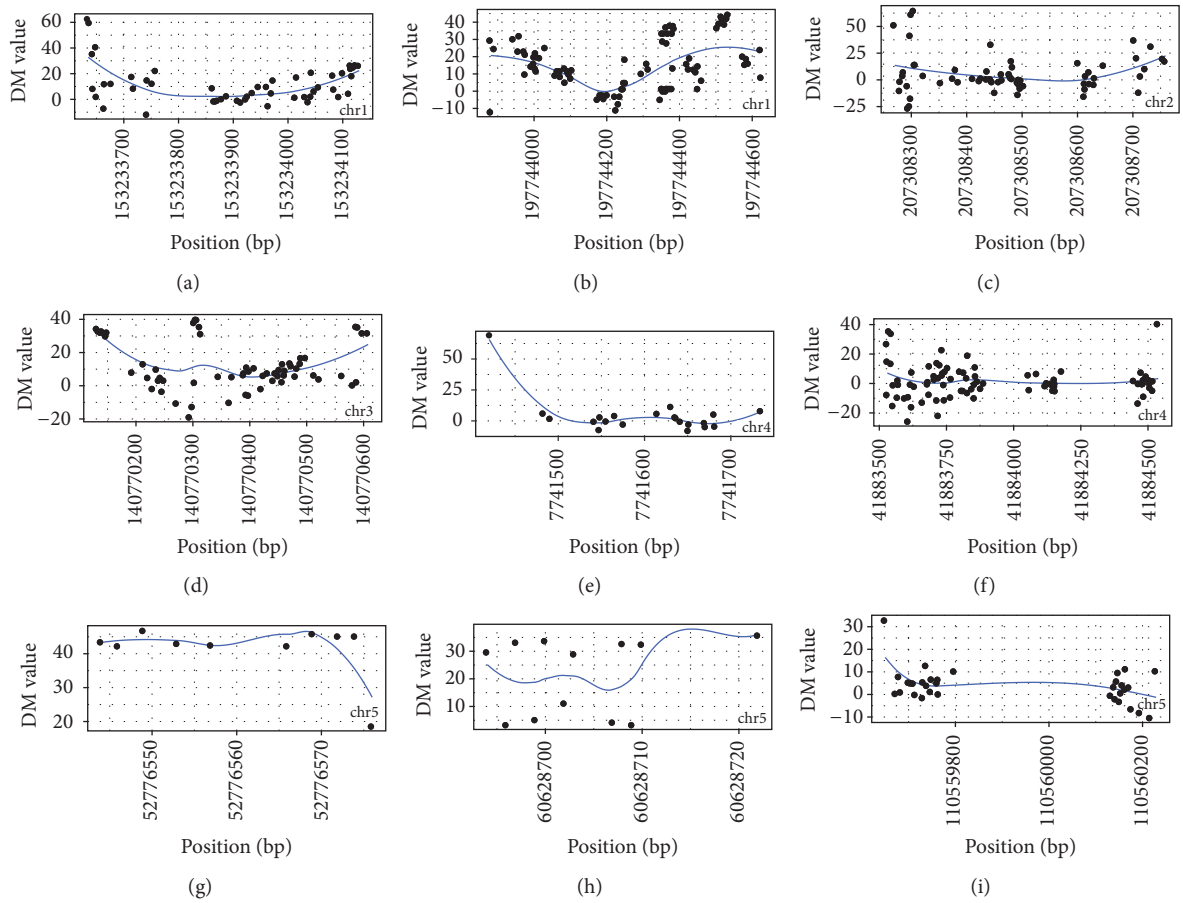


Figure 6: Schematic illustration for typical DMR candidates in chromosomes 1 to 5. The black dots denote the differential methylation value for each loci, and the blue lines represent the fitted DMR curves at each specific region. (a–i) subplots; (a) and (b) depict DMRs for chromosome 1, (c) for chromosome 2, (d) for chromosome 3, (e) and (f) for chromosome 4, and (g), (h), and (i) for chromosome 5.

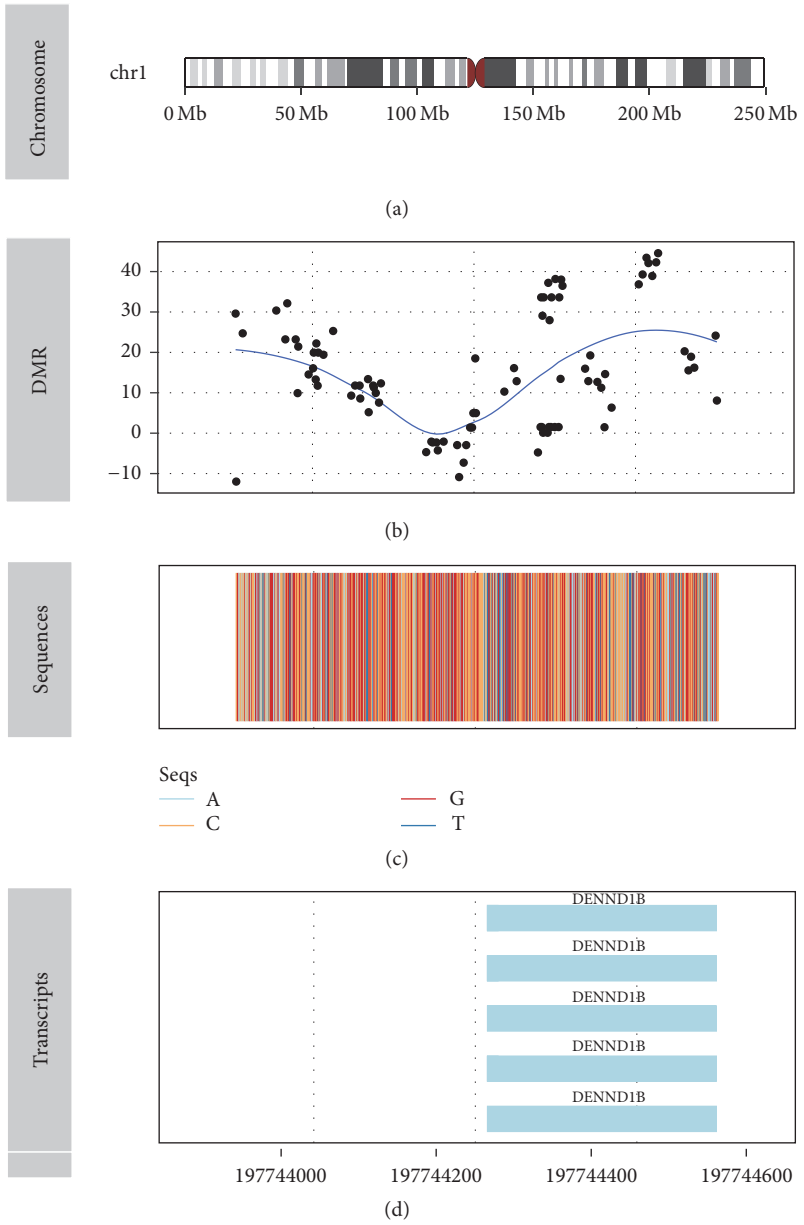


Figure 7: Schematic illustration of annotation and analysis for one typical DMR (chr1:197,743,880-197,744,626). From top to bottom, each panel depicts chromosome band, DMR and loci distribution, reference genome sequence (hg19), and annotated transcript information for DENND1B at 1q31.3.

4. Materials and Methods

4.1. Reduced Representation Bisulfite Sequencing (RRBS). Reduced representation bisulfite sequencing, or RRBS, is a large-scale random approach for analyzing and comparing genomic methylation patterns. BglII restriction fragments of 500–600 bp sized selected, together with adapters assembled, were further treated with bisulfite, PCR amplification, and clone and finally sequenced to target methylated CpG sites. From the converted and unconverted read counts at each CpG, the sample coverage and methylation level (in percentage) can be acquired [11–13].

4.2. Annotation for the Significant Differentially Methylated CpG Sites (DMC). Here we selected one cell line (T-47D, control versus treatment) as the benchmark cell line, and the annotation results are further filtered based on the lifted methylation difference threshold (at least 25% methylation difference for the paired groups).

4.3. Statistical Analysis for the Differentially Methylated Regions. We identified 16,277 DMR candidates from all the DMCs, with the adjusted q-value ≤ 0.01, CpG base methylation difference cutoff, 25, and DMR mean methylation difference cutoff, 20. Within those candidates, 8,936 entries present

hypermethylated and 7,341 hypomethylated status. With the lifted thresholds, namely, adjusted q-value ≤ 0.001 and differentially methylated CpG base count ≥ 5, we further detected 7,537 significant DMRs (Sig-DMRs), where 3,512 entries are significantly hypermethylated-DMRs (Sig-Hyper-DMRs) and 4,025 significantly hypomethylated-DMRs (Sig-Hypo-DMRs).

4.4. Tools Used in the Raw RRBS Curation and Statistical Analysis. Bowtie2 [20] was used to align sequencing reads; SAMtools [21] and BAMtools [22] were used to process the aligned sequencing reads, and methylKit [23] and META2 package were used to analyze the raw RRBS data; limma and DEseq were used in differential analysis of DNA methylation loci [24].

4.5. Generalized Mutual Information. Given two discrete random variables X and Y, the mutual information is defined as

$$\mathrm{MI}(X,Y) = H(X) + H(Y) - H(X,Y), \quad (1)$$

where $H(X)$ and $H(Y)$ are the entropy measures for X and Y and $H(X,Y)$ is the joint entropy between variables X and Y, respectively. The mutual information measure is adopted for association identification within the analysis section.

5. Conclusion

Here we present a developed toolkit, META2, for DNA methylation annotation and analysis, which aims to implement the intercellular analysis on differentially methylated loci and regions. META2 contains multiple versatile functions for annotating and analyzing DNA methylation, such as the profiling data by RRBS and other high-throughput technology.

By utilizing the toolkit on the real RRBS data from ENCODE, we performed statistical correlation and genomic loci/region annotation for all the identified differentially methylated CpGs, or DMC candidates; we further implemented statistical association analysis for DMC/R count and methylation level with respect to the preset DMR length and revealed the dynamic properties of mutual information and correlation coefficient with respect to DMR length; thus we detected three major transition zones, which provide statistical clues for further biological function investigation.

Our work provides a versatile and comprehensive analysis toolkit for epigenetic research and clinical study, especially for the genome-wide biomedical analysts, to interrogate and validate their hypothesis in an efficient and uniform way.

Further anticipated improvements including statistical annotation and analysis functions concerning cell or tissue-specific and pancancer analysis functions will be consolidated into the toolkit; thus it constitutes a versatile and evolving toolkit for biologists to easily adopt in their research.

Additional Points

Availability. RRBS profiling data for T47D is available at ENCODE; META2 and its corresponding test data and manual were deposited at https://github.com/gladex/META2.

Competing Interests

The author declares that there are no competing interests regarding the publication of this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of Jiangsu, China (BE2016655 and BK20161196), the Fundamental Research Funds for China Central Universities (2016B08914), and Changzhou Science & Technology Program (CE20155050). This work made use of the resources supported by the NSFC-Guangdong Mutual Funds for Super Computing Program (2nd Phase), and the Open Cloud Consortium- (OCC-) sponsored project resource, supported in part by grants from Gordon and Betty Moore Foundation and the National Science Foundation (USA) and major contributions from OCC members.

References

[1] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson et al., "The cancer genome Atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, pp. 1113–1120, 2013.

[2] V. N. Kristensen, O. C. Lingjærde, H. G. Russnes, H. K. M. Vollan, A. Frigessi, and A.-L. Børresen-Dale, "Principles and methods of integrative genomic analyses in cancer," *Nature Reviews Cancer*, vol. 14, no. 5, pp. 299–313, 2014.

[3] T. Witte, C. Plass, and C. Gerhauser, "Pan-cancer patterns of DNA methylation," *Genome Medicine*, vol. 6, no. 8, article 66, pp. 1–18, 2014.

[4] M. D. M. Leiserson, F. Vandin, H.-T. Wu et al., "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes," *Nature Genetics*, vol. 47, no. 2, pp. 106–114, 2015.

[5] The Cancer Genome Atlas Research Networ, "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, no. 7417, pp. 519–525, 2012.

[6] The Cancer Genome Atlas Research Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[7] A. Meissner, T. S. Mikkelsen, H. Gu et al., "Genome-scale DNA methylation maps of pluripotent and differentiated cells," *Nature*, vol. 454, no. 7205, pp. 766–770, 2008.

[8] H. Heyn, F. Carmona Javier, A. Gomez et al., "DNA methylation profiling in breast cancer discordant identical twins identifies DOK7 as novel epigenetic biomarker," *Carcinogenesis*, vol. 34, no. 1, pp. 102–108, 2013.

[9] R. Lister, E. A. Mukamel, J. R. Nery et al., "Global epigenomic reconfiguration during mammalian brain development," *Science*, vol. 341, no. 6146, Article ID 1237905, 2013.

[10] J. T. Bell, A. K. Loomis, L. M. Butcher et al., "Differential methylation of the TRPA1 promoter in pain sensitivity," *Nature Communications*, vol. 5, article no. 2978, 2014.

[11] A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch, "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis," *Nucleic Acids Research*, vol. 33, no. 18, pp. 5868–5877, 2005.

[12] H. Guo, P. Zhu, L. Yan et al., "The DNA methylation landscape of human early embryos," *Nature*, vol. 511, no. 7511, pp. 606–610, 2014.

[13] A. Kundaje, W. Meuleman, J. Ernst et al., "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–330, 2015.

[14] J. Maksimovic, L. Gordon, and A. Oshlack, "SWAN: subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips," *Genome Biology*, vol. 13, no. 6, 2012.

[15] M. Bibikova, B. Barnes, C. Tsan et al., "High density DNA methylation array with single CpG site resolution," *Genomics*, vol. 98, no. 4, pp. 288–295, 2011.

[16] R. Lister, M. Pelizzola, R. H. Dowen et al., "Human DNA methylomes at base resolution show widespread epigenomic differences," *Nature*, vol. 462, no. 7271, pp. 315–322, 2009.

[17] M. J. Ziller, H. Gu, F. Müller et al., "Charting a dynamic DNA methylation landscape of the human genome," *Nature*, vol. 500, no. 7463, pp. 477–481, 2013.

[18] A. Blattler, L. Yao, H. Witt et al., "Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes," *Genome Biology*, vol. 15, article 469, 2014.

[19] C. J. Kemp, J. M. Moore, R. Moser et al., "CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer," *Cell Reports*, vol. 7, no. 4, pp. 1020–1029, 2014.

[20] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, 2012.

[21] H. Li, B. Handsaker, A. Wysoker et al., "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[22] D. W. Barnett, E. K. Garrison, A. R. Quinlan, M. P. Strömberg, and G. T. Marth, "Bamtools: A C++ API and toolkit for analyzing and managing BAM files," *Bioinformatics*, vol. 27, no. 12, Article ID btr174, pp. 1691–1692, 2011.

[23] A. Akalin, M. Kormaksson, S. Li et al., "methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles," *Genome Biology*, vol. 13, no. 10, article R87, 2012.

[24] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, article R106, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 5621313, 8 pages
http://dx.doi.org/10.1155/2016/5621313

Research Article

Optimal Control Model of Tumor Treatment with Oncolytic Virus and MEK Inhibitor

Yongmei Su, Chen Jia, and Ying Chen

School of Mathematics and Physics, University of Science and Technology Beijing, Beijing, China

Correspondence should be addressed to Yongmei Su; suym71@ustb.edu.cn

Received 21 September 2016; Accepted 27 November 2016

Academic Editor: Tun-Wen Pai

Tumors are a serious threat to human health. The oncolytic virus is a kind of tumor killer virus which can infect and lyse cancer cells and spread through the tumor, while leaving normal cells largely unharmed. Mathematical models can help us to understand the tumor-virus dynamics and find better treatment strategies. This paper gives a new mathematical model of tumor therapy with oncolytic virus and MEK inhibitor. Stable analysis was given. Because mitogen-activated protein kinase (MEK) can not only lead to greater oncolytic virus infection into cancer cells, but also limit the replication of the virus, in order to provide the best dosage of MEK inhibitors and balance the positive and negative effect of the inhibitors, we put forward an optimal control problem of the inhibitor. The optimal strategies are given by theory and simulation.

1. Introduction

Tumors are a serious threat to human health, and chemotherapy and radiotherapy may not only kill cancer cells, but also damage human body normal cells at the same time [1]. The oncolytic virus is a kind of tumor killer virus which can infect and lyse cancer cells and spread through the tumor, while leaving normal cells largely unharmed [2]. When oncolytic viruses are inoculated into a cancer patient or directly injected into a tumor, these viruses will spread throughout the tumor and infect tumor cells. The viruses can be replicated in the infected tumor cells. When an infected tumor cell is lysed, it can burst out a mass of new oncolytic viruses. Then, these new viruses can infect much more neighboring tumor cells [3].

Experiments using oncolytic viruses such as adenovirus, CN706 [4], and ONYX-15 [5] in animal tumors show that these viruses are nontoxic and infect tumor cells specifically. Now, treatment of cancer with oncolytic virus has been clinically tested [6–8]. This treatment of cancer with oncolytic viruses has been explored by clinicians [9–11].

In recent years, in order to understand the cancer-virus dynamics and find better treatment strategies, some mathematical models have been set up [12–19]. Tian proposed a mathematical model to describe the development of a growing tumor and an oncolytic virus population as follows [18]:

$$\begin{aligned}\frac{dx}{dt} &= \lambda x\left(1-\frac{x+y}{K}\right)-\beta xv,\\ \frac{dy}{dt} &= \beta xv-\delta y,\\ \frac{dv}{dt} &= b\delta y-\beta xv-\gamma v,\end{aligned} \tag{1}$$

where variables x, y, and v stand for the population of uninfected cells, infected tumor cells, and oncolytic viruses, respectively. The coefficient β represents the infection of the virus. The tumor growth is modeled by logistic growth, and K is the maximal tumor size. λ is the per capita tumor growth rate. δ means the lysis rate of the infected tumor cells. b represents the burst size of new viruses coming out from the lysis of an infected tumor cell. γ represents the death rate of the virus.

It was shown that when the threshold $b < 1+\gamma/(K\beta)$, the equilibrium solution $(K, 0, 0)$ is globally asymptotically stable [18], indicating that the oncolytic virus therapy finally has no effect. Obviously, the smaller the value of K, the more easily $b < 1+\gamma/(K\beta)$ holds. Since K represents the total number of tumor cells, smaller tumors may be more resistant to the

treatment by oncolytic virus than large ones, which should be a contradiction. In [19], by replacing βxv with $xv/(x+y+\varepsilon)$, we proposed the model

$$\begin{aligned}\frac{dx}{dt} &= \lambda x\left(1-\frac{x+y}{K}\right)-\beta\frac{xv}{x+y+\varepsilon},\\ \frac{dy}{dt} &= \beta\frac{xv}{x+y+\varepsilon}-\delta y,\\ \frac{dv}{dt} &= b\delta y-\beta\frac{xv}{x+y+\varepsilon}-\gamma v.\end{aligned} \tag{2}$$

The meanings of variables x, y, and v and parameters λ, β, δ, γ, b, and K are the same as those in model (1), and ε is positive and sufficiently small. The threshold obtained by our model is $b < 1+(\gamma/\beta)(1+\varepsilon/K)$, which is almost independent of K when ε is sufficiently small.

On the other hand, all the above papers did not consider coxsackie-adenovirus receptor (CAR). In fact, CAR is a main receptor when oncolytic viruses enter into tumor cells [20–22]. The successful entry of viruses into cancer cells is related to the presence of CAR. When oncolytic viruses infect the tumor cells, firstly, they combine with the CAR and are absorbed into the cells.

Mitogen-activated protein kinase (MEK) inhibitors have been shown to promote CAR expression and could increase oncolytic viruses infection into tumor cells. But MEK inhibitors may also limit the replication of viruses [23–25], which will affect the treatment by oncolytic virus. With the function of MEK, [25] gave a model:

$$\begin{aligned}\frac{dx}{dt} &= \rho x(1-u)-dx-\frac{\beta zxv}{1+\varepsilon v},\\ \frac{dy}{dt} &= \frac{\beta zxv}{1+\varepsilon v}-dy-a(1-u)y,\\ \frac{dv}{dt} &= k(1-u)y-bv,\\ \frac{dz}{dt} &= \eta u(p-z)-cz.\end{aligned} \tag{3}$$

The variables x, y, and v have the same meanings as those in model (2); z represents the average expression level of CAR on the surface of the cells. The intensity of MEK inhibitor application is captured in the parameter u, $u \in [0,1]$. If $u = 0$, there is no MEK inhibitor application, and the CAR expression level will gradually decline. If $u = 1$, the MEK inhibitor has the maximum possible effect. The model assumes that exponential growth can be slowed down by the inhibitor with expression $1-u$. CAR grow at the rate of $g(p-z)$ and become extinct at the rate of c.

Based on models (2) and (3), we establish the following mathematical model:

$$\begin{aligned}\frac{dx}{dt} &= (1-u)rx\left(1-\frac{x+y}{K}\right)-\frac{\beta xvz}{x+y+\varepsilon},\\ \frac{dy}{dt} &= \frac{\beta xvz}{x+y+\varepsilon}-(1-u)\delta y,\\ \frac{dv}{dt} &= b(1-u)\delta y-\frac{\beta xvz}{x+y+\varepsilon}-\alpha v,\\ \frac{dz}{dt} &= gu(p-z)-cz.\end{aligned} \tag{4}$$

The variables x, y, v, and z have the same meanings as those in model (3). The parameters λ, β, δ, γ, b, and K are the same as those of (1). The parameter u has the same meaning as that in model (3). All the parameters are strictly positive.

Since the use of MEK inhibitors not only results in enhanced oncolytic virus entry into the tumor cells, but also renders infected cells temporarily unable to produce viruses, the maximum dosage of MEK use may not result in the best treatment effect, so the optimal control-based schedules of MEK inhibitor application should be studied. The optimal MEK inhibitor application strategy can increase the efficacy of this treatment in an economical fashion. So, first, in this paper, we let the control variable u be a constant; a stability analysis of our model is conducted, and then the optimal control strategy is discussed; we also compare the optimal control with constant control by simulation.

2. Materials and Methods

2.1. Stability Analysis. System (4) always has two equilibrium points:

$$\begin{aligned}E_0 &= \left(0,0,0,\frac{gup}{gu+c}\right),\\ E_1 &= \left(K,0,0,\frac{gup}{gu+c}\right).\end{aligned} \tag{5}$$

If ε is sufficiently small, when $b > 1+(\alpha(K+\varepsilon)/\beta K)((gu+c)/gup)$, the third steady state $E_2 = (x_2, y_2, v_2, z_2)$ exists in which

$$\begin{aligned}x_2 &= \frac{\Delta+\sqrt{\Delta^2+4(b-1)K\alpha\beta z_2 r(1-u)^2\delta\varepsilon}}{2(b-1)\beta z_2 r(1-u)},\\ y_2 &= \left[\frac{(b-1)\beta z_2}{\alpha}-1\right]x_2-\varepsilon,\\ v_2 &= \frac{(b-1)(1-u)\delta}{\alpha}y_2,\\ z_2 &= \frac{gup}{gu+c}.\end{aligned} \tag{6}$$

Here,

$$\begin{aligned}\Delta = {} & (1-u)r\alpha(K+\varepsilon)-K(b-1)(1-u)\delta\beta z_2\\ & +K\alpha(1-u)\delta.\end{aligned} \tag{7}$$

It should be noted that $b > 1 + (\alpha(K+\varepsilon)/\beta K)((gu+c)/gup)$ is equivalent to $(b-1)\beta z_2/\alpha - 1 > 0$ to ensure that $y_2 > 0$ when ε is sufficiently small.

The Jacobi matrix at point E_0 is

$$J|_{E_0} = \begin{pmatrix} r(1-u) & 0 & 0 & 0 \\ 0 & -(1-u)\delta & 0 & 0 \\ 0 & b\delta & -\alpha & 0 \\ 0 & 0 & 0 & -gu-c \end{pmatrix}. \quad (8)$$

Obviously, $\mu_1 = r(1-u) > 0$ is one eigenvalue of $J|_{E_0}$ which means E_0 is unstable. The unstable result of E_0 seems consistent with the biological meaning that, without viruses and infected tumor cells, the tumor will grow from an initial small value around E_0.

As for equilibrium point E_1, we have the following theorem.

Theorem 1. *When $b < 1+(\alpha(K+\varepsilon)/\beta K)\cdot((gu+c)/gup)$, E_1 is locally asymptotically stable. When $b > 1+(\alpha(K+\varepsilon)/\beta K)((gu+c)/gup)$, E_1 is unstable.*

Proof. At the equilibrium point E_1, the Jacobi matrix is

$$J|_{E_1} = \begin{pmatrix} -(1-u)r & -(1-u)r & -H & 0 \\ 0 & -(1-u)\delta & H & 0 \\ 0 & b(1-u)\delta & -\alpha - H & 0 \\ 0 & 0 & 0 & -gu-c \end{pmatrix}, \quad (9)$$

where $H = \beta K gup/(K+\varepsilon)(gu+c)$. The eigenvalues of $J|_{E_1}$ are $\mu_1 = -(1-u)r$, $\mu_2 = -gu-c$,

$$\mu_{3,4} = \frac{-\left((1-u)\delta + \alpha + \Delta_2\right) \pm G}{2}. \quad (10)$$

Here, $G = \sqrt{(\delta(1-u) - \alpha - \Delta_2)^2 + 4\Delta_2 b\delta(1-u)}$ in which $\Delta_2 = \beta K ugp/(K+\varepsilon)(gu+c)$.

When $b < 1 + (\alpha(K+\varepsilon)/\beta K)((gu+c)/gup)$, we have

$$\begin{aligned} G &= \left((1-u)\delta - \alpha - \Delta_2\right)^2 + 4\Delta_2 b(1-u)\delta \\ &< \left((1-u)\delta + \alpha + \Delta_2\right)^2, \end{aligned} \quad (11)$$

which ensure that μ_3 and μ_4 are negative, so E_1 is locally asymptotically stable.

Similarly, when $b > 1 + \alpha(K+\varepsilon)(gu+c)/\beta K gup$, μ_3 is positive, and E_1 is unstable.

Actually, we can prove that the equilibrium solution E_1 is globally asymptotically stable when $b < 1 + \alpha(K+\varepsilon)(gu+c)/\beta K gup$. But we need to show the boundness of system (4). From the first two equations, we obtain

$$\begin{aligned} &\frac{d\left(x(t)+y(t)\right)}{dt} \\ &= (1-u)rx(t)\left(1 - \frac{x(t)+y(t)}{K}\right) - (1-u)\delta y(t) \\ &\le r\left(x(t)+y(t)\right)\left(1 - \frac{x(t)+y(t)}{K}\right). \end{aligned} \quad (12)$$

By the comparison principle, we can obtain $\lim_{t\to\infty}\sup(x(t) + y(t)) \le K$.

From the third equation of (4), we can have

$$\begin{aligned} \frac{dv}{dt} &\le (1-u)b\delta K - \beta\frac{xvz}{x+y+\varepsilon} - \alpha v \\ &\le (1-u)b\delta K - \alpha v. \end{aligned} \quad (13)$$

It is easily shown that $v(t) \le (1-u)b\delta K/\alpha$.

Similarly, from

$$\frac{dz}{dt} = gu(p-z) - cz = gup\left(1 - \frac{gu+c}{gup}z\right), \quad (14)$$

we can get that $z(t) \le gup/(gu+c)$ holds. □

Theorem 2. *When $b < 1 + (\alpha(K+\varepsilon)/\beta K)\cdot((gu+c)/gup)$, E_1 is globally asymptotically stable.*

Proof. Consider the Lyapunov function $V = y + (1/b)v$; the derivative along a solution is given by

$$\begin{aligned} \dot{V} &= \dot{y} + \frac{1}{b}\dot{v} \\ &= \beta\frac{xvz}{x+y+\varepsilon} - (1-u)\delta y \\ &\quad + \frac{1}{b}\left(b(1-u)\delta y - \beta\frac{xvz}{x+y+\varepsilon} - \alpha v\right) \\ &= \left(1 - \frac{1}{b}\right)\frac{\beta xvz}{x+y+\varepsilon} - \frac{\alpha}{b}v. \end{aligned} \quad (15)$$

Since $0 < x \le K$, $0 < x + y \le K$, we have $xK + x\varepsilon \le xK + K\varepsilon + Ky$, which implies

$$\frac{x}{x+y+\varepsilon} \le \frac{K}{K+\varepsilon}, \quad (16)$$

and because $z \le gup/(gu+c)$, therefore,

$$\begin{aligned} \dot{V} &= \left(1 - \frac{1}{b}\right)\frac{\beta xvz}{x+y+\varepsilon} - \frac{\alpha}{b}v \\ &\le \left(1 - \frac{1}{b}\right)\frac{\beta Kv}{K+\varepsilon}\frac{gup}{gu+c} - \frac{\alpha}{b}v \\ &= \frac{\beta(b-1)Kgp - \alpha(K+\varepsilon)(gu+c)}{b(K+\varepsilon)(ug+c)}v. \end{aligned} \quad (17)$$

When $b < 1 + (\alpha(K+\varepsilon)/\beta K) \cdot ((gu+c)/gup)$, we can have $\dot{V} \le 0$.

Let $E = \{(x, y, v, z) \mid \dot{V} = 0\}$; it is clear that $E \subset \{(x, y, v, z) \mid v = 0\}$. Let M be the largest positively invariant subset of the set E; by the third equation of system (4), we can know that $y(t) = 0$, so $M = \{(x, y, v) \mid y = 0, v = 0\}$. By LaSalle invariance principal [26], we know

$$\lim_{t\to\infty} y(t) = 0,$$
$$\lim_{t\to\infty} v(t) = 0. \tag{18}$$

So, the limit equation of system (4) is

$$\frac{dx}{dt} = (1-u)\, rx\left(1 - \frac{x}{K}\right),$$
$$\frac{dz}{dt} = gu(p - z) - cz. \tag{19}$$

Therefore, $x(t) \to K$, $z \to gup/(gu+c)$ when $t \to \infty$. So, E_1 is globally attractive; note that $b < 1 + (\alpha(K+\varepsilon)/\beta K) \cdot ((gu+c)/gup)$ can also ensure the local asymptotical stability of E_1, so we can know that E_1 of system (4) is globally asymptotically stable when $b < 1 + (\alpha(K+\varepsilon)/\beta K) \cdot ((gu+c)/gup)$.

Although we can prove the global asymptotical stability of E_1, we would not want this to happen, because the global asymptotical stability means the therapy does not have any effect. When $b > 1 + (\alpha(K+\varepsilon)/\beta K)((gu+c)/gup)$ holds, the coexistent steady state $E_2 = (x_2, y_2, v_2, z_2)$ exists, but it is difficult to give the stable analysis of E_2, so we just give some simulations about it.

We choose $\beta = 0.2\,\text{day}^{-1}$, $\varepsilon = 0.009$, $\delta = 0.5\,\text{day}^{-1}$, $r = 6\,\text{cells day}^{-1}$, $\alpha = 0.5\,\text{day}^{-1}$, $p = 10$, $g = 0.1\,\text{day}^{-1}$, $c = 0.5\,\text{day}^{-1}$, and $K = 9 \times 10^8$ cells.

The initial condition is $x0 = (6\times10^7,\ 0,\ 5\times10^4,\ 4\times10^2)$, where the unit of each is cells.

We choose $b = 10\,\text{cell}^{-1}\,\text{day}^{-1}$, $b = 14\,\text{cell}^{-1}\,\text{day}^{-1}$, respectively, and $b > 1+(\alpha(K+\varepsilon)/\beta K)((gu+c)/gup) = 2.6389$ all hold; the simulation results are shown in Figures 1 and 2.

The simulation results show that oncolytic virus therapy may keep the tumor stable at some level as shown in Figure 1 or keep oscillating at a certain range as shown in Figure 2. Since the cured equilibrium is always unstable, just from our model, we could not give the condition that ensures the tumor can be cured by oncolytic virus therapy, but if we choose appropriate u which satisfies

$$b > 1 + \frac{\alpha(K+\varepsilon)}{\beta K}\frac{gu+c}{gup}, \tag{20}$$

the simulation shows that oncolytic virus therapy can prevent the tumor from getting worse and worse. Some other therapy methods should be combined to cure the tumor. □

2.2. The Optimal Control of MEK. In the simulation of Figures 1 and 2, we choose the control u as constant. Since the use of MEK inhibitors not only results in enhanced oncolytic virus entry into the cells, but also renders infected

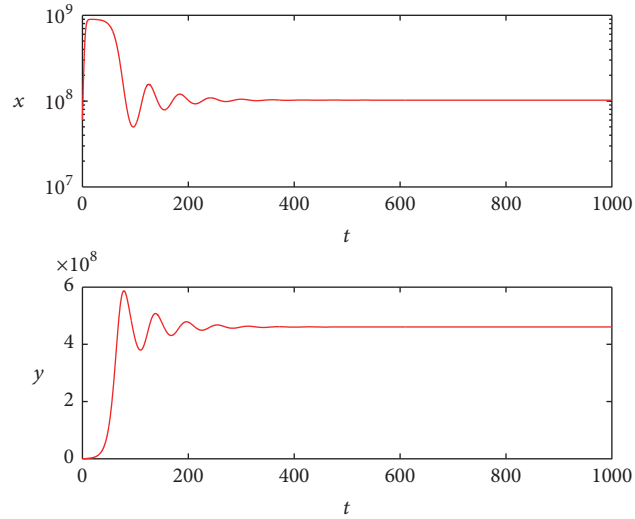


Figure 1: State dynamics for uninfected and infected tumor cells when $b = 10$.

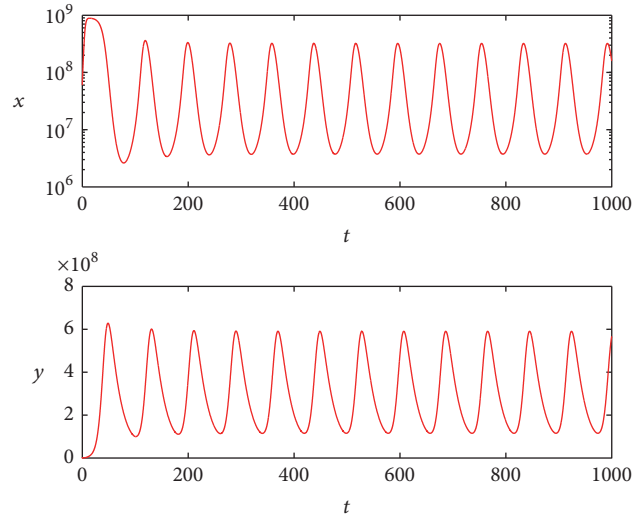


Figure 2: State dynamics for uninfected and infected tumor cells when $b = 14$.

cells temporarily unable to produce viruses, how to use the MEK inhibitors optimally should be studied. In model (4), the function of MEK inhibitors was embodied by parameter u; we use it as the control variable. The control goal is not only to formulate an objective functional which lowers the levels of tumor cells during and at the end of therapy, but also to minimize the cost of MEK, so the objective function is defined as follows:

$$J(u) = \frac{1}{2}a_{11}x^2 + \frac{1}{2}a_{22}y^2 + \frac{1}{2}\int_{t_0}^{t_f}\left(b_{11}x^2 + b_{22}y^2 + b_{33}v^2 + c_{11}u^2\right)dt, \tag{21}$$

where t_0 represents the beginning time of the treatment and t_f represents the terminal time of the treatment.

a_{11}, a_{22}, b_{11}, b_{22}, b_{33}, and c_{11} represent the cost coefficients for the variables, respectively.

For convenience, we define the state vector $X = (x, y, v, z)^T$; system (4) can be written as

$$\begin{aligned} \dot{X} &= f\left(X\left(T\right), u\left(t\right), t\right), \\ X\left(t_0\right) &= X_0. \end{aligned} \tag{22}$$

And the corresponding cost function is defined as follows:

$$J = \frac{1}{2}X^T AX + \frac{1}{2}\int_{t_0}^{t_f}\left(X^T BX + Cu^2\right)dt. \tag{23}$$

Here,

$$\begin{aligned} A &= \begin{pmatrix} a_{11} & 0 & 0 & 0 \\ 0 & a_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \\ A &= \begin{pmatrix} b_{11} & 0 & 0 & 0 \\ 0 & b_{22} & 0 & 0 \\ 0 & 0 & b_{33} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \\ C &= c_{11}. \end{aligned} \tag{24}$$

Based on the dynamic constraint $f(X(T), u(t), t)$ and the Lagrangian $L(X(t), u(t), t)$, the Hamiltonian is as follows:

$$\begin{aligned} H\left(X\left(t\right), u\left(t\right), t\right) &= L\left(X\left(t\right), u\left(t\right), t\right) \\ &\quad + \lambda^T f\left(X\left(t\right), u\left(t\right), t\right) \end{aligned} \tag{25}$$

where $L(X(t), u(t), t) = (1/2)(X^T BX + c_{11}u^2)$.

Using Pontryagin's minimum principle, the necessary conditions are given as follows:

(1^0)

$$\begin{aligned} \dot{X} &= f\left(X\left(t\right), u\left(t\right), t\right) \\ &= \begin{cases} \dfrac{dx}{dt} = (1-u)\,rx\left(1 - \dfrac{x+y}{k}\right) - \dfrac{\beta xvz}{x+y+\varepsilon} \\ \dfrac{dx}{dt} = \dfrac{\beta xvz}{x+y+\varepsilon} - (1-u)\,\delta y \\ \dfrac{dv}{dt} = b\,(1-u)\,\delta y - \dfrac{\beta xvz}{x+y+\varepsilon} - \alpha v \\ \dfrac{dz}{dt} = gu\left(p - z\right) - cz, \end{cases} \end{aligned} \tag{26}$$

(2^0)

$$X\left(t_0\right) = X_0, \tag{27}$$

(3^0)

$$\begin{aligned} \dot{\lambda} &= -\frac{\partial H}{\partial X} \\ &= -\begin{bmatrix} \lambda_1 \cdot pf_{11} + \lambda_2 \cdot pf_{21} + \lambda_3 \cdot pf_{31} + \lambda_4 \cdot pf_{41} + b_{11}x \\ \lambda_1 \cdot pf_{12} + \lambda_2 \cdot pf_{22} + \lambda_3 \cdot pf_{32} + \lambda_4 \cdot pf_{42} + b_{22}y \\ \lambda_1 \cdot pf_{13} + \lambda_2 \cdot pf_{23} + \lambda_3 \cdot pf_{33} + \lambda_4 \cdot pf_{43} + b_{33}v \\ \lambda_1 \cdot pf_{14} + \lambda_2 \cdot pf_{24} + \lambda_3 \cdot pf_{34} + \lambda_4 \cdot pf_{44} \end{bmatrix}, \end{aligned} \tag{28}$$

where

$$\begin{aligned} pf_{11} &= r\,(1-u)\left(1 - \frac{2x+y}{K}\right) - \frac{\beta\left(y+\varepsilon\right)vz}{\left(x+y+\varepsilon\right)^2}, \\ pf_{12} &= \frac{\beta xvz}{\left(x+y+\varepsilon\right)^2} - (1-u)\,\frac{r}{K}x, \\ pf_{13} &= -\frac{\beta xz}{x+y+\varepsilon}, \\ pf_{14} &= -\frac{\beta xv}{x+y+\varepsilon}, \\ pf_{21} &= \frac{\beta\left(y+\varepsilon\right)vz}{\left(x+y+\varepsilon\right)^2}, \\ pf_{22} &= -\,(1-u)\,\delta - \frac{\beta xvz}{\left(x+y+\varepsilon\right)^2}, \\ pf_{23} &= \frac{\beta xz}{x+y+\varepsilon}, \\ pf_{24} &= \frac{\beta xv}{x+y+\varepsilon}, \\ pf_{31} &= -\frac{\beta\left(y+\varepsilon\right)vz}{\left(x+y+\varepsilon\right)^2}, \\ pf_{32} &= b\,(1-u)\,\delta + \frac{\beta xvz}{\left(x+y+\varepsilon\right)^2}, \\ pf_{33} &= -\frac{\beta xz}{x+y+\varepsilon} - \alpha, \\ pf_{34} &= -\frac{\beta xv}{x+y+\varepsilon}, \\ pf_{41} &= 0, \\ pf_{42} &= 0, \\ pf_{43} &= 0, \\ pf_{44} &= -gu - c, \end{aligned} \tag{29}$$

(4^0)

$$\lambda\left(t_f\right) = AX\left(t_f\right) = \begin{bmatrix} a_{11} & 0 & 0 & 0 \\ 0 & a_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x\left(t_f\right) \\ y\left(t_f\right) \\ v\left(t_f\right) \\ z\left(t_f\right) \end{bmatrix}, \quad (30)$$

(5^0)

$$\frac{\partial H}{\partial u} = c_{11}u - \lambda_1 rx\left(1 - \frac{x+y}{K}\right) + \lambda_2\delta y - \lambda_3 b\delta y + \lambda_4 g\left(z - p\right). \quad (31)$$

3. Results and Discussion

In this part, based on the minimum principle, we will give the simulation of optimal strategy by using the Runge–Kutta fourth-order scheme and the steepest gradient method [27].

We choose 100 days as the control time. The efficacy u can theoretically lie between 0 and 1, where 0 corresponds to no effectiveness of the MEK and 1 corresponds to full effectiveness of the MEK. However, the perfect efficacy of MEK is unlikely to be achieved totally, so we suppose that the maximum effect is 0.98.

It is difficult to choose the parameters exactly based on biological meaning without experiment data, since the global stability of E_1 means the therapy has no effect when

$$b < 1 + \frac{\alpha\left(K+\varepsilon\right)}{\beta K} \cdot \frac{gu + c}{gup}. \quad (32)$$

We choose $\beta = 0.2\,\text{day}^{-1}$, $\varepsilon = 0.009$, $\delta = 0.5\,\text{day}^{-1}$, $r = 6$ cells day^{-1}, $\alpha = 0.5\,\text{day}^{-1}$, $p = 10$, $g = 0.1\,\text{day}^{-1}$, $c = 0.5\,\text{day}^{-1}$, $K = 9 \times 10^8$ cells, $b = 4\,\text{cell}^{-1}\,\text{day}^{-1}$, $a_{11} = a_{22} = 1$, $b_{11} = b_{22} = b_{33} = 10^{-5}$, and $c_{11} = 8000$.

Even if we use the maximum constant $u = 0.98$, $b > 1 + (\alpha(K+\varepsilon)/\beta K)((gu+c)/gup) = 2.5255$ also holds.

We give and compare two control strategies with the same initial conditions:

$$x0 = \left(6 \times 10^7,\ 0,\ 5 \times 10^4,\ 4 \times 10^2\right). \quad (33)$$

The optimal control simulation results are shown in Figure 3. The corresponding state dynamics for uninfected and infected tumor cells under the optimal control are shown in Figure 4 with solid line.

We choose constant control $u = 0.98$ to compare with the optimal control effect; the state dynamics for uninfected and infected tumor cells under the constant control are shown in Figure 4 with dotted line.

From the simulation, we can see that even if we use the maximum constant $u = 0.98$, the tumor cells still increase at the former stage and then keep stable at some level. But if we use the optimal control strategy as shown in Figure 3, that is to say, we need not use the maximum dosage of MEK all the time, though the tumor cells increase quickly with the lower dosage of MEK at the beginning stage, about 5 days later, it will begin to decrease and keep lower than that of constant control. As for the infected tumor, it is apparent that more tumor cells are infected with the optimal strategy; this in turn can help tumor cells keep at a contrarily lower level.

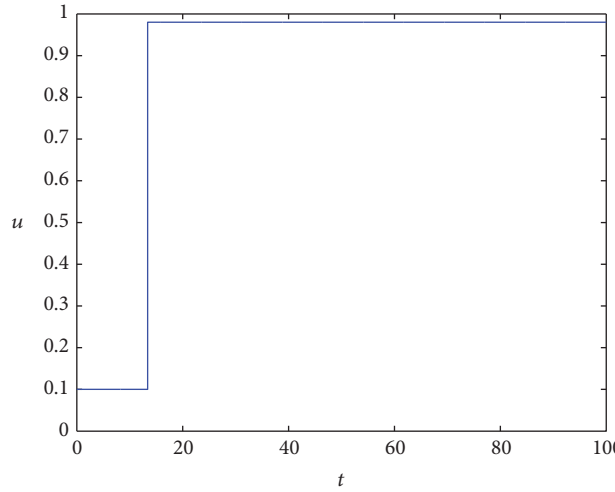


Figure 3: The optimal control of MEK.

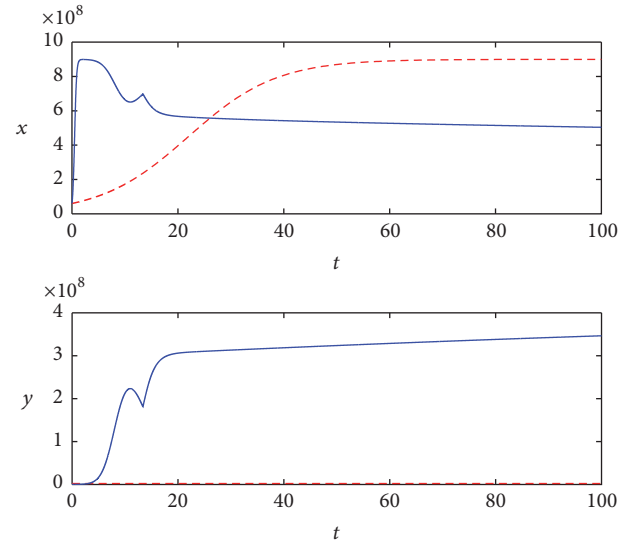


Figure 4: State dynamics for uninfected and infected tumor cells with different control strategies.

4. Conclusion

This paper introduces a new mathematical model of tumor therapy with oncolytic virus and MEK inhibitor. The stability of the equilibrium points is analyzed. Because inhibitors (MEK) can not only lead to greater oncolytic virus infection into cancer cells, but also cause cell cycle to stop, from theoretical analysis and numerical simulations, we compare optimal control strategy about the dosage of MEK inhibitor and constant control strategy with the same initial conditions. Simulations show that the optimal control has better control effect than constant control. But it should be pointed out that

our model has no cure equilibrium point, so, just from our model, we could not say that the tumor can be cured only by oncolytic virus therapy. But the optimal control strategy can help to prevent the tumor from getting worse and worse. Some other therapy methods should be combined to cure the tumor.

As we cannot get the exact parameters based on biological meaning, more work should be done about the modeling and simulations.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

This paper is supported by 2015 National Traditional Chinese Medicine Clinical Research Base Business Construction Special Topics (JDZX2015299).

References

[1] L. Zhou, W. W. He, Z. N. Zhu et al., "The clinical research progress for oncolytic adenovirus targeting cancer therapy," *China Biotechnology*, vol. 33, no. 12, pp. 105–113, 2013.

[2] N. L. Komarova and D. Wodarz, "ODE models for oncolytic virus dynamics," *Journal of Theoretical Biology*, vol. 263, no. 4, pp. 530–543, 2010.

[3] E. Kelly and S. J. Russell, "History of oncolytic viruses: genesis to genetic engineering," *Molecular Therapy*, vol. 15, no. 4, pp. 651–659, 2007.

[4] R. Rodriguez, E. R. Schuur, H. Y. Lim, G. A. Henderson, J. W. Simons, and D. R. Henderson, "Prostate attenuated replication competent adenovirus (ARCA) CN706: a selective cytotoxic for prostate-specific antigen-positive prostate cancer cells," *Cancer Research*, vol. 57, no. 13, pp. 2559–2563, 1997.

[5] I. Ganly, D. Kirn, G. Eckhardt et al., "A phase I study of ONYX-015, an E1B attenuated adenovirus, administered intratumorally to patients with recurrent head and neck cancer," *Clinical Cancer Research*, vol. 6, no. 3, pp. 798–806, 2000.

[6] R. M. Eager and J. Nemunaitis, "Clinical development directions in oncolytic viral therapy," *Cancer Gene Therapy*, vol. 18, no. 5, pp. 305–317, 2011.

[7] O. G. Donnelly, F. Errington-Mais, R. Prestwich et al., "Recent clinical experience with oncolytic viruses," *Current Pharmaceutical Biotechnology*, vol. 13, no. 9, pp. 1834–1841, 2012.

[8] S. J. Russell, K.-W. Peng, and J. C. Bell, "Oncolytic virotherapy," *Nature Biotechnology*, vol. 30, no. 7, pp. 658–670, 2012.

[9] T. S. Miest and R. Cattaneo, "New viruses for cancer therapy: meeting clinical needs," *Nature Reviews Microbiology*, vol. 12, no. 1, pp. 23–34, 2014.

[10] M. R. Patel and R. A. Kratzke, "Oncolytic virus therapy for cancer: the first wave of translational clinical trials," *Translational Research*, vol. 161, no. 4, pp. 355–364, 2013.

[11] M. Aghi and R. L. Martuza, "Oncolytic viral therapies—the clinical experience," *Oncogene*, vol. 24, no. 52, pp. 7802–7816, 2005.

[12] W. Si and W. Zhang, "Control exponential growth of tumor cells with slow spread of oncolytic virus," *Journal of Theoretical Biology*, vol. 367, pp. 111–129, 2015.

[13] D. Wodarz, "Viruses as antitumor weapons: defining conditions for tumor remission," *Cancer Research*, vol. 61, no. 8, pp. 3501–3507, 2001.

[14] B. S. Choudhury and B. Nasipuri, "Efficient virotherapy of cancer in the presence of immune response," *International Journal of Dynamics and Control*, vol. 2, no. 3, pp. 314–325, 2014.

[15] Z. Bajzer, T. Carr, K. Josić, S. J. Russell, and D. Dingli, "Modeling of cancer virotherapy with recombinant measles viruses," *Journal of Theoretical Biology*, vol. 252, no. 1, pp. 109–122, 2008.

[16] A. S. Novozhilov, F. S. Berezovskaya, E. V. Koonin, and G. P. Karev, "Mathematical modeling of tumor therapy with oncolytic viruses: regimes with complete tumor elimination within the framework of deterministic models," *Biology Direct*, vol. 1, article 6, pp. 1–18, 2006.

[17] Y. Wang, J. P. Tian, and J. Wei, "Lytic cycle: a defining process in oncolytic virotherapy," *Applied Mathematical Modelling*, vol. 37, no. 8, pp. 5962–5978, 2013.

[18] J. P. Tian, "The replicability of oncolytic virus: defining conditions in tumor virotherapy," *Mathematical Biosciences and Engineering. MBE*, vol. 8, no. 3, pp. 841–860, 2011.

[19] Y. Chen and Y. M. Su, "An improved model of tumor therapy with oncolytic virus," *Journal of Henan University of Science & Technology*, vol. 37, no. 4, pp. 92–96, 2016.

[20] K. A. Rauen, D. Sudilovsky, J. L. Le et al., "Expression of the coxsackie adenovirus receptor in normal prostate and in primary and metastatic prostate carcinoma: potential relevance to gene therapy," *Cancer Research*, vol. 62, no. 13, pp. 3812–3818, 2002.

[21] M. D. Lacher, M. I. Tiirikainen, E. F. Saunier et al., "Transforming growth factor-β receptor inhibition enhances adenoviral infectability of carcinoma cells via up-regulation of coxsackie and adenovirus receptor in conjunction with reversal of epithelial-mesenchymal transition," *Cancer Research*, vol. 66, no. 3, pp. 1648–1657, 2006.

[22] M. Anders, C. Christian, M. McMahon, F. McCormick, and W. M. Korn, "Inhibition of the Raf/MEK/ERK pathway up-regulates expression of the coxsackievirus and adenovirus receptor in cancer cells," *Cancer Research*, vol. 63, no. 9, pp. 2088–2095, 2003.

[23] G. Cherubini, T. Petouchoff, M. Grossi, S. Piersanti, E. Cundari, and I. Saggio, "E1B55K-deleted Adenovirus (ONYX-015) overrides G1/S and G 2/M checkpoints and causes mitotic catastrophe and endoreduplication in p53-proficient normal cells," *Cell Cycle*, vol. 5, no. 19, pp. 2244–2252, 2006.

[24] N. Bagheri, M. Shiina, D. A. Lauffenburger, and W. M. Korn, "A dynamical systems model for combinatorial cancer therapy enhances oncolytic adenovirus efficacy by MEK-inhibition," *PLoS Computational Biology*, vol. 7, no. 2, Article ID e1001085, 2011.

[25] R. Zurakowski and D. Wodarz, "Model-driven approaches for in vitro combination therapy using ONYX-015 replicating oncolytic adenovirus," *Journal of Theoretical Biology*, vol. 245, no. 1, pp. 1–8, 2007.

[26] J. P. LaSalle, *The Stability of Dynamical Systems*, Regional Conference Series in Applied Mathematics, SIAM, Philadelphia, Pa, USA, 1976.

[27] G. Pachpute and S. P. Chakrabarty, "Dynamics of hepatitis C under optimal therapy and sampling based analysis," *Communications in Nonlinear Science and Numerical Simulation*, vol. 18, no. 8, pp. 2202–2212, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 1850404, 10 pages
http://dx.doi.org/10.1155/2016/1850404

Research Article

Multichannel Convolutional Neural Network for Biological Relation Extraction

Chanqin Quan,[1] Lei Hua,[2] Xiao Sun,[2] and Wenjun Bai[1]

[1]*Graduate School of System Informatics, Kobe University, Kobe, Japan*
[2]*Department of Computer and Information Science, Hefei University of Technology, Hefei, China*

Correspondence should be addressed to Lei Hua; hualeilxf@163.com

Received 22 June 2016; Accepted 9 November 2016

Academic Editor: Oliver Ray

The plethora of biomedical relations which are embedded in medical logs (records) demands researchers' attention. Previous theoretical and practical focuses were restricted on traditional machine learning techniques. However, these methods are susceptible to the issues of *"vocabulary gap"* and data sparseness and the unattainable automation process in feature extraction. To address aforementioned issues, in this work, we propose a multichannel convolutional neural network (MCCNN) for automated biomedical relation extraction. The proposed model has the following two contributions: (1) it enables the fusion of multiple (e.g., five) versions in word embeddings; (2) the need for manual feature engineering can be obviated by automated feature learning with convolutional neural network (CNN). We evaluated our model on two biomedical relation extraction tasks: drug-drug interaction (DDI) extraction and protein-protein interaction (PPI) extraction. For DDI task, our system achieved an overall f-score of 70.2% compared to the standard linear SVM based system (e.g., 67.0%) on DDIExtraction 2013 challenge dataset. And for PPI task, we evaluated our system on Aimed and BioInfer PPI corpus; our system exceeded the state-of-art ensemble SVM system by 2.7% and 5.6% on f-scores.

1. Introduction

DDI and PPI are two of the most typical tasks in the field of biological relation extraction. DDI task aims to extract the interactions among two or more drugs when these drugs are combined and act with each other in human body; the hidden drug interactions may seriously affect the health of human body. Therefore, it is significant to further understand the interactions of drugs to reduce drug-safety accidents. Different from DDI task, PPI task aims to extract the interaction relations among proteins, and it has captured much interest among the study of biomedical relations recently [1, 2]. There are a number of databases which have been created for DDI (DrugBank [3, 4]) and PPI (MINT [5], IntAct [6]). However, with the rapid growth of biomedical literatures (e.g., MedLine has doubled in size within decade), it is hard for these databases to keep up with the latest DDI or PPI. Consequently, efficient DDI and PPI extraction systems become particularly important.

Previous studies have explored many different methods for DDI and PPI tasks. The dominant techniques generally fall under three broad categories: cooccurrence based method [7], rule-pattern based method [8, 9], and statistical machine learning (ML) based method [10–13]. Cooccurrence based method considers two entities interacting with each other if entities occur in the same sentence. A major weakness of this method is its tendency for having a high recall but a low precision.

The rule and pattern based methods employ predefined patterns and rules to match the labeled sequence. Although having achieved high accuracy among traditional rule and pattern based methods, their sophistication in pattern design and attenuated recall performance deviate them from practical usage. Besides the rule and pattern based methods, ML based techniques view DDI or PPI task as a standard supervised classification problem, that is, to decide whether there is an interaction (binary classification) or what kinds of relations (multilabel classification) between two entities.

Compared with cooccurrence and rule-pattern based methods, ML based methods show much better performance and generalization, and the state-of-the-art results for DDI [14] and PPI [2] are all achieved by ML based methods.

Traditional ML based methods usually collect words around target entities as key features, such as unigram, bigram, and trigram, and then these features are put into a *bag-of-words* model and encoded into *one-hot* (https://en.wikipedia.org/wiki/One-hot) type representations; after that, these representations are fed to a traditional classifier such as SVM. However, such representations are unable to capture semantic relations among words or phrases and fail in generalizing the long context dependency [15]. The former issue is rendered as *"vocabulary gap"* (e.g., the words *"depend"* and *"rely"* (these words are considered as the cue words or interaction verbs [8] which are important in biomedical relation extraction) are different in *one-hot* representations, albeit their similar linguistic functions). The latter one is introduced due to the n-order Markov restriction that attempts to alleviate the issue of *"curse of dimensionality."* Moreover, the inability to extract features automatically leads to the laborious manual efforts in designing features, which hinders the practical use of traditional ML based methods in extracting biomedical relation features.

To tackle these issues, in this work, we employ word embedding [16, 17] (also known as distribution representations) to represent the words. Different from *one-hot* representation, word embedding could map words to dense vectors of real numbers in a low-dimensional space, and thus the *"vocabulary gap"* problem can be well solved by the dot product of two word vectors. Compared to *one-hot* model, which merely allows the binary coding fashion in words (e.g., yes or no), our employment of the word embedding was able to output the similarity of two words via dot product. Such representation also yield neurological underpinning and is more in consistent with the way of human thinking.

Based on the previous researches on word embedding, this research builds a model on distributed word embedding and proposes a multichannel convolutional neural network (MCCNN) for biomedical relation extraction. The concept *"channel"* in MCCNN is inspired by three-channel RGB image processing [18], which means different word embedding represents different channel and different aspect of input words. The proposed MCCNN integrates different versions of word embeddings for better representing the input words. The only input for MCCNN is the sentences which contain drug-drug pairs (in DDI task) and protein-protein pairs (in PPI task). By looking up different versions of word embedding, input sentences will be initialized and transformed into multichannel representations. After that, the robust neural network method (CNN) will be applied to automatically extract features and feed them to a Softmax layer for the classification.

In sum, our proposed MCCNN model has yield threefold contributions:

(1) We propose a new model MCCNN to tackle DDI and PPI tasks and demonstrate that MCCNN model which relies on multichannel word embedding is effective in extracting biomedical relations features; the proposed model allows the automated feature extraction process. We tested our proposed model on DDIExtraction 2013 challenge dataset and achieved an overall f-score 70.2% that outperformed the current best system in DDIExtraction challenge by 5.1% and recent [14] state-of-the-art linear SVM based method by 3.2%.

(2) We also evaluated the proposed model on Aimed and BioInfer PPI extraction tasks. The attained F-scores 72.4% and 79.6% which outperform the state-of-the-art ensemble SVM system by 2.7% and 5.6%, respectively.

(3) We release our code (https://github.com/coddinglxf/DDI) taking into account the model's simplicity and good performance.

In remaining sections, Section 2 details proposed MCCNN methods, Section 3 demonstrates and discusses the experiments results, Section 4 briefly concludes this work, and Section 5 details the implementation of MCCNN.

2. Method

In this section, firstly, we briefly describe the concept and training algorithm for word embedding. And then, we introduce the multichannel word embedding and CNN model for relation extraction in detail; at last, we show how to train proposed MCCNN model.

2.1. Word Embedding. Word embedding which could capture both syntactical and semantic information from a large unlabeled corpus has shown its effectiveness in many NLP tasks. The basic assumption for word embedding is that words which occur in similar contexts tend to have similar meanings. Many models had been proposed to train the word embedding, such as NNLM [16], LBL [19], Glove [20], and CBOW. CBOW model (also known as a part of word2vec [17] (https://code.google.com/archive/p/word2vec/)) is employed to train our own word embedding in this work due to its simplicity and effectiveness. CBOW model takes the average embedding of the context words as the context representation, and it reduces the training time by replacing the last traditional Softmax layer with a hierarchical Softmax. In addition, CBOW could further reduce time consumption by negative samples. An outline architecture of CBOW is shown by Figure 1.

2.2. Multichannel Word Embedding Input Layer. Word embedding reflects the distributions of words in unlabeled corpus. In order to ensure the maximum coverage of the word embeddings, the articles from PubMed, PMC, MedLine, and Wikipedia are used for training word embedding. Five versions of word embedding are generated based on these corpora. The first four word embeddings are released by Pyysalo et al. [21], while the fifth word embedding is trained by CBOW on MedLine corpus (http://www.nlm.nih.gov/databases/journal.html) (see Figure 1 for more details).

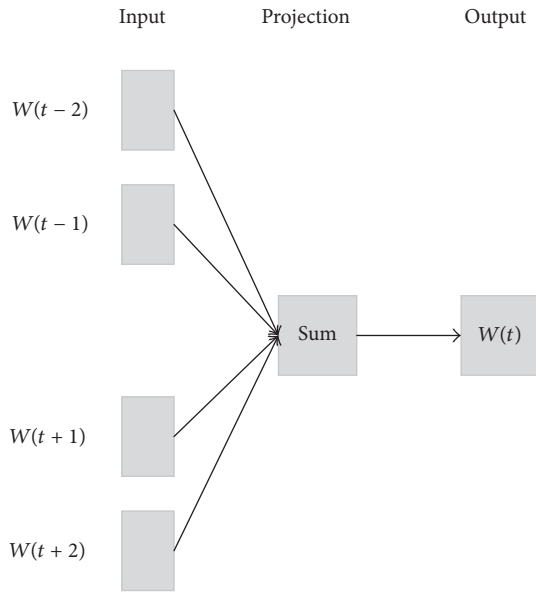


Figure 1: The architecture of CBOW model [17].

Table 1: Statistics for five word embeddings (all with 200 dimensions).

	Vocabulary size	Training corpus
1	2515686	PMC
2	2351706	PubMed
3	4087446	PMC and PubMed
4	5443656	Wikipedia and PubMed
5	650187	MedLine

The statistics of the five word embeddings are rendered in Table 1.

There are several advantages to use multichannels word embeddings. (1) PMC, MedLine, and PubMed corpus cover most of the literatures in the field of biology; thus these word embeddings can in large extent be used to extract biomedical relation features. (2) Some frequent words may occur in all of the five word embeddings, such kind of words has more information (weight) to leverage. (3) Word information can be shared among different word embeddings. Multichannel word embeddings could enlarge the coverage of vocabulary based on different ways of word embedding and decrease the number of unknown words.

The architecture of our proposed MCCNN is showed by Figure 2. c is defined as the number of the channels, v is the corpora's vocabulary size, N (N is the max length of the input sentence) is the length of input sentences, and d is the word embedding dimension. By looking up the pretrained multichannel word embeddings $\mathbf{D} \in R^{c\times v\times d}$, the multichannel inputs $\mathbf{V}$ can be represented as a 3-dimensional array with size $c \times N \times d$; the subsequent convolutional layer would take $\mathbf{V}$ as input and extract the features.

2.3. Convolutional Layer. The convolution operation could be considered to apply different filters $\mathbf{W} \in R^{c\times h\times d}$ to the

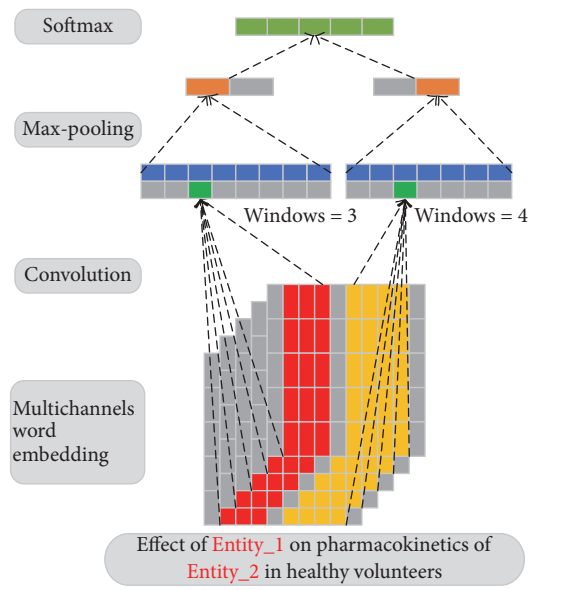


Figure 2: The architecture of the proposed MCCNN. In this example, the length of input sentence is 10, the input word embedding dimension is 5, and there are 5-word embedding channels. Therefore, the size of multichannel inputs is $5\times10\times5$. Two windows sizes 3 and 4 are used in this example. The green part is generate by (1). The orange part, representing the max-pooling result, is generated by take the maximum value of the blue part through (3). Since there are 2 filters for each window size, 2 features are produced. These extracted features are then concatenated together and fed to a Softmax layer for classification.

h-word windows in each channel of the input $\mathbf{V}$. Suppose $\mathbf{W}^i \in R^{h\times d}$ donates the filter for channel i and $\mathbf{V}^i \in R^{N\times d}$ is one of input word embeddings for channel i; a features $\mathbf{m}_k$ could be generated by (1), where $\mathbf{V}^i[k : k+h-1]$ (the red and yellow parts in Figure 2) is generated by parallel connecting row k to row $k+h-1$ in $\mathbf{V}^i$, f is an activation function, b is a bias term, and $\odot$ is element-wise multiplication

$$\mathbf{m}_k = f\left(\sum_{i=1}^{c}\mathbf{V}^i[k : k+h-1] \odot \mathbf{W}^i + b\right). \quad (1)$$

By applying an filter to each window in input sentence through (1), the model could produce a new feature $\mathbf{C}$ called feature map by

$$\mathbf{C} = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, \ldots, \mathbf{m}_{N-h+1}]. \quad (2)$$

Intuitively, convolutional layer is equal to applying filters on *n-grams* of input sentence. With different window size h, convolutional layer could extract various *n-grams* information.

2.4. Max-Pooling Layer. Max-pooling [26] operation by taking the maximum value over $\mathbf{C}$ (see (3)) brings two advantages: (1) it could extract the most important local features; (2) it reduces the computational complexity by reducing the feature dimension. A filter $\mathbf{W}$ would produce

a feature $\mathbf{C}^*$ (see (1), (2), and (3)), and thus M filters would generate M features. All of these features are represented by $\mathbf{r}^* = [\mathbf{C}_1^*, \mathbf{C}_2^*, \mathbf{C}_3^*, \ldots, \mathbf{C}_M^*]$

$$\mathbf{C}^* = \max(\mathbf{C}). \tag{3}$$

A single window size h can only capture fixed-size context information, by applying different window sizes, the model could learn more abundant features, suppose we use K to represent the number of window sizes, by concatenating the generated $\mathbf{r}^*$ for each window size, and the full feature $\mathbf{r} \in R^{KM\times 1}$ (the second last layer in Figure 2) is represented by

$$\mathbf{r} = [\mathbf{r}_1^*, \mathbf{r}_2^*, \mathbf{r}_3^*, \ldots, \mathbf{r}_K^*]. \tag{4}$$

2.5. Softmax Layer for Classification. Before feeding distributed representation $\mathbf{r}$ to the last Softmax layer for classifying the DDI or PPI type, original features space is transformed into confidence space $\mathbf{I} \in R^{O\times 1}$ by

$$\mathbf{I} = \mathbf{W}_2\mathbf{r}, \tag{5}$$

where $\mathbf{W}_2 \in R^{O\times KM}$ can be considered as a transformation matrix and O is the number of classes.

Each value in $\mathbf{I}$ represents the confidence of the current sample belongs to each class. A Softmax layer can normalize the confidences to $[0, 1]$ and thus can view the confidence from the perspective of probability. Given $\mathbf{I} = [i_1, i_2, \ldots, i_O]$, the output of Softmax layer $\mathbf{S} = [s_1, s_2, \ldots, s_O]$. The Softmax operation can be calculated by (6). Both s_j and $p(j \mid \mathbf{X})$ represent probability of an entity pair $\mathbf{x}$ which belongs to class j

$$s_j = p(j \mid \mathbf{x}) = \frac{e^{i_j}}{\sum_{k=1}^{O} e^{i_k}}. \tag{6}$$

2.6. Model Training. There are several parameters which need to be tuned during the training: the multichannel word embeddings $\mathbf{D}$, the multifilters $\mathbf{W}$, the transformation matrix $\mathbf{W}_2$, and the bias terms b. All the parameters are represented by $\theta = (\mathbf{D}, \mathbf{W}, \mathbf{W}_2, b)$. For training, we use Negative Log-Likelihood (NLL) in (7) as loss function ($\mathbf{y}_i$ is annotated label for the input sentence $\mathbf{x}_i$, and L is the minibatches size which means L samples will be fed to model in each training time). In order to minimize the loss function, we use gradient descent (GD) based method to learn the network parameters. In each training time, for L input samples $\langle \mathbf{x}_i, \mathbf{y}_i \rangle$, we firstly calculate the gradient (using the chain rules) of each parameter relative to loss and then update each parameter with learning rate λ by (8). It is notable that fixed learning rate λ would lead to unstable loss in training. In this work, we use an improved GD based algorithm Adadelta [27] to update the parameters in each training step; Adadelta can dynamically adjust the learning rate

$$\text{loss} = \sum_{i=1}^{L} -\log p(\mathbf{y}_i \mid \mathbf{x}_i), \tag{7}$$

$$\theta = \theta - \lambda \frac{\partial \text{loss}}{\theta}. \tag{8}$$

Table 2: An example for preprocessing of sentence *"Caution should be exercised when administering nabumetone with warfarin since interactions have been seen with other NSAIDs"* in DDI task. There are 3 entities in this example, and thus 3 entity pairs would be generated.

Entity1	Entity2	Generated inputs
Nabumetone	warfarin	Caution should be exercised when administering Entity1 with Entity2 since interactions have been seen with other EntityOther
Nabumetone	NSAIDs	Caution should be exercised when administering Entity1 with EntityOther since Interactions have been seen with other Entity2
Warfarin	NSAIDs	Caution should be exercised when administering EntityOther with Entity1 since interactions have been seen with other Entity2

3. Experiments

In this section, we firstly demonstrate the preprocessing method for both train and test corpora in DDI and PPI tasks. Secondly, the experimental results on DDI and PPI tasks are reported, respectively, for each task, we start from a baseline model with one-channel randomly initialized word embedding, and then, we show the results of one-channel word embedding; after that, we conduct the experiments on multichannel CNN model. In discussion part, we analyze the effects of hyperparameters settings as well as the typical errors caused by MCCNN.

3.1. Preprocessing for Corpora. The standard preprocessing includes sentence splitting and word tokenise. If there are n entities in a sentence, then, C_n^2 entity pairs would be generated. To reduce the sparseness and ensure the generalization of features, we share the similar preprocessing method as [11, 14] by replacing two target entities with special symbols *"Entity1"* and *"Entity2,"* respectively, and entities which are not target entities in inputs are all represented as *"EntityOther."* Table 2 demonstrates an example of preprocessing method.

The preprocessing method mentioned above may also produce some noise instances. For instance, entity pairs referred to the same name are unlikely to interact with each other. Such noise instances may (1) cause the imbalance distribution of the data, (2) hurt the performance of classifier, and (3) increase the training time. We define two rules to filter the noise instances. The rules are listed as follows. Table 3 shows the examples of noise instance for the rules.

Rule 1. Entity pairs referred to the same name or an entity which is an abbreviation of the other entity should be removed.

Rule 2. Entity pairs which are in a coordinate structure should be discarded.

Table 3: Examples of noise instance for defined rules; the mentioned entities are in italic.

Rule 1	*Anesthetics*, general: exaggeration of the hypotension induced by general *anesthetics*
Rule 2	To minimize CNS depression and possible potentiation, *barbiturates*, *antihistamines*, *narcotics*, *hypotensive* agents or phenothiazines should be used with caution

3.2. Evaluation on DDI Task

3.2.1. Datasets. DDIExtraction 2013 challenge (https://www.cs.york.ac.uk/semeval-2013/task9/) provides the benchmark corpora and annotations for DDI task [28]. The main purpose of this task is to pursue the classification of each drug-drug interaction according to one of the following four types: *advice*, *effect*, *mechanism*, and *int*; therefore, DDI is a 5-label (four interaction types plus one negative type) classification task. We shortly describe each interaction type and give an example for each type:

(1) *advice*: a recommendation or advice regarding the concomitant use of two drugs. For example, interaction may be expected, and *UROXATRAL* should not be used in combination with other *alpha-blockers*;

(2) *effect*: a description for the effect of drug-drug interaction. For example, *Methionine* may protect against the ototoxic effects of *gentamicin*;

(3) *mechanism*: pharmacodynamic or pharmacokinetic interactions between drug pairs. For example, *Grepafloxacin*, like other *quinolones*, may inhibit the metabolism of *caffeine* and *theobromine*;

(4) *int*: an interaction simply stated or described in a sentence. For example, the interaction of *omeprazole* and *ketoconazole* has been established.

(5) *negative*: no interaction between two entities. For example, concomitantly given thiazide *diuretics* did not interfere with the absorption of a tablet of *digoxin*.

The training and testing corpora in DDIExtraction 2013 consist of two parts: DrugBank and MedLine. A detailed description for these corpus could be found in Table 4. As can be seen from Table 4, our filtering rules are effective. In train datasets, the negative noise instances are reduced by 34.0% from 23665 to 15624 and only 22 out of 4020 (about 0.5%) positive instances are falsely filtered out. As for testing data, 35.0% of noise instances are discarded, while only 3 positive instances are mistaken. Such simple preprocessing method is beneficial to our system; especially it can reduce training time and avoid unbalanced classes.

3.2.2. Pretrained Word Embedding. As mentioned before, five versions of pretrained word embeddings are used in MCCNN as shown in Table 5. There are 13767 words (some of drug entities consisted with multiwords are all considered as single words) in DDI corpus. As a result, unknown words in smaller PMC and MedLine can be *"made up"* by word embedding with larger vocabulary coverage such as Wikipedia and PubMed.

3.2.3. Experimental Settings and Results. The experimental settings for DDI task are as follows: 200 filters are chosen for convolutional layer; minibatches size is set with 20; and window size h is set by 6, 7, 8, and 9, respectively. We select Relu as the activation function for convolutional layer due to its simplicity and good performance. Gaussian noise with mean 0.001 is added to the input multichannel word embedding, to overcome and prevent overfitting; we also add the weight constraint 5 to the last Softmax layer weight. Discussion section gives the details on parameter selection as well as the impact of the parameters.

Table 6 shows experimental results of baseline, one-channel, and the proposed MCCNN. As shown in Table 6, for each interaction type, we calculate the precision (P), recall (R), and the f-scores (F). We also report the overall micro-f-scores which has been used as a standard evaluation method in DDIExtraction 2013 challenge.

The baseline model utilizes randomly initialized word embedding, and the semantic similarity between words is not considered. Table 6 shows that one-channel with pretrained word embedding model performed much better than the baseline model and improved the overall f-scores from 60.12 to 66.90. This demonstrates that semantic information is crucial in DDI.

From Table 6, we can also find that, compared with one-channel model, MCCNN model achieved better results and improved the overall f-scores by 3.31%. For individual interaction type classification, MCCNN model also achieved the best f-scores. This demonstrates the effectiveness of the use of multichannel word embedding and richer semantic information.

We also trained the model on the corpus without preprocessing; the results could be found in Table 7. As we can see, preprocessing is important, which can improve the f-scores by 2.21% through reducing the potentially misleading examples.

Another aspect to note is that all three models behave worst on interaction type *"Int,"* such results are consistent with other systems [29–31], and the poor performance is mainly due to the lack of training samples (only 188 samples for training data and 96 samples for test data in Table 4).

In conclusion, (1) semantic information is important in DDI task, (2) rich semantic information can improve the performance, (3) preprocessing rules are crucial in DDI task, and (4) data scale would affect the model performance.

3.2.4. Performance Comparison. In this section, we compare the proposed MCCNN model with the top 3 approaches in DDIExtraction 2013 challenge (FBK-irst [29], WBI [29], and UTurku [31]). We also compare with the recently [14] novel linear kernel based SVM method. All of the four systems use SVM as the basic classifier. Both the FBK-irst and Kim's system detected the DDI at first (binary classification) and then classified the interaction into a specific

Table 4: Statistics for DDIExtraction 2013 challenge corpus. The entities pairs interacting with each other are labeled as positive, otherwise negative. The abstract indicates the number of article abstracts in datasets.

	Train			Test		
	DrugBank	MedLine	Overall	DrugBank	MedLine	Overall
Abstract	572	142	714	158	33	191
Positive	3788	232	4020	884	95	979
Negative	22118	1547	23665	4367	345	4712
Advice	818	8	826	214	7	221
Effect	1535	152	1687	298	62	360
Mechanism	1257	62	1319	278	24	302
Int	178	10	188	94	2	96
After preprocessing and filtering rules						
Positive	3767	231	3998	884	92	976
Negative	14445	1179	15624	2819	243	3062
Advice	815	7	822	214	7	221
Effect	1517	152	1669	298	62	360
Mechanism	1257	62	1319	278	21	299
Int	178	10	188	94	2	96

Table 5: Vocabulary included in five pretrained word embeddings.

	Vocabulary size	Word embedding
1	9984	PMC
2	10273	PubMed
3	10399	PMC and PubMed
4	10432	Wikipedia and PubMed
5	9639	Medline

type (multilabel classification). Different from FBK-irst's one-against-all strategy, Kim et al. utilized the one-against-one strategy for DDI type classification. They claimed the strategy could reduce the effect of unbalanced classes. WBI and UTurku ignored strategies problem by using multiclass SVM. The characteristics of the four approaches and the result comparisons are all listed in Tables 8 and 9.

As we can see, feature engineering still accounts for a large proportion of these systems. The features like word-levels features, dependency graphs, and parser trees are commonly used. In addition, syntax and dependency analysis are not effective for long sentences. The proposed MCCNN is able to avoid these problems by using word embedding and CNN. As shown by Table 9, MCCNN performs better than other methods for detecting interaction types *"Advice," "Effect,"* and *"Mechanism"* and further improves the state-of-the-art overall f-scores by 3.2%.

In addition, for interaction detection subtask (DEC), MCCNN achieved the second best f-scores compared to the FBK-irst's 80.0. DEC is a binary classification task, focusing on distinguishing the negative and positive instances. For most of the traditional methods, the most direct way is using cue words as they are not likely to be included in negative instances; in other words, "vocabulary gap" problem is not serious in these traditional methods. But in the problem of fine-grained interaction type classification, semantic information shows importance to classify different types. MCCNN showed its effectiveness on fine-grained classification by combing richer semantic information.

3.2.5. Compared with Other CNN Based Models. It is notable that CNN was also utilized by Zhao et al. [32] recently; they combined traditional CNN and external features such as contexts, shortest path, and part-of-speech to classify the interaction type and achieved an overall f-scores 68.6 which was similar to our results. The differences between [32] and our model lie on two aspects: (1) feature engineering still plays an important part in [32] model, whereas our model demands no manually feature sets; (2) multichannel word embeddings in our model contain richer semantic information which has been proved to be much useful in fine-grained interaction classification task.

3.2.6. Evaluation on Separated DrugBank and MedLine Corpus. Table 10 shows the performances of MCCNN on separated DrugBank and MedLine corpus. As shown in Table 10, MCCNN obtained f-scores 70.8 (compared to Kim's 69.8, FBK-irst's 67.6) on DrugBank and a sharp decline f-scores 28.0 (compared to Kim's 38.2, FBK-irst's 39.8). Reference [29] pointed out that such worse performance on MedLine might be caused by the presence of the cue words. From our point of view, the smaller number of training sentences in MedLine could also lead to the poor performances, as a proof, the MCCNN performed much better on MedLine (52.6) when trained on larger DrugBank and much worse (10.0) on DrugBank when trained on smaller MedLine in Table 10. As mentioned earlier, the scale of the data still has a great impact on the final results.

Table 6: Experimental results of baseline, one-channel, and the proposed MCCNN on DDI task. *Baseline*: with one-channel randomly initialized word embedding. *One-channel*: with one-channel Wikipedia and PubMed word embedding.

	Baseline			One-channel			MCCNN		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Advice	**89.39**	53.88	67.24	80.77	67.12	73.32	82.99	73.52	**77.97**
Effect	56.32	57.42	56.87	60.46	73.67	66.41	**67.03**	69.47	**68.23**
Mechanism	78.33	53.36	63.47	64.72	70.81	67.63	**85.00**	62.75	**72.20**
Int	**93.55**	30.21	45.67	82.05	33.33	47.41	75.51	38.54	**51.03**
Overall (micro)	70.00	52.68	60.12	66.50	67.31	66.90	**75.99**	65.25	**70.21**

Table 7: Performances of model with and without preprocessing.

	F-score
MCCNN (with preprocessing)	**70.21**
MCCNN (without preprocessing)	67.80

Table 8: Feature sets for four approaches.

Method	Feature sets
Kim	Word features, dependency graph features Word pair features, parse tree features Noun phrase constrained coordination features
FBK-irst	Linear features, path-enclosed tree kernels Shallow linguistic features
WBI	Features combination of other DDI methods
UTurku	Linear features, external resources Word features, graph features

Table 9: Comparisons with other systems on f-scores. ADV, EFF, MEC, and INT donate advice, effect, mechanism, and int, respectively, while DEC refers to interaction detection.

	ADV	EFF	MEC	INT	DEC	Overall
Kim	72.5	66.2	69.3	48.3	77.5	67.0
FBK-irst	69.2	62.8	67.9	**54.7**	**80.0**	65.1
WBI	63.2	61.0	61.8	51.0	75.9	60.9
UTurku	63.0	60.0	58.2	50.7	69.6	59.4
MCCNN	**78.0**	**68.2**	**72.2**	51.0	79.0	**70.2**

Table 10: Evaluation results (overall f-scores) on separated DrugBank and MedLine corpus. The first column corresponds to the training data set, while the first row corresponds to the test data set.

	DrugBank	MedLine
DrugBank	70.8	52.6
MedLine	10.0	28.0

3.3. Evaluation on PPI Task

3.3.1. Datasets and Pretrained Word Embedding. Two PPI datasets Aimed and BioInfer (http://mars.cs.utu.fi/PPICorpora/) are used to evaluate MCCNN. Aimed was manually tagged by Bunescu et al. [33] which included about 200 medical abstracts with around 1900 sentences and was considered as a standard dataset for PPI task. BioInfer [34] was developed by Turku BioNLP group (http://bionlp.utu.fi/clinicalcorpus.html) which contained about 1100 sentences. For corpora preprocessing, we do not use the filter rules in PPI task because of the limited size of corpus. The statistics of two datasets could be found in Table 11. We also report the vocabulary included in five pretrained word embeddings in Table 12.

Table 11: Statistics for Aimed and BioInfer datasets after preprocessing.

Datasets	Positive	Negative
BioInfer	2512	7010
Aimed	995	4812

Table 12: Vocabulary in pretrained word embedding.

	Aimed	BioInfer	Word embedding
All	6276	5461	—
1	5293	4666	PMC
2	5363	4712	PubMed
3	5404	4749	PMC and PubMed
4	5414	4762	Wikipedia and PubMed
5	4977	4328	MedLine

3.3.2. Changes of Performance from Baseline to MCCNN. For PPI experimental settings, the only difference from DDI task is the window size. Because the average sentence length in PPI task (42 in BioInfer, 36 in Aimed) is shorter than sentence length in DDI task (51), we set windows size h as 3, 4, 5, and 6.

Table 13 shows the experimental results of baseline, one-channel, and the proposed MCCNN on PPI task. We used 10-fold cross validation method for evaluation. As can be seen from Table 13, one-channel model performed much better than baseline model and improved the f-scores by 1.31% and 4.73% on Aimed and BioInfer, respectively. MCCNN achieved the best f-scores and improved the f-scores by 6.87% and 2.55% on Aimed and BioInfer when compared with one-channel.

3.3.3. Performance Comparison. Table 14 shows the comparisons with other systems on Aimed and BioInfer corpus. Kernel methods have been proved efficient in recent

Table 13: Change of performances from baseline to MCCNN on Aimed and BioInfer datasets, respectively.

	Baseline			One-channel			MCCNN		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Aimed	71.62	61.25	64.27	72.28	60.82	65.58	76.41	69.00	**72.45**
BioInfer	78.13	73.00	72.34	76.06	79.43	77.07	81.30	78.10	**79.62**

Table 14: Comparisons with other systems (f-scores) on Aimed and BioInfer.

	Aimed	BioInfer
Choi and Myaeng [22]	67.0	72.6
Yang et al. [23]	64.4	65.9
Li et al. [2]	69.7	74.0
Erkan et al. [11]	59.6	—
Miwa et al. [24]	60.8	68.1
Miwa et al. [25]	64.2	67.6
MCCNN (the proposed)	**72.4**	**79.6**

researches. Reference [22] proposed a single convolutional parse tree kernel and gave an in-depth analysis about the tree pruning and tree kernel decay factors. Reference [11] made full use of the shortest dependency path and proposed the edit-distance kernel. It has been verified that a combination of multiple kernels could improve effectiveness of kernel based PPI extraction methods. References [23–25] proposed hybrid kernel by integrating various kernels, such as bag-of-word kernel, subset tree kernel, graph kernel, and POS path kernel; they all achieved competitive results on PPI task.

It is notable that the word embedding information was also integrated by Li et al. [2]. They assigned a category to each word by clustering the word embedding, which can be used as a distributed representation feature. They also made full use of brown cluster and instance representation by words clustering method. The relationship between two words is no longer a simple yes or no; words with similar meanings are clustered and assigned with the same class label. The methods are essential to weaken *"vocabulary gap"* and proved to significantly improve the performance in their experiments (7.1% and 4.9% f-scores improvement on Aimed and BioInfer compared with their baseline model). Through combining the other features such as bag-of-words and syntactic features, they obtained remarkable results on Aimed and BioInfer.

Distributed representation features proposed by Li et al. [2] could be considered as a *"hard"* assignment: a cluster label for each word, but the extracted features are still discrete. As a benefit from word embedding and CNN, the proposed MCCNN model is able to be trained in a continuous space and manual assignment is not necessary. Compared with existing kernel based methods, the baseline model yielded a comparable performance. By replacing the randomly initialized word embedding with pretrained one, the one-channel model achieved better results and improved the state-of-the-art f-scores by 3% on BioInfer corpora. Furthermore, by integrating multichannel word embedding, the proposed MCCNN model exceeded 2.7% and 5.6% compared with [2] approach on Aimed and BioInfer.

3.4. Discussions. In this section, we firstly investigate the effects of hyperparameters, and then we carefully analyze the errors caused by MCCNN as well as the possible solutions to errors.

3.4.1. Hyperparameter Settings. The hyperparameters of neural network have great impact on the experimental results. In this work, three parameters including window size h, filter numbers M, and minibatches size need to be adjusted. To find the best hyperparameters, we split the training datasets into two parts: one for training and the other for validation. The basic method is to change one of the parameters while the other parameters remain unchanged. Filter numbers are set by $[10, 20, 50, 100, 200, 400]$, and the value range of minibatches size is $[10, 20, 50, 100]$; in addition, windows size h is set by $[3, 5, 7, 9, 11, 13]$. Experimental results show that the best settings for system are as follows: M is 200, minibatches size is 20, and h is 7 (7 in DDI task and 3 in PPI task). According to the suggestion that the best window size combination is usually close to each other by Zhang and Wallace [35], we set the windows size h as $[5, 6, 7, 8]$ in DDI task and $[3, 4, 5, 6]$ in PPI task.

Two methods are used to train a more robust model as well as prevent model from overfitting. The first method is to add Gaussian noise to the multichannel word embedding inputs. Considering the example in Table 2, the only differences of the three instances are the positions of Entity1, Entity2, and EntityOther; Gaussian noise could help to distinguish these instances. Experimental results showed that Gaussian noise can improve the performance by 0.5% in DDI task. In addition, according to [36], Gaussian noise could prevent overfitting. The other method is to add the weight constraint 5 to the last Softmax layer weight which could prevent overfitting.

3.4.2. Errors Analysis. Subjected to the complexity and diversity of the biomedical expressions, extracting relations from biological articles remain a big challenge. In this subsection, we carefully analyze the errors caused by MCCNN and list the two typical errors as follows:

(1) An input sentence is very long (more than 60 words), and Entity1 in this sentence is very close to Entity2.

(2) An input sentence is very long (more than 70 words), and Entity1 in this sentence is far from Entity2.

As the only input for MCCNN is a whole sentence, Entity1 and Entity2 are likely to be included in the same word window

Table 15: Configurations of machine.

GPU	NVIDIA GeForce GTX TITAN X
CPU	Intel(R) Xeon CPU E5-2620 v3 @ 2.4 GHz
System	Windows 7
memory	8 G

if Entity1 is very close to Entity2. In addition, due to the long context, the irrelevant word windows also have the chance to be chosen, and noise windows could hurt the system's performance. In the second case, a fixed window size such as 7 might fail to capture long sentence context when two entities are far from each other. A possible solution to avoid the above two errors might introduce dependency parser or parse tree information that would be able to capture the syntax information no matter the distance of the two entities.

4. Conclusion

In this work, we focused on three issues in biological relation extraction. The first is the *"vocabulary gap"* problem that would affect the performance of the biological extraction system; the second is how integration of semantic information will improve the performance of the system; and the third is the investigation of a mean to avoid the manual feature selection. The first two issues could be solved by introducing word embedding, especially the multichannel word embedding. By integrating CNN with aforementioned multichannel word embedding, the third problem could be well solved, and the experimental results show that our proposed MCCNN is at least effective for the two typical types of biomedical relation extraction tasks: drug-drug interaction (DDI) extraction and protein-protein interaction (PPI) extraction. In error analysis section, we notice that the proposed MCCNN is not capable of dealing with long sentences. In our future work, we would like to design and evaluate our relation extraction system by making full use of multichannel word embeddings, CNN, and syntax information.

5. Implementation

We use Keras (https://keras.io/) to implement our model. The configurations of our machine are listed in Table 15. It takes about 400 seconds to finish an epoch in training and 21 seconds to predict the results during the test. In order to get the best result, 10 iterations over train corpus are usually required.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This research has been partially supported by National Natural Science Foundation of China under Grant no. 61472117.

References

[1] C. Quan, M. Wang, and F. Ren, "An unsupervised text mining method for relation extraction from biomedical literature," *PLoS ONE*, vol. 9, no. 7, Article ID e102039, 2014.

[2] L. Li, R. Guo, Z. Jiang, and D. Huang, "An approach to improve kernel-based Protein-Protein Interaction extraction by learning from large-scale network data," *Methods*, vol. 83, pp. 44–50, 2015.

[3] C. Knox, V. Law, T. Jewison et al., "DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D1035–D1041, 2011.

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

[5] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, "Mint: a molecular interaction database," *FEBS Letters*, vol. 513, no. 1, pp. 135–140, 2002.

[6] S. Kerrien, B. Aranda, L. Breuza et al., "The IntAct molecular interaction database in 2012," *Nucleic Acids Research*, vol. 40, pp. D841–D846, 2012.

[7] R. Bunescu, R. Mooney, A. Ramani, and E. Marcotte, "Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from medline," in *Proceedings of the HLT-NAACL Workshop on Linking Natural Language Processing and Biology (BioNLP '06)*, New York, NY, USA, 2006.

[8] K. Fundel, R. Küffner, and R. Zimmer, "RelEx—relation extraction using dependency parse trees," *Bioinformatics*, vol. 23, no. 3, pp. 365–371, 2007.

[9] I. Segura-Bedmar, P. Martínez, and C. de Pablo-Sánchez, "A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents," *BMC Bioinformatics* , vol. 12, supplement 2, p. S1, 2011.

[10] B. Cui, H. Lin, and Z. Yang, "SVM-based protein-protein interaction extraction from medline abstracts," in *Proceedings of the 2nd International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA '07)*, pp. 182–185, IEEE, Zhengzhou, China, September 2007.

[11] G. Erkan, A. Özgür, and D. R. Radev, "Semi-supervised classification for extracting protein interaction sentences using dependency parsing," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07)*, vol. 7, pp. 228–237, June 2007.

[12] C. Sun, L. Lin, and X. Wang, "Using maximum entropy model to extract protein-protein interaction information from biomedical literature," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues: Third International Conference on Intelligent Computing, ICIC 2007 Qingdao, China, August 21–24, 2007 Proceedings*, vol. 4681 of *Lecture Notes in Computer Science*, pp. 730–737, Springer, Berlin, Germany, 2007.

[13] I. Segura-Bedmar, P. Martínez, and C. de Pablo-Sánchez, "Using a shallow linguistic kernel for drug-drug interaction extraction," *Journal of Biomedical Informatics*, vol. 44, no. 5, pp. 789–804, 2011.

[14] S. Kim, H. Liu, L. Yeganova, and W. J. Wilbur, "Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach," *Journal of Biomedical Informatics*, vol. 55, pp. 23–30, 2015.

[15] K. Arora and A. Rangarajan, "A compositional approach to language modeling," https://arxiv.org/abs/1604.00100.

[16] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J. L. Gauvain, "Neural probabilistic language models," in *Innovations in Machine Learning*, Studies in Fuzziness and Soft Computing, pp. 137–186, Springer, Berlin, Germany, 2006.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," https://arxiv.org/abs/1301.3781.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, NIPS Proceedings, pp. 1097–1105, Neural Information Processing Systems Foundation, 2012.

[19] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, pp. 641–648, ACM, Corvallis, Ore, USA, June 2007.

[20] J. Pennington, R. Socher, and C. D. Manning, "Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '14)*, vol. 14, pp. 1532–1543, Doha, Qatar, October 2014.

[21] S. Pyysalo, F. Ginter, F. Moen, and T. Salakoski, "Distributional semantics resources for biomedical text processing," in *Proceedings of the Languages in Biology and Medicine (LBM '13)*, pp. 39–44, Tokyo, Japan, December 2013.

[22] S.-P. Choi and S.-H. Myaeng, "Simplicity is better: revisiting single kernel ppi extraction," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling '10)*, pp. 206–214, Association for Computational Linguistics, Beijing, China, August 2010.

[23] Z. Yang, N. Tang, X. Zhang, H. Lin, Y. Li, and Z. Yang, "Multiple kernel learning in protein-protein interaction extraction from biomedical literature," *Artificial Intelligence in Medicine*, vol. 51, no. 3, pp. 163–173, 2011.

[24] M. Miwa, R. Sætre, Y. Miyao, and J. Tsujii, "Protein-protein interaction extraction by leveraging multiple kernels and parsers," *International Journal of Medical Informatics*, vol. 78, no. 12, pp. e39–e46, 2009.

[25] M. Miwa, R. Sætre, Y. Miyao, and J. Tsujii, "A rich feature vector for protein-protein interaction extraction from multiple corpora," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, pp. 121–130, Association for Computational Linguistics, August 2009.

[26] R. Collobert, R. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *The Journal of Machine Learning Research*, vol. 12, no. 2-1, pp. 2493–2537, 2011.

[27] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," https://arxiv.org/abs/1212.5701.

[28] I. Segura-bedmar, P. Martínez, and M. Herrero-zazo, "2013 SemEval-2013 task 9: extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)," in *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval '13)*, Association for Computational Linguistics, Atlanta, Ga, USA, June 2013.

[29] Md. F. M. Chowdhury and A. Lavelli, "FBK-irst: a multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information," in *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: 7nth International Workshop on Semantic Evaluation (SemEval '13)*, pp. 351–355, Atlanta, Ga, USA, June 2013.

[30] P. Thomas, M. Neves, T. Rocktäschel, and U. Leser, "WBI-DDI: drug-drug interaction extraction using majority voting," in *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval '13)*, vol. volume 2, pp. 628–635, Atlanta, Ga, USA, June 2013.

[31] J. Björne, S. Kaewphan, and T. Salakoski, "Uturku: drug named entity recognition and drug-drug interaction extraction using svm classification and domain knowledge," in *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (SemEval '13)*, vol. 2, pp. 651–659, 2013.

[32] Z. Zhao, Z. Yang, L. Luo, H. Lin, and J. Wang, "Drug drug interaction extraction from biomedical literature using syntax convolutional neural network," *Bioinformatics*, vol. 32, no. 22, pp. 3444–3453, 2016.

[33] R. Bunescu, R. Ge, R. J. Kate et al., "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial Intelligence in Medicine*, vol. 33, no. 2, pp. 139–155, 2005.

[34] S. Pyysalo, F. Ginter, J. Heimonen et al., "BioInfer: a corpus for information extraction in the biomedical domain," *BMC Bioinformatics*, vol. 8, no. 1, article 50, 2007.

[35] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," https://arxiv.org/abs/1510.03820.

[36] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103, ACM, July 2008.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 3962761, 11 pages
http://dx.doi.org/10.1155/2016/3962761

Research Article

Differentially Coexpressed Disease Gene Identification Based on Gene Coexpression Network

Xue Jiang, Han Zhang, and Xiongwen Quan

College of Computer and Control Engineering, Nankai University, Tianjin 300350, China

Correspondence should be addressed to Han Zhang; zhanghan@nankai.edu.cn

Received 23 September 2016; Accepted 26 October 2016

Academic Editor: Tun-Wen Pai

Screening disease-related genes by analyzing gene expression data has become a popular theme. Traditional disease-related gene selection methods always focus on identifying differentially expressed gene between case samples and a control group. These traditional methods may not fully consider the changes of interactions between genes at different cell states and the dynamic processes of gene expression levels during the disease progression. However, in order to understand the mechanism of disease, it is important to explore the dynamic changes of interactions between genes in biological networks at different cell states. In this study, we designed a novel framework to identify disease-related genes and developed a differentially coexpressed disease-related gene identification method based on gene coexpression network (DCGN) to screen differentially coexpressed genes. We firstly constructed phase-specific gene coexpression network using time-series gene expression data and defined the conception of differential coexpression of genes in coexpression network. Then, we designed two metrics to measure the value of gene differential coexpression according to the change of local topological structures between different phase-specific networks. Finally, we conducted meta-analysis of gene differential coexpression based on the rank-product method. Experimental results demonstrated the feasibility and effectiveness of DCGN and the superior performance of DCGN over other popular disease-related gene selection methods through real-world gene expression data sets.

1. Introduction

High throughput biotechnologies have been routinely used in biological and biomedical researches. As a result, tremendous amounts of large-scale omics data have been generated, providing not only great opportunities but also challenges for understanding the molecular mechanism of complex diseases. Screening disease-related genes by analyzing gene expression data represents one of these opportunities and challenges.

Differentially expressed gene analysis represents one of the most fundamental methods for disease-related gene identification by using gene expression data. Differentially expressed gene analysis methods select the genes which give the greatest contribution to diseases classification by comparing the changes of gene expression levels between normal samples and disease samples [1]. Those selected differentially expressed genes are considered as candidates to play a pathogenic role, termed disease-related genes or disease genes. The papers [2–4] firstly conducted gene expression analysis using statistical test, then ranked the genes in descending order according to the statistics which define the degree of gene differential expression, and finally selected the top genes as disease genes. The papers [5, 6] reconstructed gene expression data using nonnegative matrix factorization and conducted analysis of differentially expressed genes according to the new constructed matrix. The papers [7, 8] selected differentially expressed disease-related genes by minimizing the prediction error of classification. The papers [9, 10] obtained different disease-related gene subsets by using different samples and then got the optimal disease-related gene subset by integrating multiple disease-related gene subsets. This strategy in [9, 10] improved the correctness and robustness of disease-related genes. Though differential expression genes have high correlation with disease phenotypes and diseases classification, these methods may not fully consider the changes of interactions between genes in different cell states and the dynamic processes of gene expression levels

during disease development and progression for disease gene selection [11]. It is reported that complex diseases are often related to the changes of interactions between genes. Thus, some disease-related genes may not be identified by only finding differentially expressed genes.

Differentially coexpressed genes (DCG) analysis is different from the individual differentially expressed gene analysis methods. Differentially coexpressed genes are highly correlated under one cell state but uncorrelated under another cell state [12, 13]. Since the normal functions of genes are destroyed in disease cell state, the coexpression patterns in normal cell state are broken down [14]. Differential coexpression gene identification is very helpful for discovering potential biomarkers and understanding the pathophysiology of complex disease. The existing methods for identifying differentially coexpressed genes focused on gene-gene coexpression analysis or gene coexpression modules analysis. The earliest related research [15] proposed an additive model and a stochastic search algorithm to investigate differentially coexpressed genes. The paper [16] selected pairs of differentially coexpressed genes using a statistical method. The paper [17] constructed gene network by measuring the correlation between genes using mutual information and conducted clique analysis to get the differentially coexpressed genes.

As the normal interactions between genes would be greatly affected by abnormal protein in neurodegenerative diseases, such as Huntington disease, the symptoms of the disease grow progressively more severe and are debilitated with time, eventually leading to death. The disease gene (IT15) of Huntington disease which produces the abnormal disease protein (Htt) has already been discovered [18]. However, there is still no cure for this disease. In fact, the exact pathogenesis of Huntington disease has not yet been illustrated completely. The changes of interactions between genes caused by the abnormal protein are reflected as the changes of gene expression level. It is well known that the similar expression patterns represent the same biological process or function [19–21]. The changes of interactions between genes can be reflected by the changes of expression patterns in coexpression network, as gene coexpression network is constructed by using gene expression data. Thus, we can identify the differentially coexpressed disease-related genes by studying and analyzing the dynamic changes of gene coexpression patterns in phase-specific gene coexpression networks. This is of great significance to understand the pathogenesis of neurodegenerative diseases.

In this study, we developed a differentially coexpressed disease gene identification method based on gene coexpression network (DCGN) for identifying differential coexpression disease-related genes. We firstly constructed a series of phase-specific gene coexpression networks using gene expression data of different time points and defined the conception of differential coexpression of genes in coexpression network. Then, we designed two metrics to measure the value of gene differential coexpression according to the change of local topological structures between different phase-specific coexpression networks. Finally, we conducted meta-analysis of gene differential coexpression according to the rank-product method [22]. This paper provided a novel framework and a method to evaluate the value of differential coexpression for each gene rather than gene pairs or genes modules. Experimental results demonstrated the feasibility and effectiveness of DCGN and the superior performance of DCGN over other popular disease-related gene selection methods through real gene expression data sets.

The rest of this study was organized as follows: the DCGN was presented in Section 2. Experiments that demonstrated the performance of DCGN were reported in Section 3. The overall discussion with some suggestions for future research was presented in the last section.

2. Method

In this section, we firstly presented the overview of the novel framework for differentially coexpressed disease gene identification. The framework was shown in Figure 1.

Next, the gene coexpression network was introduced and the construction of gene coexpression network by using WGCNA software package [23, 24] was briefly described. Then, the conception of gene differential coexpression in coexpression network was defined and two metrics were proposed to measure the value of gene differential coexpression according to the change of local topological structures between phase-specific networks. Finally, the meta-analysis of gene differential coexpression based on the rank-product method was described.

2.1. Gene Coexpression Network. The gene coexpression network is usually constructed by measuring the gene expression similarity, which represents the coexpression relationships between genes [25]. Each node in the network represents a single gene. Each edge connecting two genes indicates the coexpression.

Let $X^t = [x_{ijt}] \in R^{n\times m}$ denote gene expression data in t-phase. x_{ijt} represents expression level of gene i in sample j at t-phase. n and m denote the number of genes and number of samples, respectively.

In order to study the dynamic changes of interactions between genes, we firstly constructed phase-specific gene coexpression network by using the WGCNA software package [23, 24], ensuring that the network is scale-free [26]. In the coexpression network $G = (V, E)$, V is the set of nodes, where one node corresponds to a gene. E is the set of edges, showing the mutual interactions between genes. w_{ij} is the weight of the edge connecting nodes i and j, $w_{ij} \in (0, 1)$. It should be noted that the stronger the Pearson correlation is, the larger the weight is. $W = [w_{ij}]$ is the weight matrix of gene coexpression network. The adjacency matrix is $A = [a_{ij}]$, where a_{ij} represents the interactions between nodes i and j. The calculation of a_{ij} is given by

$$a_{ij} = \begin{cases} 1, & \text{if } w_{ij} \neq 0; \\ 0, & \text{else.} \end{cases} \quad (1)$$

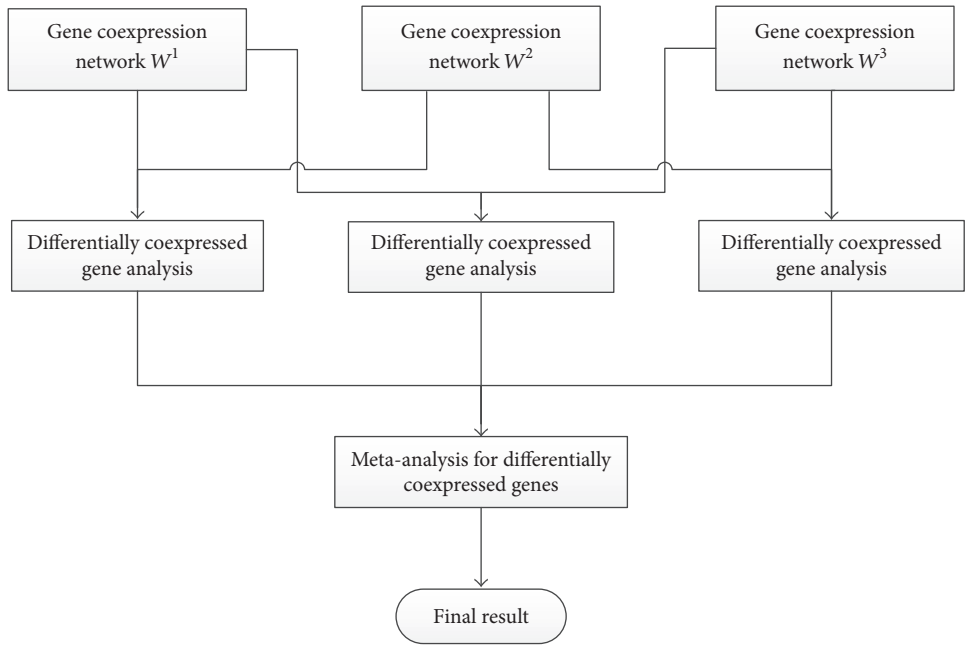


Figure 1: An overview of the novel framework for differentially coexpressed disease gene identification.

The transition matrix is $M = [m_{ij}]$, where m_{ij} denotes the probability of transition from node i to node j. The calculation of m_{ij} is given by

$$m_{ij} = \begin{cases} \dfrac{w_{ij}}{\sum_{j \in N_i} w_{ij}}, & \text{if } N_i \neq \varnothing; \\ 0, & \text{else.} \end{cases} \quad (2)$$

Here, N_i is the set of neighboring nodes of i in gene coexpression network W.

2.2. Gene Differential Coexpression Analysis. In this subsection, according to the change of local topological structures between different phase-specific gene coexpression networks, the gene differential coexpression analysis was conducted. The conception of gene differential coexpression was defined and two metrics were proposed to measure the value of gene differential coexpression.

Definition 1. Gene differential coexpression: in gene coexpression network W^t, $SW_i^{t,k}$ represents the i-centric system, a subnet including gene i and its k-level neighboring nodes (nodes that can be reached within k steps from node i). W^{t_1} and W^{t_2} denote the gene coexpression network in t_1-phase and t_2-phase, respectively. The differential coexpression of gene i represents the change of topological structures between $SW_i^{t_1,k}$ and $SW_i^{t_2,k}$.

In this paper, we designed two metrics to measure the value of gene differential coexpression. The first one is to evaluate the value of gene differential coexpression based on the local topological structures similarity. The second one is to evaluate the value of gene differential coexpression based on the variation of local topological information.

(1) The Value of Gene Differential Coexpression Based on the Local Topological Structures Similarity. In this subsection, we firstly defined the conception of the value of gene differential coexpression based on the local topological structures similarity between two different phase-specific gene coexpression networks.

Definition 2. The value of differential coexpression of gene i based on the local topological structures similarity between coexpression networks W^{t_1} and W^{t_2} is the topological similarity between $SW_i^{t_1,k}$ and $SW_i^{t_2,k}$. The value can be calculated according to the following equation:

$$d_i^{t_1 t_2, k} = 1 - \frac{\left| N_i^{t_1,k} \cap N_i^{t_2,k} \right|}{\left| N_i^{t_1,k} \cup N_i^{t_2,k} \right|}. \quad (3)$$

Here, $N_i^{t,k}$ is the set of connections between genes in $SW_i^{t,k}$.

(2) The Value of Gene Differential Coexpression Based on the Variation of Local Topological Information. In this subsection, the information of an edge was firstly described. Then, the conception of the value of gene differential coexpression based on the variation of local topological information between two different phase-specific gene coexpression networks was proposed.

In gene coexpression network W^t, we designed a function (shown as (6)) to evaluate the information of an edge. Then, according to (7), the value of gene differential coexpression

based on the variation of local topological information can be calculated.

Definition 3. The value of differential coexpression of gene i based on the variation of local topological information between coexpression networks W^{t_1} and W^{t_2} is the total variation of topological information caused by the topological structures differences between $SW_i^{t_1,k}$ and $SW_i^{t_2,k}$. The value can be calculated according to (7).

The details of computing the value of gene differential coexpression based on the variation of local topological information are shown below.

Step 1. In gene coexpression network W^t, extract submatrix $P^t = [p_{ij}^t]$, which denotes the subnetwork $SW_i^{t,k}$, from the transition matrix M^t, $j \in N_i^{tk}$. Here, N_i^{tk} is the set of genes i and their k-level neighboring nodes in gene coexpression network W^t.

Step 2. In subnetwork $SW_i^{t,k}$, calculate the maximum probability of transition from node i to node j with least steps. We use $p_{ij}^{\max}$ to denote the maximum transition probability, $j \neq i$, $j \in N_i^{tk}$.

Step 3. Normalize the probability of transition from node i to node j. The normalized probability of transition from nodes i to j is calculated by

$$p_{ij} = \frac{p_{ij}^{\max}}{\sum_{j \in N_i^{tk}, j \neq i} p_{ij}^{\max}}, \quad j \in N_i^{tk}. \tag{4}$$

After the above three steps, according to the topological information of $SW_i^{t,k}$, we transformed the i-centric subnetwork $SW_i^{t,k}$ into a network $DW_i^{t,k}$. In $DW_i^{t,k}$, node i connects to node j directly with the transition probability p_{ij}, $j \in N_i^{tk}$. It needs to be noted that, in the i-centric network $DW_i^{t,k}$, there are no connections between other nodes. To get the value of gene differential coexpression, we still need to do the following steps.

Step 4. To ensure that the strong coexpressed interactions between genes carry larger amount of information, we need to modify p_{ij} as

$$p_{ij}^{t,k} = \frac{1/p_{ij}}{\sum_{j \in N_i^{tk}, j \neq i} 1/p_{ij}}, \quad j \in N_i^{tk}. \tag{5}$$

Step 5. In i-centric subnetwork $SW_i^{t,k}$, the information that represents the connection between node i and node j is $I_{ij}^{t,k}$. The calculation of $I_{ij}^{t,k}$ is given by

$$I_{ij}^{t,k} = -\frac{\ln\left(p_{ij}^{t,k}\right)}{p_{ij}^{t,k}}, \quad j \in N_i^{tk}. \tag{6}$$

The value of differential coexpression of gene i based on the variation of topological information between coexpression networks W^{t_1} and W^{t_2} is calculated by

$$I_i^{t_1,t_2} = \sum_{j \in N_i^{t_1 k} - N_i^{t_2 k}} I_{ij}^{t_1,k} + \sum_{j \in N_i^{t_2 k} - N_i^{t_1 k}} I_{ij}^{t_2,k}. \tag{7}$$

The variation of topological information can be also interpreted as the total information change when one topological structure is replaced by another topological structure.

2.3. Meta-Analysis of Gene Differential Coexpression. After getting the value of gene differential coexpression according to any two different phase-specific coexpression networks, we ranked the genes in descending order according to the value of gene differential coexpression. $r_i^{t_1,t_2}$ denotes the ranking of gene i based on the coexpression network W^{t_1} and W^{t_2}. It needs to be noted that the larger the value of gene differential coexpression is, the higher the ranking of gene is. That means high ranking gene is of large probability of being disease-related gene. According to the rank-product method [22], the comprehensive ranking of gene i is

$$R_i = \left(\prod_{t_1, t_2 \in T, t_1 \neq t_2} r_i^{t_1 t_2} \right)^{(1/C)}. \tag{8}$$

Here, $C = N(N-1)/2$, where N is the number of coexpression networks. Then, rank the R_i, $i \in V$, in ascending order to get the final rank list of genes. It is important to note that the higher the ranking of gene is, the larger the probability of differentially coexpressed disease-related gene is.

3. Experimental Results

In this section, experiments were conducted to verify the feasibility of the novel framework for disease gene identification and the effectiveness of DCGN proposed in this paper. Two time-series real data sets were used in our study, one is of Huntington disease (HD) and the other one is of type 2 diabetes mellitus (T2DM). We firstly described the analysis process by using gene expression data of HD in detail. Then, the results in the gene expression data sets of T2DM were analyzed. Compared with other statistical disease gene selection methods, the superior performance of DCGN was illustrated. Finally, to explore the characters of DCGN based on different measures, a case study was conducted.

3.1. Gene Expression Data of HD. The gene expression data of HD used in our study was RNA-seq data from http://www.hdinhd.org/. It was obtained from striatum tissue of Huntington disease mice. Huntington disease is one kind of neurodegenerative diseases. It is due to a triplet repeat elongation in the Huntington gene (IT15), which leads to neuronal malfunction and degeneration through a large scale of different interactions between genes and a number of different molecular pathways. The symptoms of the disease grow progressively more sever and are debilitated with time, eventually leading to death.

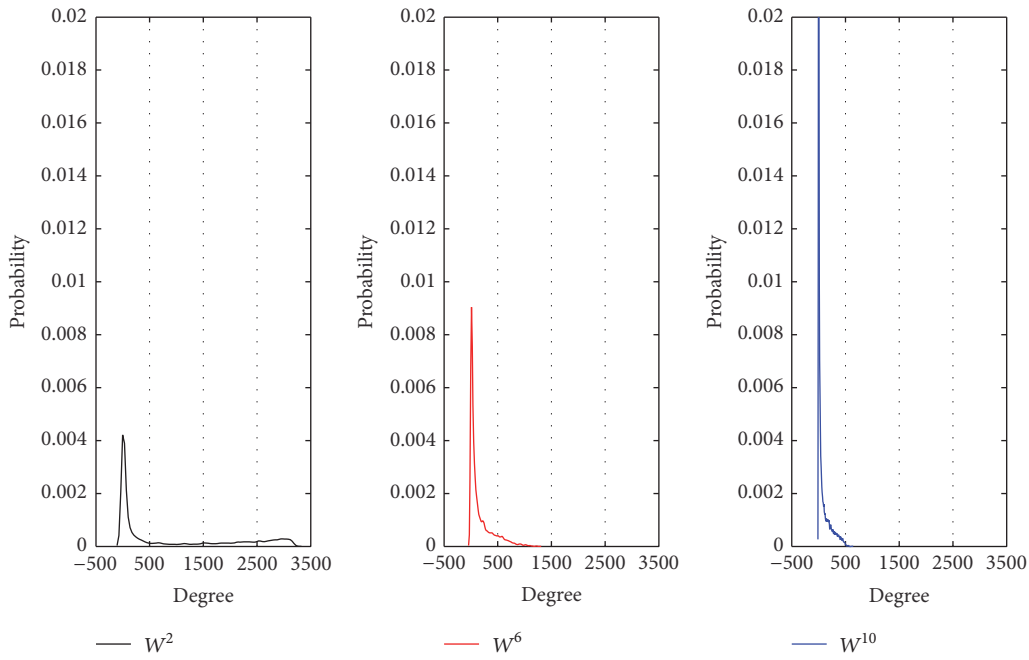


Figure 2: The distribution of degrees in W^2, W^6, and W^{10}.

Table 1: RNA-seq data of Huntington disease mice.

Tissue	Genotype	Age
Striatum	polyQ 92	2-month-old
	polyQ 111	6-month-old
	polyQ 140	10-month-old
	polyQ 175	

Table 2: Topological information of gene coexpression networks.

Network	W^2	W^6	W^{10}
Nodes number	8815	8815	8815
Edges number	7797404	1433744	628150
Average degree	1024.09	168.08	89.58
Average weight	0.423	0.339	0.336
Scatters number	1201	285	803

In the gene expression data, there are 4 genotypes, including polyQ 92, polyQ 111, polyQ 140, and polyQ 175. Each genotype has 8 replications. Thus, the gene expression data has 32 samples totally. According to the age of experimental mouse, there are 3 gene expression data sets in different phases, including gene expression data of 2-month-old HD mouse, gene expression data of 6-month-old HD mouse, and gene expression data of 10-month-old HD mouse. In order to clearly demonstrate the information of the experimental data, Table 1 was carried out. In order to filter out noise genes, we conducted a preprocessing step and selected 8815 genes from the total 23351 genes in the gene expression data. The data of modifier genes were from [27], which contained 520 genes in training set, including 89 disease genes and 431 nondisease genes.

3.2. The Topological Information of Gene Coexpression Network. The gene coexpression network was constructed by using the WGCNA software package [23, 24]. W^t denotes the gene coexpression network constructed with gene expression data of t-month-old HD mouse. The topological information of the three phase-specific networks is shown in Table 2. As shown in Table 2, there exist big differences between the topological structures of the three gene coexpression networks though we used the same standard to construct these networks.

To illustrate differences of the three networks, we analyzed the distribution of degrees, weighted degrees, and weights in each gene coexpression network.

Investigating the similarity between different coexpression networks, we can know that the similarity between W^2 and W^6 is only 0.032, the similarity between W^2 and W^{10} is 0.042, and the similarity between W^6 and W^{10} is 0.111.

From Figures 2, 3, and 4, we can get the following information and conclusions. Firstly, for W^2, there are denser connections (Figures 2 and 3) and the degrees of hub nodes in W^2 are about 3000 while most nodes have large degrees (Figure 2). At the same time, the connections between genes are also stronger (Figure 4). The above topological information of W^2 suggests that the interactions between genes are very active in 2-month-old Huntington disease mouse. Secondly, for W^6, compared with W^2, W^6 has quite sparse connections (Figures 2 and 3) and the degrees of hub nodes in W^6 are about 800 while only few nodes have large degrees (Figure 2). Moreover, the most connections between genes

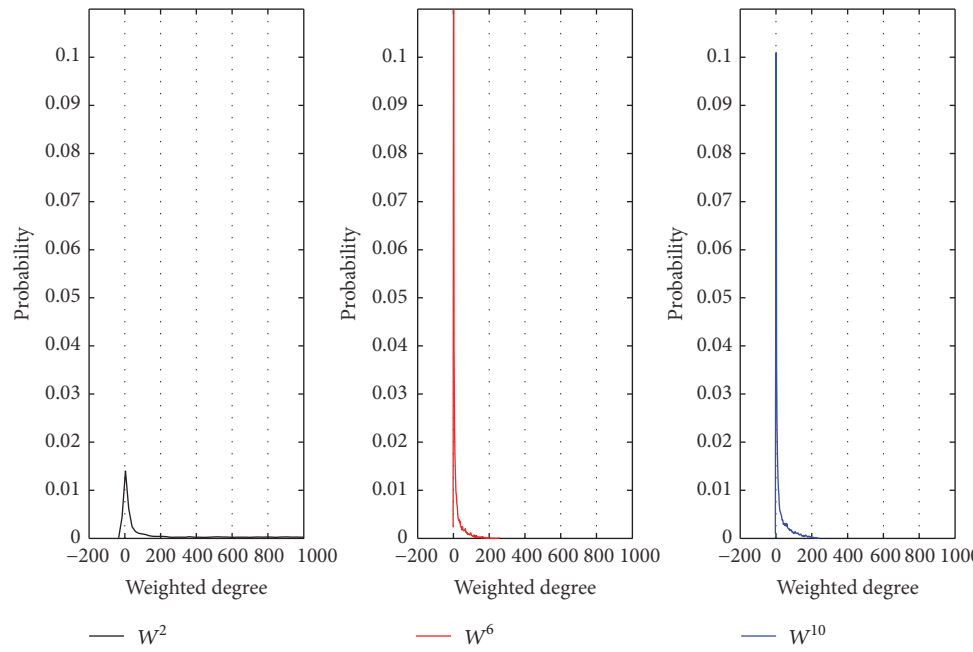


Figure 3: The distribution of weighted degrees in W^2, W^6, and W^{10}.

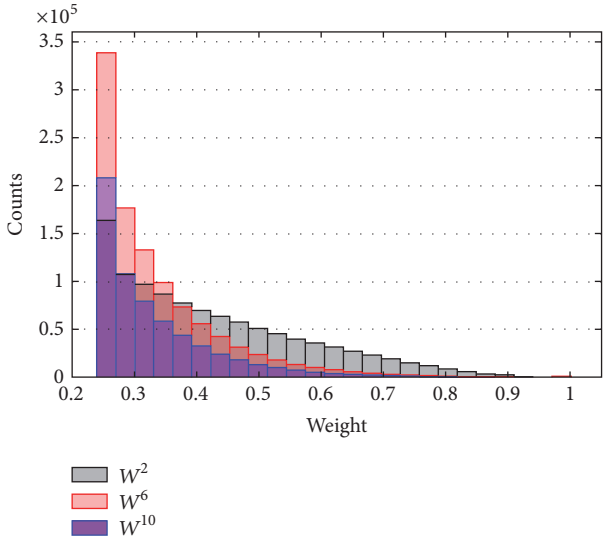


Figure 4: The distribution of weights in W^2, W^6, and W^{10}.

are not strong (Figure 4). This topological information of W^6 suggests that the interactions between genes in 6-month-old Huntington disease mouse are greatly changed. Thirdly, for W^{10}, the connections in W^{10} are sparser and weaker (Figures 2, 3, and 4), indicating that the interactions between genes in 10-month-old are not obvious.

The differences between the topological structures of W^2, W^6, and W^{10} stem from the fact that the expression of most genes has been affected by the Huntington disease as time goes on. The big differences between phase-specific gene coexpression networks indicate that the analysis of differentially coexpressed gene according to the changes of the topological structures of different networks may be helpful for understanding the changes of interactions between genes as the disease gets worse.

3.3. Performance Analysis of DCGN. According to Definition 2, we denoted the identification of differentially coexpressed genes based on the topological structure similarity by using (3) as DCGN-S. According to Definition 3, we denoted the identification of differentially coexpressed genes based on the variation of topological information by using (7) as DCGN-I. There is a parameter k, the level of the neighboring nodes, which needs to be preset in practice. In our paper, we set $k = 1$, $k = 2$, and $k = 3$ to test the performance of DCGN with different measures to evaluate the value of gene differential coexpression, including DCGN-S and DCGN-I. The following criteria were used to evaluate the identification accuracy of disease-related genes: the true positive rate (TPR), which is defined as the ratio of correctly predicted disease genes to all disease genes, and the false positive rate (FPR), which is defined as the ratio of incorrectly predicted disease genes to all nondisease genes. The receiver operating characteristic (ROC) curve was created by plotting TPR versus FPR. The area under the curve (AUC) [28] was also used as a measure of the identification accuracy.

As illustrated in Figure 5, with different k (the level of neighboring nodes), the ROC curves of DCGN-S with different k are approximate. From Figure 6, it is clear that the ROC curves of DCGN-I with different k are also approximate. These results suggest that the performances of DCGN-S and DCGN-I are insensitive to k. From Table 3, it can be seen that the AUCs of DCGN-S and DCGN-I with $k = 1$ are better than $k \geq 2$. This indicates that we may introduce redundancy information when $k \geq 2$. Thus the performances of DCGN-S and DCGN-I get poor when $k \geq 2$. In addition, this also

Table 3: The AUC of each experiment.

Method	$k = 1$	$k = 2$	$k = 3$
MFSN-S	**0.7118**	0.7062	0.7083
MFSN-I	**0.7101**	0.7081	0.7094
RP-FC	0.5856		
RP-t	0.5513		

Note. Bold indicates the best values.

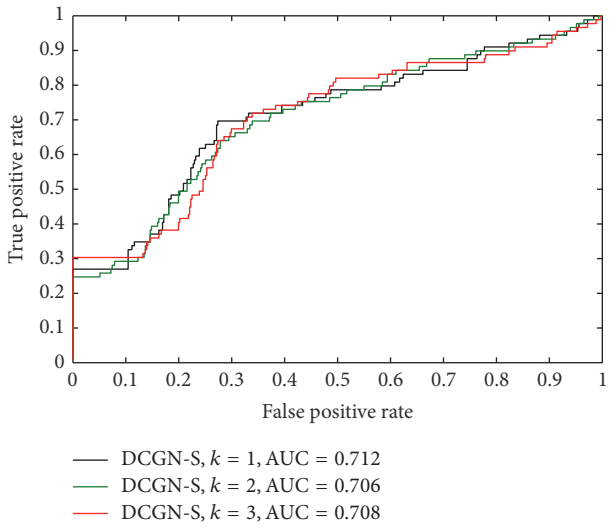


Figure 5: The ROC curves of DCGN-S with different k.

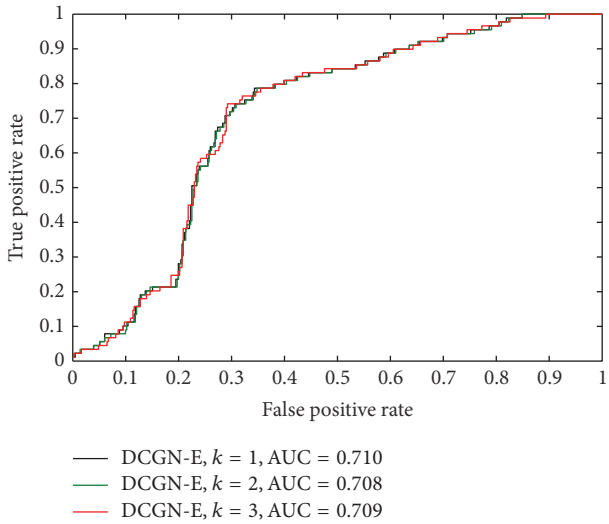


Figure 6: The ROC curves of DCGN-I with different k.

increases the computational complexity of DCGN when $k \geq 2$. Therefore, we suggest to use $k = 1$ in other experiments.

Comparing Figure 5 with Figure 6, it can be clearly seen that the ROC curves of DCGN-S are greatly different from DCGN-I. As illustrated in Figure 5, DCGN-S can distinguish disease genes from nondisease genes accurately for high ranking genes. We checked the rank lists and found that these high ranking genes share the same ranking in rank list, which means the values of differential coexpression of these genes are equal. It suggests that DCGN using the topological structure similarity can not precisely reflect the dynamic changes of the interactions between genes. As shown in Figure 6, though DCGN-I can hardly distinguish disease genes from nondisease genes for high ranking genes, the accuracy is greatly improved when FPR in [0.2, 0.4]. Though the nodes with large degrees are prone to get a higher rank by using DCGN-I (the analysis is shown in Section 3.6), DCGN-I fails to accurately distinguish disease genes from nondisease genes for high ranking genes. This suggests that there is no strong and significant correlation between hub nodes and disease genes. However, the ratio of disease genes to nondisease genes in training set is approximate 1 : 5. The ratio of TPR to FPR in hub nodes (high ranking genes) is approximate 1 : 1. It demonstrates that the hub nodes are more likely to be disease genes.

Table 4: The number of overlapped genes (the degree of overlap) between different rank lists in the same ranking range.

Ranking range Unit: 10^3	DCGN-S $k = 1, 2, 3$	DCGN-I $k = 1, 2, 3$	RP-FC ∩ RP-t —
[0, 1]	489 (0.49)	**926 (0.93)**	325 (0.33)
[1, 2]	153 (0.15)	**830 (0.83)**	186 (0.19)
[2, 3]	118 (0.12)	**809 (0.81)**	169 (0.17)
[3, 4]	86 (0.09)	**830 (0.83)**	224 (0.22)
[0, 2]	1269 (0.63)	**1896 (0.95)**	991 (0.50)
[0, 3]	2242 (0.75)	**2904 (0.97)**	1913 (0.64)
[0, 4]	3152 (0.79)	**3919 (0.98)**	2768 (0.69)

Table 5: The number of overlapped genes (the degree of overlap) by using different methods.

Ranking range Unit: 10^3	DCGN-S ∩ DCGN-I	DCGN-S ∩ RP-FC ∩ RP-t	DCGN-I ∩ RP-FC ∩ RP-t
[0, 1]	1 (0.001)	4 (0.004)	**13 (0.013)**
[0, 2]	5 (0.003)	48 (0.024)	**78 (0.039)**
[0, 3]	164 (0.055)	232 (0.077)	**278 (0.093)**
[0, 4]	509 (0.013)	637 (0.016)	**743 (0.019)**

3.4. The Performance Comparison of DCGN, RP-FC, and RP-t. To illustrate the effectiveness of our methods, we compared it with a rank-product method based on fold-change criteria [22], denoted as RP-FC, and a rank-product model based on t-test, denoted as RP-t [4]. The comparison of ROC curves of DCGN-S, DCGN-I, RP-FC, and RP-t is shown in Figure 7. We also investigated the differentially coexpressed genes obtained by using DCGN and the differentially expressed genes obtained by using RP-FC and RP-t, and the comparison results are shown in Tables 4 and 5. We used the following criteria to compare the results of different methods and to evaluate the performance of different methods: the number of overlapped genes between different rank lists in the same ranking range, which was used to test the robustness of the results, and the percentage of overlap, which is defined as the ratio of the number of overlapped genes in the same ranking range to the length of the ranking range.

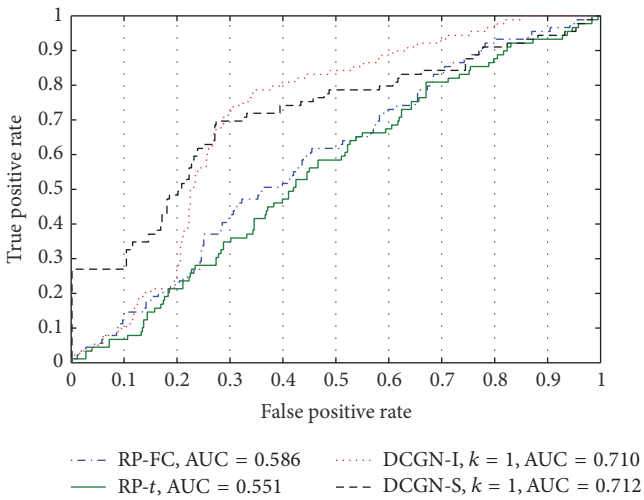


Figure 7: Performance comparison between RP-FC, RP-t, DCGN-S, and DCGN-I.

Table 6: Topological information of spatial-specific gene coexpression networks for T2DM.

Network	W^4	W^8	W^{12}	W^{16}	W^{20}
Nodes number	5555	5555	5555	5555	5555
Edges number	1712916	1656238	1312428	1167228	1104506
Ave degree	308.4	298.2	236.3	210.1	198.8
Ave weight	0.660	0.655	0.650	0.646	0.644
Scatters number	0	0	0	0	0

Table 7: The similarity of topological structures between any two phase-specific gene coexpression networks for T2DM.

Network	W^8	W^{12}	W^{16}	W^{20}
W^4	0.029	0.037	0.037	0.034
W^8		0.035	0.029	0.025
W^{12}			0.031	0.036
W^{16}				0.030

As illustrated in Figure 7, the AUCs of DCGN, including DCGN-S and DCGN-I, are far better than RP-FC and RP-t. From Table 3, it can be known that the AUC of DCGN-I with $k = 1$ is improved by more than 21.2% compared with the AUC of RP-FC. This indicates that the accuracy of the differentially coexpressed disease genes obtained by using DCGN is much better than that of differentially expressed genes obtained by using RP-FC and RP-t. Thus, the results of experiments verify the effectiveness of DCGN.

As illustrated in Table 4, when identifying differential coexpression genes by using the measure of the variation of topological information, the degree of overlap between different rank lists in the same ranking range is more than 80%, which is much higher compared to the results by other methods. It suggests that the result of DCGN-I is robust. When identifying differential coexpression genes by using the measure of topological structure similarity, the degree of overlap between different rank lists in the same ranking range is poor. It indicates that the rank of gene differential coexpression is greatly affected by parameter k. However, the fluctuation of the ranking of a gene is mostly controlled within 500. When identifying differentially expressed genes by using RP-FC and RP-t, respectively, the degree of overlap between the two rank lists in the same ranking range is very poor. It indicates the poor robustness of differentially expressed genes obtained by different methods.

We conducted further analysis of the overlapped genes in the same ranking range. As shown in Table 5, the degree of overlap between the differentially coexpressed genes obtained by using DCGN-S and the differentially coexpressed genes obtained by using DCGN-I is very poor. It illustrates that there exist big differences between the differentially coexpressed genes by using different measures to evaluate the value of gene differential coexpression. The degree of overlap between the differentially coexpressed genes obtained by using DCGN-S or DCGN-I and the differentially expressed genes obtained by using RP-FC and RP-t is also very poor. It suggests that there exist big differences between the differentially expressed genes which derived from the changes of gene expression levels and the differentially coexpressed genes which derived from the changes of gene coexpression networks.

3.5. Results in Type 2 Diabetes Mellitus Gene Expression Data. To test the feasibility and effectiveness of DCGN, we conducted experiment by using another time-series gene expression dataset of type 2 diabetes mellitus (T2DM) [29, 30]. The gene expression data was obtained from the Gene Expression Omnibus database (GSE 13271) of National Center for Biotechnology Information (NCBI). It was obtained from the white adipose tissue of disease rats aged from 4 weeks to 20 weeks, and the time interval was 4 weeks. There are 5 samples in each time point. In order to filter out noise genes, we conducted a preprocessing step and selected 5555 genes from the total 31099 genes in the gene expression data.

The T2DM related genes were downloaded from http://rgd.mcw.edu/wg/home. Totally, 202 disease-related genes were used, which were part of gene expression data in our experiment.

It is important to be noted that as there are only 5 samples in every time point, we improved the threshold to filter out large amount of false positive connections in the construction process of phase-specific gene coexpression networks. The topological information of the networks is shown in Table 6. The similarity of topological structures between any two networks is shown in Table 7.

Two rank lists of gene differential coexpressions were obtained by conducting gene differential coexpression analysis based on DCGN-I and DCGN-S with $k = 1$, respectively. Two rank lists of gene differential expression were obtained by conducting gene differential expression analysis based on RP-FC and RP-t, respectively. It needs to be noted that the high ranking of a gene in the rank list represents high probability of being a disease-related gene. We analyzed the distribution of rankings of disease-related genes in the training set, and the results are shown in Figure 8. From Figure 8, it can be

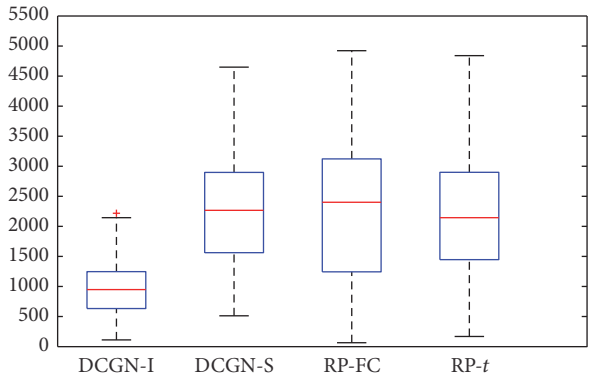


Figure 8: Boxplot of the rankings of disease-related genes.

seen that the average ranking of disease genes in training set by using DCGN-I is much higher than that by using other methods. From Table 5, the average degree of each gene coexpression network is in the range of [190, 310], indicating that there are dense connections in those networks. For networks with such feature, DCGN-I is more suitable than DCGN-S for analyzing those networks (the analysis is shown in Section 3.6). Experimental results also show that the effectiveness of DCGN-I is much better compared to DCGN-S.

In total, by analyzing local topological structures of gene coexpression networks, DCGN-I can screen differentially coexpressed genes. Compared with traditional differentially expressed gene identification methods, DCGN-I can effectively improve the accuracy of disease-related genes selection.

3.6. A Case Study. In order to analyze the characters of gene differential coexpression by using different measures, we conducted a case analysis in this subsection. Figure 9 illustrated the topological structure changes of a node, which has a larger degree in comparison with the node in Figure 10, from network W1 to network W2. Figure 10 showed the topological structure changes of a node with small degree from network W1 to network W2.

By using DCGN-S, we obtained that the value of differential coexpression of node 1 in Figure 9 is $S_1^{12} = 0.719$ and the value of differential coexpression of node $1'$ in Figure 9 is $S_{1'}^{12} = 0.813$. $S_1^{12} < S_{1'}^{12}$ means that node $1'$ in Figure 10 is of larger probability of being differentially coexpressed disease gene compared to node 1 in Figure 9.

By using DCGN-I, we obtained that the value of differential coexpression of node 1 in Figure 9 is $I_1^{12} = 998.6$ and the value of differential coexpression of node $1'$ in Figure 10 is $I_1^{12} = 310.6$. $I_1^{12} > I_{1'}^{12}$ means that node 1 in Figure 9 is of larger probability of being differentially coexpressed disease gene compared to node $1'$ in Figure 10.

In gene coexpression network W^t, if the degree of node i is large, the probability of transition from node i to its most neighboring nodes will be getting small. This is because $\sum_{j \in N_i^{tk}} p_{ij} = 1$. Since the information of an edge (see (6)), which is used to evaluate the information of a connection between nodes, is a monotone decreasing function, thus the changes of connections between large degree nodes could generate a greater value. Therefore, the identification of differentially coexpressed genes based on the variation of topological information is prone to give nodes with large degree (e.g., hub nodes) larger differential coexpression values. So, we can conclude that the nodes with significant network property of hub nodes are more likely to be screened as differentially coexpressed disease-related genes by using DCGN-I. The above characters of DCGN-I may contribute to improving the identification accuracy of disease genes [31]. From the above, the identification of differentially coexpressed genes based on the variation of topological information is more suitable for disease gene analysis of highly connected network.

From the case study it can also be seen that, for nodes with small degree, slight differences in two networks may generate large differential coexpression value when screening differentially coexpressed genes by using the measure of topological structure similarity, while, for nodes with large degree, great differences in two networks only generate small differential coexpression value. The above characters of DCGN-S may result in low accuracy of the identification of disease genes. It can be concluded that the identification of differentially coexpressed genes based on the topological structure similarity is more suitable for gene differential coexpression analysis of sparsely connected network.

In brief, the DCGN can effectively improve the accuracy of disease gene selection, while there exist large differences between the selected differentially coexpressed genes by using different measures to evaluate the value of gene differential coexpression. From the above analysis, it is also clear that DCGN-S and DCGN-I can be used to analyze networks with different topological structures.

4. Conclusion

Existing disease gene prediction methods mostly focus on cancer diagnosis and classification. For complex diseases with complex etiology, such as neurodegenerative diseases and diabetes mellitus, it is hard to find disease-related genes by traditional computing methods, making it difficult to discover and understand the development mechanism of these diseases.

In this paper, we designed a novel framework to identify disease-related genes and developed a differential coexpression analysis method by using time-series gene expression data. Compared with traditional analysis methods for differential expression disease-related genes, the effectiveness of DCGN for differential coexpression disease-related genes is verified.

It is reported that there usually exist a lot of false connections in gene coexpression network; thus the simulation results of coexpression network may have a great departure from real situation [32]. Therefore, constructing gene networks which accurately reflect the interactions between genes will greatly improve the performance of DCGN. In addition, the robustness of differentially coexpressed genes may be improved by integrating other information, such as weights of edges or the properties information of nodes, owing to the low percentage of overlap between the differentially

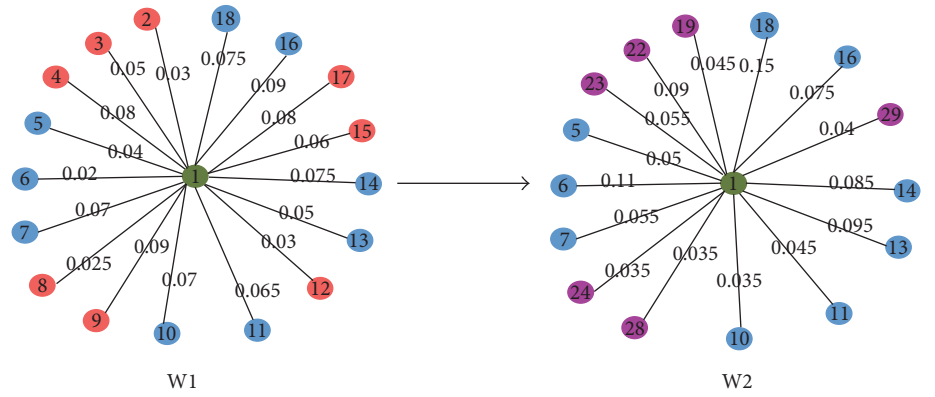


Figure 9: The topological structure changes of a node with large degree from W1 to W2.

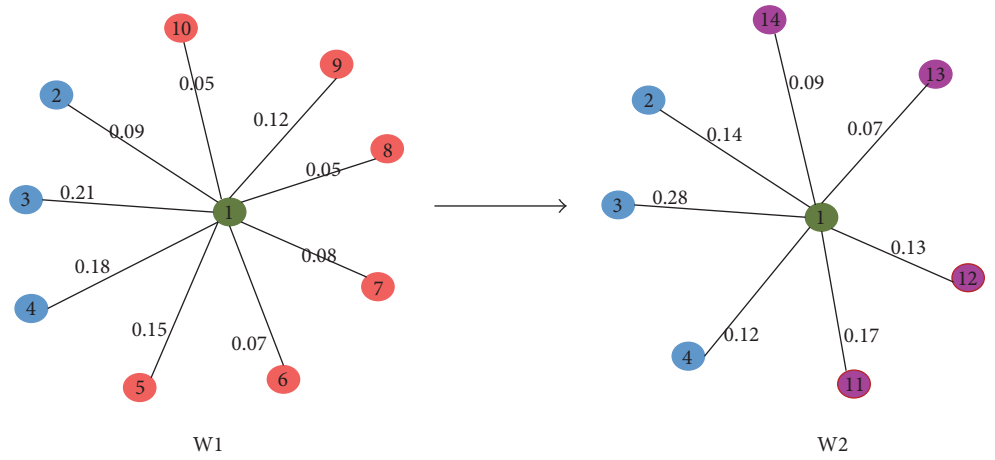


Figure 10: The topological structure changes of a node with small degree from W1 to W2.

coexpressed genes obtained by using DCGN-S and the differentially coexpressed genes obtained by using DCGN-I. As the percentage of overlap between the differentially coexpressed genes obtained by using DCGN and the differentially expressed genes obtained by using RP-FC and RP-*t* is poor, the identification accuracy of disease genes may be greatly improved by integrating the differential expression information of nodes into the process of differential coexpression analysis. We will conduct relevant studies about the strategies mentioned above.

Disclosure

The manuscript entitled "Differentially Coexpressed Disease Gene Identification Based on Gene Coexpression Network" is an extended and modified version of a previously presented paper in "The 10th International Conference on Systems Biology (ISB 2016)," which was entitled "Meta-Analysis for Feature Selection Based on Gene Co-Expression Network."

Competing Interests

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

Acknowledgments

The authors would like to thank Associate Professor Feng Duan of Naikai University for his comments that helped improve the manuscript, Associate Professor Zhiping Liu of Shandong University for his help and explanation of T2DM gene expression data, and Haibin Sun for the assistance and discussions. This work is supported by the Natural Science Foundation of Tianjin (15JCYBJC18900), the National Science Foundation of China (61403213), and the Key Program of Science Foundation of Tianjin (14JCZDJC31800).

References

[1] J. C. Ang, A. Mirzal, H. Haron, and H. N. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 5, pp. 971–989, 2016.

[2] C. Soneson and M. Delorenzi, "A comparison of methods for differential expression analysis of RNA-seq data," *BMC Bioinformatics*, vol. 14, no. 5, pp. 775–775, 2013.

[3] P. Baldi and A. D. Long, "A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, no. 6, pp. 509–519, 2001.

[4] F. Hong and R. Breitling, "A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments," *Bioinformatics*, vol. 24, no. 3, pp. 374–382, 2008.

[5] B. Liao, Y. Jiang, W. Liang et al., "On efficient feature ranking methods for high-throughput data analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 6, pp. 1374–1384, 2015.

[6] H.-Q. Wang, C.-H. Zheng, and X.-M. Zhao, "jNMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data," *Bioinformatics*, vol. 31, no. 4, pp. 572–580, 2015.

[7] P. A. Mundra and J. C. Rajapakse, "SVM-RFE with MRMR filter for gene selection," *IEEE Transactions on Nanobioscience*, vol. 9, no. 1, pp. 31–37, 2010.

[8] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[9] Q. Shen, R. Diao, and P. Su, "Feature selection ensemble," *Pure Collection*, vol. 10, pp. 289–306, 2012.

[10] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2009.

[11] D. Kostka and R. Spang, "Finding disease specific alterations in the co-expression of genes," *Bioinformatics*, vol. 20, no. 1, pp. i194–i199, 2004.

[12] D. L. Fuente and A. From, "differential expression to differential networking identification of dysfunctional regulatory networks in diseases," *Cell*, vol. 26, Article ID 326333, 2010.

[13] V. Varadan and D. Anastassiou, "Inference of disease-related molecular logic from systems-based microarray analysis," *PLoS Computational Biology*, vol. 2, no. 6, article e68, pp. 2–16, 2006.

[14] M. Watson, "CoXpress: differential co-expression in gene expression data," *BMC Bioinformatics*, vol. 7, article 509, 2006.

[15] D. Kostka and R. Spang, "Finding disease specific alterations in the co-expression of genes," *Bioinformatics*, vol. 20, no. 1, pp. i194–i199, 2004.

[16] Y. Lai, B. Wu, L. Chen, and H. Zhao, "A statistical method for identifying differential gene-gene co-expression patterns," *Bioinformatics*, vol. 20, no. 17, pp. 3146–3155, 2004.

[17] H. Zhang, X. Song, H. Wang, and X. Zhang, "MIClique: an algorithm to identify differentially coexpressed disease gene subset from microarray data," *Journal of Biomedicine and Biotechnology*, vol. 2009, Article ID 642524, 9 pages, 2009.

[18] The Huntington Disease Collaborative Research Group, "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes," *Cell*, vol. 72, no. 6, pp. 971–983, 1993.

[19] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "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.

[20] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[21] E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg, "A combined algorithm for genome-wide prediction of protein function," *Nature*, vol. 402, no. 6757, pp. 83–86, 1999.

[22] F. Hong, R. Breitling, C. W. McEntee, B. S. Wittner, J. L. Nemhauser, and J. Chory, "RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis," *Bioinformatics*, vol. 22, no. 22, pp. 2825–2827, 2006.

[23] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, pp. 1–45, 2005.

[24] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, article 559, 2008.

[25] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[26] A.-L. Barabsi, "Scale-free networks: a decade and beyond," *Science*, vol. 325, no. 5939, pp. 412–413, 2009.

[27] P. Langfelder, J. P. Cantle, D. Chatzopoulou et al., "Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice," *Nature Neuroscience*, vol. 19, no. 4, pp. 623–633, 2016.

[28] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[29] S.-Y. Sun, Z.-P. Liu, T. Zeng, Y. Wang, and L. N. Chen, "Spatio-temporal analysis of type 2 diabetes mellitus based on differential expression networks," *Scientific Reports*, vol. 3, no. 2, article 2268, pp. 468–473, 2013.

[30] B. Xue, S. Sukumaran, J. Nie, W. J. Jusko, D. C. DuBois, and R. R. Almon, "Adipose tissue deficiency and chronic inflammation in diabetic Goto-Kakizaki rats," *PLoS ONE*, vol. 6, no. 2, Article ID e17386, 2011.

[31] S. Deng, L. Zhu, and D. Huang, "Predicting hub genes associated with cervical cancer through gene co-expression networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 1, pp. 27–35, 2016.

[32] C. Liu, J. Yu, S. Yu et al., "MicroRNA-21 acts as an oncomir through multiple targets in human hepatocellular carcinoma," *Journal of Hepatology*, vol. 53, no. 1, pp. 98–107, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 8905675, 9 pages
http://dx.doi.org/10.1155/2016/8905675

Research Article

Identification of Five Novel *Salmonella* Typhi-Specific Genes as Markers for Diagnosis of Typhoid Fever Using Single-Gene Target PCR Assays

Yuan Xin Goay,[1] Kai Ling Chin,[1] Clarissa Ling Ling Tan,[1] Chiann Ying Yeoh,[1] Ja'afar Nuhu Ja'afar,[1] Abdul Rahman Zaidah,[2] Suresh Venkata Chinni,[3] and Kia Kien Phua[1]

[1]*Institute for Research in Molecular Medicine (INFORMM), Universiti Sains Malaysia (USM), Health Campus, 16150 Kubang Kerian, Kelantan, Malaysia*
[2]*Department of Medical Microbiology and Parasitology, Universiti Sains Malaysia (USM), Health Campus, 16150 Kubang Kerian, Kelantan, Malaysia*
[3]*Faculty of Applied Sciences, AIMST University, Jalan Bedong-Semeling, 08100 Bedong, Kedah, Malaysia*

Correspondence should be addressed to Kia Kien Phua; kkphua7@gmail.com

Received 22 June 2016; Revised 27 September 2016; Accepted 18 October 2016

Academic Editor: Hao-Teng Chang

Salmonella Typhi (*S.* Typhi) causes typhoid fever which is a disease characterised by high mortality and morbidity worldwide. In order to curtail the transmission of this highly infectious disease, identification of new markers that can detect the pathogen is needed for development of sensitive and specific diagnostic tests. In this study, genomic comparison of *S.* Typhi with other enteric pathogens was performed, and 6 *S.* Typhi genes, that is, STY0201, STY0307, STY0322, STY0326, STY2020, and STY2021, were found to be specific *in silico*. Six PCR assays each targeting a unique gene were developed to test the specificity of these genes *in vitro*. The diagnostic sensitivities and specificities of each assay were determined using 39 *S.* Typhi, 62 non-Typhi *Salmonella*, and 10 non-*Salmonella* clinical isolates. The results showed that 5 of these genes, that is, STY0307, STY0322, STY0326, STY2020, and STY2021, demonstrated 100% sensitivity (39/39) and 100% specificity (0/72). The detection limit of the 5 PCR assays was 32 pg for STY0322, 6.4 pg for STY0326, STY2020, and STY2021, and 1.28 pg for STY0307. In conclusion, 5 PCR assays using STY0307, STY0322, STY0326, STY2020, and STY2021 were developed and found to be highly specific at single-gene target resolution for diagnosis of typhoid fever.

1. Introduction

To date, there are more than 2,500 serotypes identified within the *Salmonella enterica* species [1]. Most are harmless to humans but one serotype, *Salmonella enterica* subspecies *enterica* serovar Typhi (*S.* Typhi), causes typhoid fever, a severe and life-threatening systemic infection in humans. Worldwide, typhoid fever causes 269,000 deaths from 26.9 million new cases each year [2]. Travellers, children, the elderly, and immune-compromised individuals are especially at risk [3, 4]. The clinical manifestations of typhoid fever are similar to other febrile illnesses. Therefore, diagnosis based on clinical signs and symptoms alone is difficult [5]. The emergence of multidrug-resistant *S.* Typhi strains and development of the typhoid carrier state have further complicated the management of typhoid fever [6, 7]. Delay in diagnosis and initiation of antibiotic treatment can cause serious clinical complications and fatality [8]. Thus, early and correct laboratory diagnosis of typhoid fever is critical to reduce the morbidity and mortality, as well as curtail transmission of the disease.

DNA-based detection methods, such as polymerase chain reaction (PCR), have proven to be sensitive, specific, and rapid compared to conventional culture-based methods for the diagnosis of many infectious diseases [9–11]. Several target genes have been used for *S.* Typhi identification using

PCR, such as the O antigen somatic genes (*tyv* and *prt*) [12], H antigen flagellar gene (*fliC*-d) [13], and Vi capsular antigen gene (*viaB*) [14]. However, these genes cannot stand alone as single *S*. Typhi-specific diagnostic marker since they are not specific to *S*. Typhi and are also found in other *Salmonella* serotypes. Thus, these markers provide provisional rather than differential diagnosis of typhoid fever. For example, the *fliC*-d gene of *S*. Typhi shares the same nucleic acid sequence as *S*. Muenchen [15]; the *prt* gene is present in *S*. Typhi, *S*. Paratyphi A, and *S*. Enteritidis [12]; and the *viaB* gene is found not only in *S*. Typhi but also in *S*. Dublin, a few strains of *S*. Paratyphi C [16] and *Citrobacter freundii* [17]. Due to the lack of specificity of these target genes, a combination of different pairs of primers using multiplex PCR [18] or nested PCR [19] are needed to increase the sensitivity and specificity of the PCR diagnostic test. This, however, will increase the cost, time, and complexity of the laboratory diagnosis.

Diagnostic markers which can detect pathogens at single-gene target resolution could lead to a simpler, cost-effective, and more functional DNA-based detection method since less primers are needed for target detection. Many approaches, such as subtractive hybridization [20], next generation sequencing [21], and microarray [22] techniques, have been used to identify genes that are specific or unique to a pathogen. However, these high-end technologies are cumbersome and expensive and sometimes yield false negative or false positive results [23]. Since bacterial genome databases have expanded tremendously over the past decade and advancement in computing technologies has made nucleic acid sequence alignment services readily accessible at NCBI, *in silico* comparative hybridization approach coupled with *in vitro* PCR (wet-lab) validation is sufficient to facilitate the translation of genomic data into diagnostic marker discoveries. In this study, a low-cost and simple attempt was made to identify new DNA diagnostic markers specific for *S*. Typhi by utilizing genome data (stored in NCBI databases) and nucleic acid sequence alignment tools (BLASTn) that are readily available in the public domain. The diagnostic sensitivities and specificities of the primers designed for amplifying whole gene sequences can be validated using a panel of confirmed bacteria isolates selected from *S*. Typhi, non-Typhi *Salmonella*, and non-*Salmonella* clinical isolates. To serve as a control for the PCR reaction, 16S rRNA gene, that is ubiquitous among bacteria species, can be used as a PCR amplification control [24].

2. Materials and Methods

2.1. Bacterial Strains. A total of 111 bacteria isolates including 39 *S*. Typhi, 62 non-Typhi *Salmonella* serotypes, and 10 non-*Salmonella* strains were used in this study. *S*. Typhi strains consisted of 1 *S*. Typhi reference strains (ATCC 7251) and 38 different pulsed-field types (PFTs) representing all strains in the state of Kelantan in Malaysia. These 38 PFTs were the result of screening 279 *S*. Typhi clinical isolates using pulsed-field gel electrophoresis (PFGE) [25]. Non-Typhi *Salmonella* serotypes were closely related *Salmonella* species made up of 26 different serotypes (Table 2) and 10 ATCC strains including *S*. Paratyphi A (ATCC 9150), *S*. Paratyphi B (ATCC BAA 1250), *S*. Paratyphi C (ATCC 9068), *S*. Enteritidis (ATCC 13076), *S*. Typhimurium (ATCC 14028), *S*. Weltevreden (NCTC 6534), *S*. Agona (ATCC 51957), *S*. Heidelberg (ATCC 8326), *S*. Poona (ATCC 04840), and *S*. Braenderup (ATCC BAA-664). In addition, 10 other non-*Salmonella* strains such as *Shigella dysenteriae*, *Shigella flexneri*, *Shigella boydii*, *Shigella sonnei*, *Vibrio cholera, Enterohemorrhagic E. coli*, *Enteropathogenic E. coli*, *Aeromonas hydrophila*, *Yersinia enterocolitica,* and *Klebsiella pneumonia* were also included. All clinical strains were procured from the Department of Clinical Microbiology and Parasitology, Hospital Universiti Sains Malaysia (HUSM), Kelantan, Malaysia, and the Biobank of the Institute for Research in Molecular Medicine (INFORMM), Kelantan, Malaysia. All bacteria strains were stored in glycerol stocks at −80°C until being ready for use. Ethical clearance for this project was obtained from the Human Research Ethics Committee, Universiti Sains Malaysia (reference number USMKK/PPP/JEPeM [235.3.(16)]).

2.2. Culture Conditions and Confirmation Tests. All bacteria isolates used in this study were confirmed by traditional culture, biochemical, and serotyping methods as described in ISO6579 with some modifications. Bacteria isolates were revived from frozen glycerol stocks by pipetting 100 μL thawed cells into 10 mL nutrient broth and incubated at 37°C for 18 hours in an orbital shaker at 200 rpm. The bacteria were streaked on Xylose Lysine Deoxycholate (XLD) selective agar and incubated at 37°C for 18 hours. Colonies grown on the agar were tested with a panel of biochemical tests, including Triple Sugar Iron (TSI), urease, Methyl Red Voges Proskauer (MRVP), citrate, and indole tests. Suspected *Salmonella* isolates were then sent to the *Salmonella* Reference Centre, Institute for Medical Research (IMR), Malaysia, to confirm their serotypes using specific antisera and latex agglutination method.

2.3. Identification of S. Typhi-Specific Genes Using Bioinformatics (In Silico). Full genome sequence of *S*. Typhi CT18 (GenBank accession number AL513382) was downloaded from the National Center for Biotechnology Information database (NCBI) and used as the reference genome. The 2 plasmids, namely, pHCM1 and pHCM2, which resided in *S*. Typhi CT18 were excluded since plasmids are genetically unstable. The 6 complete *S*. Typhi whole-genome sequences available in NCBI were used for data mining. They comprised CT18 (Genbank accession number AL513382) [27], Ty2 (Genbank accession number AE014613) [28], P-stx-12 (Genbank accession number CP003278) [29], Ty21a (Genbank accession number CP002099) [30], B/SF/13/03/195 (Genbank accession number CP012151) [31], and PM016/13 (Genbank accession number CP012091) [32]. In order to ascertain whether the genomic regions were conserved and specific to *S*. Typhi, the nucleotide Basic Local Alignment Search Tool (BLASTn), a free online software for nucleic acid analysis, was used to compare the whole-genome sequence of *S*. Typhi CT18 with the other 5 complete *S*. Typhi genomes and other bacteria genomes in the NCBI database (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Genes found in

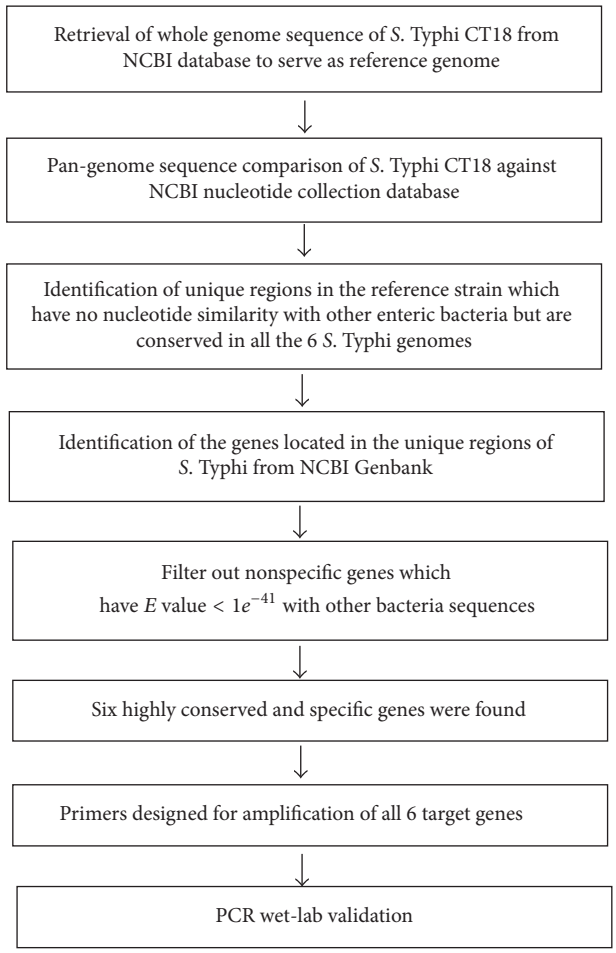


Figure 1: Experimental workflow describing the comparative genomic and wet-lab approaches used to identify and validate *S.* Typhi-specific DNA diagnostic markers.

unique regions which have no nucleotide similarity with other enteric organisms were identified and retrieved from the Genebank of NCBI. These genes were further screened individually using similarity searches against the NCBI non-redundant nucleotide (nr/nt) database to reconfirm their specificities. The program was set for "somewhat similar sequences search," which allowed nucleotide sequence matching down to 7 bases (the smaller the nucleotide size, the more sensitive the result). Realizing the high genome similarity among the enteric pathogens and the possibility that different geographical areas may result in different bacterial genotypes, only genes which have 100% sequence conservation (an *E*-value threshold = 0.0) in all 6 complete *S.* Typhi genomes and had little or no similarity (*E*-value threshold $\geq 1e^{-41}$) to other bacterial sequences in the NCBI database were considered as potential targets and were subjected to wet-lab analysis. The experimental pipeline is as shown in Figure 1.

2.4. Design of Oligonucleotide Primers for PCR Amplification. Primers were designed manually to amplify the *S.* Typhi-specific genes identified previously, including the start and the stop codons. A pair of primers specific for 16S rRNA gene amplification as described by Marchesi and colleagues [24] were also incorporated into each PCR assay to serve as an internal amplification control (IAC). This is a universal gene target which is highly conserved in bacteria [24]. All primers were synthesized by Integrated DNA Technologies (IDT) Pte. Ltd., Malaysia.

2.5. Template DNA Extraction. DNA from all bacteria isolates were extracted using DNeasy Blood & Tissue kit® (Qiagen, USA) according to the manufacturer's instructions. The purity and concentration of the extracted DNA were determined using Nanodrop Spectrophotometer ND-1000 (Thermo Fisher Scientific, USA). DNA concentration was measured from the absorbance at 260 nm. Ratio of the absorbance at 260 and 280 nm ($A_{260/280}$) and ratio of the absorbance at 230 and 260 nm ($A_{230/260}$) were used to evaluate the DNA quality. The extracted DNAs were diluted to a final stock concentration of 50 ng/μL using ultrapure water and stored at −20°C until ready for PCR amplification.

2.6. Optimization of PCR. Each PCR assay was optimized using a modified Taguchi method as described by Cobb and Clarkson [33]. The effects and interactions of the 4 main PCR components (IAC primers, *S.* Typhi-specific gene primers, $MgCl_2$, and annealing temperatures) each at 3 different levels (IAC primers: 0.05, 0.10, and 0.15 μM; *S.* Typhi primers: 1.00, 1.50, and 2.00 μM; $MgCl_2$: 2.00, 2.50, and 3.00 mM, and annealing temperatures: 50, 55, and 60°C) were investigated in a balanced orthogonal array of 9 experimental combinations. The PCR amplifications were carried out in a total reaction volume of 20 μL, and the PCR products were analysed on a 1.2% (w/v) agarose gel containing SYBR® Safe DNA Gel Stain (Invitrogen, USA), visualized using a blue-light transilluminator (Syngene, UK).

2.7. Analytical Specificities of Genes Unique to S. Typhi. Analytical specificities of the PCR assays were assessed by running each PCR assay on a panel of bacteria strains consisting of 39 *S.* Typhi, 62 non-Typhi *Salmonella*, and 10 non-*Salmonella* clinical isolates.

2.8. Detection Limit of the PCR Assays. Detection limit of the PCR assays was defined as the minimum amount of *S.* Typhi DNA (ng/μL) that yielded positive PCR amplicons. The assay sensitivities were determined by amplification of a 5-fold serial dilution of *S.* Typhi ATCC 7251 DNA, ranging from 50 ng to 25.6 fg. Two microliters of the DNA was subjected to PCR amplification. The analytical sensitivity was indicated by the presence of visible PCR product bands on the agarose gel using the transilluminator as described above.

2.9. DNA Sequencing. To confirm the PCR products were indeed derived from the *S.* Typhi strains, PCR amplicons from all assays produced using Phusion® High-Fidelity DNA Polymerase (New England Biolabs, USA) were purified and sent to First BASE Laboratories Pte. Ltd., Malaysia, for sequencing. The resultant nucleotide sequences were compared with the reference *S.* Typhi CT18 gene sequences in NCBI using BioEdit software.

Table 1: List of primers targeting *S.* Typhi-specific genes for the development of 6 PCR assays.

Target genes	Primer labels	Primer sequences (5′-3′)	Target lengths (bp)
STY0201	0201F	ATGCTTTTAAAAAACACAACATGG	1176
	0201R	TTACGGATAGGTGATTGAAAATTG	
STY0307	0307F	ATGAAACCTTTATTCTCAGTGC	495
	0307R	TTAGCGTAATTCCCAGAACC	
STY0322	0322F	ATGAAATATAAAAAAATAAGAG	678
	0322R	CTATGGATTCATTTCCATTTC	
STY0326	0326F	ATGAATACGAATAATTCACC	261
	0326R	TTACCCTCCCCATGTCAC	
STY2020	2020F	ATGCCTGTTATGCATAATTG	429
	2020R	TTATGCTGTTAACGAGTCGTC	
STY2021	2021F	ATGAGTTTAGCGCAGCCTAAATCC	732
	2021R	TTAGAAGTCTCCTGCCTGGAAAC	
16S rRNA[a]	16SF	CAGGCCTAACACATGCAAGTC	1362
	16SR	GGGCGGTGTGTACAAGGC	

[a]16S ribosomal RNA gene served as internal amplification control (IAC) [24].
F represents forward primer.
R represents reverse primer.

3. Results

Using the bioinformatic method for whole-genome comparison (Figure 1), 6 potential diagnostic markers with NCBI locus tags, STY0201, STY0307, STY0322, STY0326, STY2020, and STY2021, were found. They exhibit 100% query coverage and identity (E-value = 0) with all 6 *S.* Typhi gene sequences but had low or no significant similarity (E-value $\geq 1e^{-41}$) with other enteric bacteria nucleotide sequences as of 11 March 2016. These genes were found to be (bioinformatically) highly conserved and specific and thus were selected for further wet-lab validation using PCR method. The primers designed to amplify these selected genes are shown in Table 1.

The results showed that all 6 designed primer pairs successfully amplified their target genes with amplicon sizes of 1176, 495, 678, 261, 429, and 732 bps, respectively. DNA sequencing results of the amplicons showed 100% identity with their corresponding *S.* Typhi genes, confirming the fidelity and sensitivity of the primers.

The 6 single-gene target PCR assays were then optimized using Taguchi method with the incorporation of IAC which targeted the 16S rRNA gene. The optimized master mix for the PCR assays targeting STY0201, STY0307, and STY2020 genes consisted of 1x Green GoTaq Flexi Buffer, 2.0 mM $MgCl_2$, 0.2 mM dNTPs, 1.5 μM *S.* Typhi-specific gene primers, 0.10 μM IAC primers, 0.75 U GoTaq Flexi DNA Polymerase (Promega, USA), and 5% glycerol in a total volume of 20 μL. Two microliters of test DNA (50 ng/μL) was added to the master mix and amplified using the following optimized thermal-cycling parameters: initial denaturation at 95°C for 1 min, followed by 30 cycles elongation at 95°C for 30 s, 55°C for 30 s, 72°C for 1 min and a final extension at 72°C for 5 min. Similar PCR conditions were used for amplification of STY0322, STY0326, and STY2021 genes except for the concentration of $MgCl_2$ and IAC primers which were set at 3.0 mM and 0.15 μM, respectively. The optimal annealing temperature was set at 50°C. Under these conditions, the IAC primer pair produced an amplicon of 1,362 bp for all bacteria isolates tested (111/111).

The optimized PCR assays for STY0307, STY0322, STY0326, STY2020, and STY2021 correctly identified all *S.* Typhi (39/39) isolates, whereas none of the non-Typhi *Salmonella* (0/62) and none of the non-*Salmonella* (0/10) isolates were detected. This showed a 100% sensitivity and 100% specificity for the PCR assays (Table 2) and indicate that the 5 genes were unique to *S.* Typhi (Figures 2, 3, and 4).

The results of serial dilution of *S.* Typhi genomic DNA showed that the detection limit of the optimized PCR assays was 32 pg for gene STY0322, 6.4 pg for genes STY0326, STY2020, and STY2021, and 1.28 pg for gene STY0307.

Although gene STY0201 exhibited 100% sensitivity (detection of 39/39 *S.* Typhi isolates), it showed cross-reactivity with *S.* Oslo and *S.* Kissi (Table 2), resulting in a specificity of only 97.2% (detection of 2/72 of non-Typhi isolates). Sequencing of their PCR products showed a substitution of nucleotide C → T at position 89 and T → C at positions 354 and 1,026 for both *S.* Kissi and *S.* Oslo. The sequence variation between *S.* Kissi and *S.* Oslo with the *S.* Typhi CT18 reference genome was very small (only 3 nucleotide differences), indicating that the false positive results were due to sequence similarity among themselves.

4. Discussion

The diagnosis of typhoid fever based on clinical signs and symptoms is often ambiguous, while phenotypic detection of *S.* Typhi bacteria based on biochemical and serotyping methods is laborious and time-consuming. Thus, rapid molecular detection methods, such as nucleic acid-based amplification, such as PCR assay, is critically needed to help diagnose this contagious disease. Development of this test requires diagnostic markers that are sensitive and specific.

Table 2: Evaluation of the specificities of the 6 target genes for identification of *S.* Typhi using PCR (total of 111 clinical isolates).

Test bacteria strains	Positive PCR amplification for each target gene					
	STY0201	STY0307	STY0322	STY0326	STY2020	STY2021
S. Typhi ($n = 39$)	39/39	39/39	39/39	39/39	39/39	39/39
S. Paratyphi A ($n = 10$)	0/10	0/10	0/10	0/10	0/10	0/10
S. Paratyphi B ($n = 10$)	0/10	0/10	0/10	0/10	0/10	0/10
S. Paratyphi C ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Enteritidis ($n = 10$)	0/10	0/10	0/10	0/10	0/10	0/10
S. Typhimurium ($n = 10$)	0/10	0/10	0/10	0/10	0/10	0/10
S. Weltevreden ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Agona ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Hadar ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Heidelberg ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Poona ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Braenderup ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Albany ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Oslo ($n = 1$)	1/1	0/1	0/1	0/1	0/1	0/1
S. Kibi ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Newport ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Tshiongwe ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Uppsala ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Richmond ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Bardo ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Emek ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Kissi ($n = 1$)	1/1	0/1	0/1	0/1	0/1	0/1
S. Virchow ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Bordeaux ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Regent ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Java ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
S. Farsta ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
Shigella dysenteriae ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
Shigella flexneri ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
Shigella sonnei ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
Shigella boydii ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
Vibrio cholerae ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
Enterohemorrhagic E. coli ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
Enteropathogenic E. coli ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
Aeromonas hydrophila ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
Yersinia enterocolitica ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1
Klebsiella pneumoniae ($n = 1$)	0/1	0/1	0/1	0/1	0/1	0/1

This is the first report on the use of genes STY0307, STY0322, STY0326, STY2020, and STY2021 as *S.* Typhi-specific diagnostic markers. Unlike other *S.* Typhi PCR targets that were selected based on immunological properties, these genes are individually highly specific for *S.* Typhi and therefore can be used as single-gene target PCR assays without the need for nested or multiplex PCR. Also, these targets are whole gene sequences (from start to stop codon for the purpose of whole gene amplification) unlike other diagnostic markers which are only partial gene sequences. The idea of using this strategy is that if the whole gene sequence is specific to the bacteria then primers can be designed at any location of the gene. Thus, these gene sequences not only serve as specific targets for PCR assay, but also are suitable for more advance diagnostic tests that require multiple DNA sites, such as loop-mediated isothermal amplification (LAMP) and strand displacement amplification (SDA) which requires multiple primer annealing sites [34]. These genes could be utilized for the development of innovative Point-of-Care (POC) diagnostics to address the need for low-cost, simple, rapid, and accurate diagnostics for low resource settings.

The gene STY0201 has been used as a PCR target, and the PCR assays that were developed based on this gene were reported to be 100% sensitivity and specificity [35, 36].

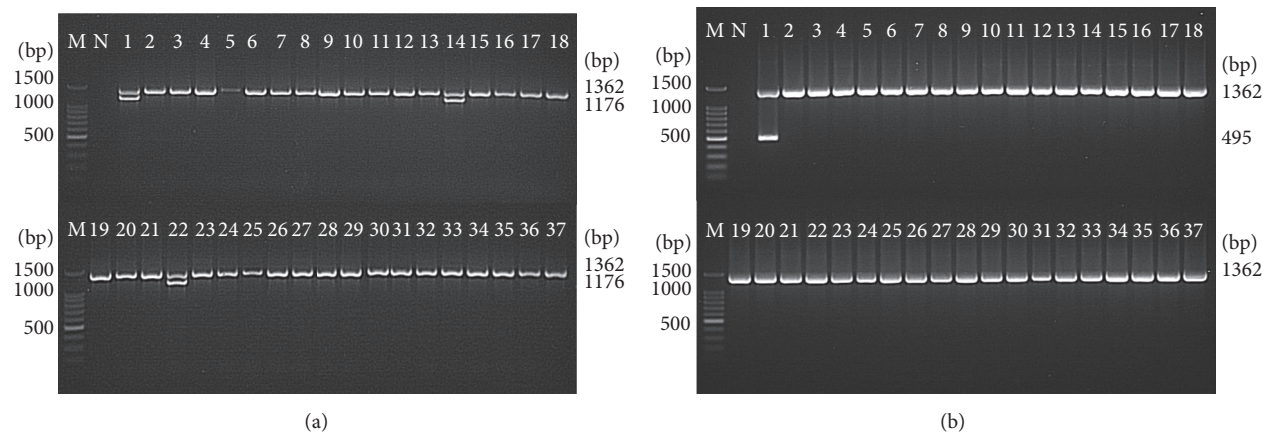


Figure 2: Analytical specificity of the PCR assay for detection of (a) STY0201 and (b) STY0307 genes (representative figures). Lane: M = 100 bp DNA ladder (Promega); N = negative control; 1 = *S.* Typhi; 2 = *S.* Paratyphi A; 3 = *S.* Paratyphi B; 4 = *S.* Paratyphi C; 5 = *S.* Enteritidis; 6 = *S.* Typhimurium; 7 = *S.* Weltevreden; 8 = *S.* Agona; 9 = *S.* Heidelberg; 10 = *S.* Poona; 11 = *S.* Hadar; 12 = *S.* Braenderup; 13 = *S.* Albany; 14 = *S.* Oslo; 15 = *S.* Kibi; 16 = *S.* Newport; 17 = *S.* Tshiongwe; 18 = *S.* Uppsala; 19 = *S.* Richmond; 20 = *S.* Bardo; 21 = *S.* Emek; 22 = *S.* Kissi; 23 = *S.* Virchow; 24 = *S.* Bordeaux; 25 = *S.* Regent; 26 = *S.* Java; 27 = *S.* Farsta; 28 = *Shigella dysenteriae*; 29 = *Shigella flexneri*; 30 = *Shigella sonnei*; 31 = *Shigella boydii*; 32 = *Vibrio cholerae*; 33 = *Enterohemorrhagic Escherichia coli*; 34 = *Enteropathogenic Escherichia coli*; 35 = *Aeromonas hydrophila*; 36 = *Yersinia enterocolitica*; and 37 = *Klebsiella pneumonia*. The PCR amplicon sizes for genes 16S rRNA, STY0201, and STY0307 were 1326 bp, 1176 bp, and 732 bp, respectively.

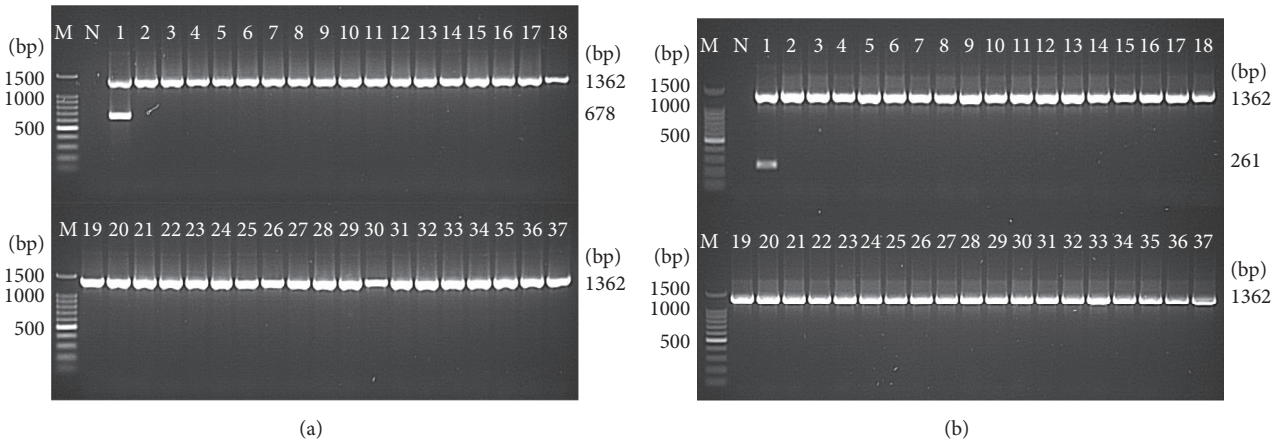


Figure 3: Analytical specificity of the PCR assay for detection of (a) STY0322 and (b) STY0326 genes, respectively (representative gels). Lane: M = 100 bp DNA ladder (Promega); N = negative control; 1 = *S.* Typhi; 2 = *S.* Paratyphi A; 3 = *S.* Paratyphi B; 4 = *S.* Paratyphi C; 5 = *S.* Enteritidis; 6 = *S.* Typhimurium; 7 = *S.* Weltevreden; 8 = *S.* Agona; 9 = *S.* Heidelberg; 10 = *S.* Poona; 11 = *S.* Hadar; 12 = *S.* Braenderup; 13 = *S.* Albany; 14 = *S.* Oslo; 15 = *S.* Kibi; 16 = *S.* Newport; 17 = *S.* Tshiongwe; 18 = *S.* Uppsala; 19 = *S.* Richmond; 20 = *S.* Bardo; 21 = *S.* Emek; 22 = *S.* Kissi; 23 = *S.* Virchow; 24 = *S.* Bordeaux; 25 = *S.* Regent; 26 = *S.* Java; 27 = *S.* Farsta; 28 = *Shigella dysenteriae*; 29 = *Shigella flexneri*; 30 = *Shigella sonnei*; 31 = *Shigella boydii*; 32 = *Vibrio cholerae*; 33 = *Enterohemorrhagic Escherichia coli*; 34 = *Enteropathogenic Escherichia coli*; 35 = *Aeromonas hydrophila*; 36 = *Yersinia enterocolitica*; and 37 = *Klebsiella pneumonia*. The PCR amplicon sizes for genes 16S rRNA, STY0322, and STY0326 were 1,326 bp, 678 bp, and 261 bp, respectively.

However, this study found that this gene was only 97.2% specific and cross-reacted with *S.* Oslo and *S.* Kissi. The incorrect bioinformatic prediction of the specificity of gene STY0201 may be due to the incomplete genome sequence available for the 2 bacteria in the NCBI database that limit the matching accuracy of the BLASTn search. This is a limitation of the alignment-based marker identification method, as it relies on the availability of a complete genome sequence. Thus, whenever new sequence data becomes available for the target organism, the bioinformatic analysis should be repeated to align the current diagnostic markers with the new gene sequence to ensure the specificity.

The other 5 genes identified in this study showed no sequence homology to proteins of known function using protein BLAST (BLASTp) programs. Genes STY0307, STY0322, and STY0326 encode for hypothetical proteins, while genes STY2020 and STY2021 encode for putative bacteriophage proteins. Interestingly, genes STY0307, STY0322, and STY0326 are located in the *Salmonella* Pathogenicity Island 6 (SPI-6). Yet, their role in bacteria virulence and

Table 3: Details of the 5 target genes and their description, antigenicity prediction, protein coverage, and identity with *S.* Paratyphi A.

Number	Target genes (NCBI locus tag)	Gene description	GC content (%)	Antigenicity prediction*	Protein coverage with *S.* Paratyphi A (%)	Protein identity with *S.* Paratyphi A (%)
1	STY0307	Hypothetical protein	43	0.66	0	0
2	STY0322	Hypothetical protein	29	0.37	21	33
3	STY0326	Conserved hypothetical protein	37	0.79	0	0
4	STY2020	Putative bacteriophage protein	42	0.66	0	0
5	STY2021	Putative bacteriophage protein	42	0.27	0	0

* Antigenicity of the *S.* Typhiproteins predicted using SCRATCH Protein Prediction software [26].

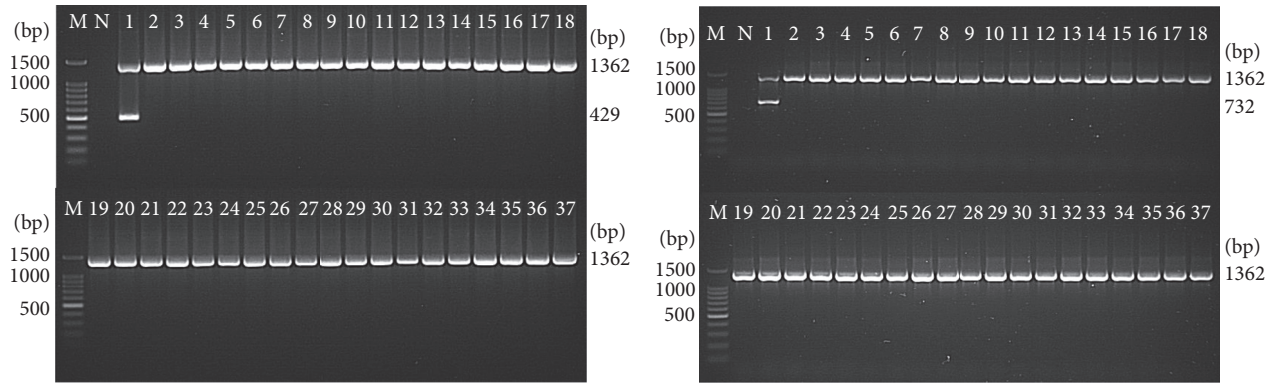


Figure 4: Analytical specificity of the PCR assay for detection of (a) STY2020 and (b) STY2021 genes, respectively (representative figures). Lane: M = 100 bp DNA ladder (Promega); N = negative control; 1 = *S.* Typhi; 2 = *S.* Paratyphi A; 3 = *S.* Paratyphi B; 4 = *S.* Paratyphi C; 5 = *S.* Enteritidis; 6 = *S.* Typhimurium; 7 = *S.* Weltevreden; 8 = *S.* Agona; 9 = *S.* Heidelberg; 10 = *S.* Poona; 11 = *S.* Hadar; 12 = *S.* Braenderup; 13 = *S.* Albany; 14 = *S.* Oslo; 15 = *S.* Kibi; 16 = *S.* Newport; 17 = *S.* Tshiongwe; 18 = *S.* Uppsala; 19 = *S.* Richmond; 20 = *S.* Bardo; 21 = *S.* Emek; 22 = *S.* Kissi; 23 = *S.* Virchow; 24 = *S.* Bordeaux; 25 = *S.* Regent; 26 = *S.* Java; 27 = *S.* Farsta; 28 = *Shigella dysenteriae*; 29 = *Shigella flexneri*; 30 = *Shigella sonnei*; 31 = *Shigella boydii*; 32 = *Vibrio cholerae*; 33 = *Enterohemorrhagic Escherichia coli*; 34 = *Enteropathogenic Escherichia coli*; 35 = *Aeromonas hydrophila*; 36 = *Yersinia enterocolitica*; and 37 = *Klebsiella pneumonia*. The PCR amplicon sizes for genes 16S rRNA, STY2020, and STY2021 were 1,326 bp, 429 bp, and 732 bp, respectively.

pathogenicity remains unknown. More importantly, antigenicity prediction scores using SCRATCH protein prediction software [26] showed that genes STY0201, STY0207, STY0307, STY0326, and STY2020 were highly antigenic and may have potential to serve as antigens for serodiagnosis of typhoid fever (Table 3). When compared with the deduced amino acid sequence of *S.* Paratyphi A, which is the closest relative of *S.* Typhi [37], the putative proteins showed weak or no similarity to *S.* Typhi (Table 3). These findings provide an opportunity for gene cloning and protein expression to investigate their serodiagnostic value for development of low-cost antibody-based diagnostic tests or vaccines for typhoid fever.

In conclusion, 5 *S.* Typhi-specific genes, namely, STY0307, STY0322, STY0326, STY2020, and STY2021, were found to be highly conserved among *S.* Typhi strains. Wet-lab experiments found no false positive reaction with non-Typhi serotypes or non-*Salmonella* enteric pathogens. These genes could serve as useful diagnostic markers for development of DNA-based diagnostics for sensitive and specific detection of typhoid fever.

Competing Interests

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

Acknowledgments

This work was supported by Universiti Sains Malaysia (USM) Research University Individual (RUI) Grant

(1001/CIPPM/812096) and the Ministry of Higher Education Prototype Research Grant Scheme (PRGS) (203/CIPPM/6740030). Yuan Xin Goay was a recipient of the National Science Fellowship (NSF) (M/0071/03/2010/S&T) and USM Fellowship [P-NFD0004/12(R)]. The authors would like to thank the Department of Clinical Microbiology and Parasitology, Hospital Universiti Sains Malaysia (HUSM), and INFORMM Biobank, Kelantan, Malaysia, for providing the bacterial strains.

References

[1] M. Akiba, M. Kusumoto, and T. Iwata, "Rapid identification of *Salmonella enterica* serovars, typhimurium, choleraesuis, infantis, hadar, enteritidis, dublin and gallinarum, by multiplex PCR," *Journal of Microbiological Methods*, vol. 85, no. 1, pp. 9–15, 2011.

[2] G. C. Buckle, C. L. Walker, and R. E. Black, "Typhoid fever and paratyphoid fever: systematic review to estimate global morbidity and mortality for 2010," *Journal of Global Health*, vol. 2, article 10401, 2012.

[3] T. C. Darton, C. J. Blohmke, and A. J. Pollard, "Typhoid epidemiology, diagnostics and the human challenge model," *Current Opinion in Gastroenterology*, vol. 30, no. 1, pp. 7–17, 2014.

[4] J. N. Ja'afar, Y. X. Goay, N. F. Mohammed Zaidi et al., "Epidemiological analysis of typhoid fever in Kelantan from a retrieved registry," *Malaysian Journal of Microbiology*, vol. 9, no. 2, pp. 147–151, 2013.

[5] T. V. T. Nga, A. Karkey, S. Dongol et al., "The sensitivity of real-time PCR amplification targeting invasive *Salmonella* serovars in biological specimens," *BMC Infectious Diseases*, vol. 10, article 125, 2010.

[6] M. D. Ganjali, P. Abdeshahian, K. Sudesh, and K. K. Phua, "Optimization of *Salmonella* Typhi biofilm assay on polypropylene microtiter plates using response surface methodology," *Biofouling*, vol. 32, no. 4, pp. 477–487, 2016.

[7] S. A. Zaki and S. Karande, "Multidrug-resistant typhoid fever: a review," *Journal of Infection in Developing Countries*, vol. 5, no. 5, pp. 324–337, 2011.

[8] A. I. Ugochukwu, O. C. Amu, and M. A. Nzegwu, "Ileal perforation due to typhoid fever—review of operative management and outcome in an urban centre in Nigeria," *International Journal of Surgery*, vol. 11, no. 3, pp. 218–222, 2013.

[9] M. Hayashi, T. Natori, S. Kubota-Hayashi et al., "A new protocol to detect multiple foodborne pathogens with PCR dipstick DNA chromatography after a six-hour enrichment culture in a broad-range food pathogen enrichment broth," *BioMed Research International*, vol. 2013, Article ID 295050, 10 pages, 2013.

[10] G. Kumar, C. B. Pratap, O. P. Mishra, K. Kumar, and G. Nath, "Use of urine with nested PCR targeting the flagellin gene (*fliC*) for diagnosis of typhoid fever," *Journal of Clinical Microbiology*, vol. 50, no. 6, pp. 1964–1967, 2012.

[11] J. Wang, Z. Q. Xu, P. H. Niu et al., "A two-tube multiplex reverse transcription PCR assay for simultaneous detection of viral and bacterial pathogens of infectious diarrhea," *BioMed Research International*, vol. 2014, Article ID 648520, 9 pages, 2014.

[12] K. Hirose, K.-I. Itoh, H. Nakajima et al., "Selective amplification of tyv (rfbE), prt (rfbS), viaB, and fliC genes by multiplex PCR for identification of *Salmonella enterica* serovars Typhi and Paratyphi A," *Journal of Clinical Microbiology*, vol. 40, no. 2, pp. 633–636, 2002.

[13] J.-H. Song, H. Cho, M. Y. Park, D. S. Na, H. B. Moon, and C. H. Pai, "Detection of Salmonella typhi in the blood of patients with typhoid fever by polymerase chain reaction," *Journal of Clinical Microbiology*, vol. 31, no. 6, pp. 1439–1443, 1993.

[14] S. Kolyva, H. Waxin, and M. Y. Popoff, "The Vi antigen of *Salmonella typhi*: molecular analysis of the viaB locus," *Journal of General Microbiology*, vol. 138, no. 2, pp. 297–304, 1992.

[15] G. Frankel, S. M. Newton, G. K. Schoolnik, and B. A. Stocker, "Intragenic recombination in a flagellin gene: characterization of the *H1-j* gene of *Salmonella typhi*," *The EMBO Journal*, vol. 8, pp. 3149–3152, 1989.

[16] H. M. Seth-Smith, "SPI-7: salmonella's Vi-encoding Pathogenicity Island," *Journal of Infection in Developing Countries*, vol. 2, no. 4, pp. 267–271, 2008.

[17] E. M. Daniels, R. Schneerson, W. M. Egan, S. C. Szu, and J. B. Robbins, "Characterization of the Salmonella paratyphi C Vi polysaccharide," *Infection and Immunity*, vol. 57, pp. 3159–3164, 1989.

[18] A. Kumar, Y. Balachandran, S. Gupta, S. Khare, and Suman, "Quick PCR based diagnosis of typhoid using specific genetic markers," *Biotechnology Letters*, vol. 32, no. 5, pp. 707–712, 2010.

[19] S. Khan, B. N. Harish, G. A. Menezes, N. S. Acharya, and S. C. Parija, "Early diagnosis of typhoid fever by nested PCR for flagellin gene of salmonella enterica serotype typhi," *Indian Journal of Medical Research*, vol. 136, no. 5, pp. 850–854, 2012.

[20] P. G. Agron, R. L. Walker, H. Kinde et al., "Identification by subtractive hybridization of sequences specific for salmonella enterica serovar enteritidis," *Applied and Environmental Microbiology*, vol. 67, no. 3–12, pp. 4984–4991, 2001.

[21] P. Leekitcharoenphon, E. M. Nielsen, R. S. Kaas, O. Lund, and F. M. Aarestrup, "Evaluation of whole genome sequencing for outbreak detection of *Salmonella enterica*," *PLoS ONE*, vol. 9, no. 2, Article ID e87991, 2014.

[22] S. Porwollik, R. M.-Y. Wong, and M. McClelland, "Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 13, pp. 8956–8961, 2002.

[23] F. Poly, D. Threadgill, and A. Stintzi, "Identification of *Campylobacter jejuni* ATCC 43431-specific genes by whole microbial genome comparisons," *Journal of Bacteriology*, vol. 186, no. 14, pp. 4781–4795, 2004.

[24] J. R. Marchesi, T. Sato, A. J. Weightman et al., "Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA," *Applied and Environmental Microbiology*, vol. 64, no. 2, pp. 795–799, 1998.

[25] N. F. Kamaruzzaman, N. J. Jaafar, M. H. Hani et al., "Pulsed-field gel electrophoresis analysis of *Salmonella enterica* serovar typhi isolates in the north-east region of Peninsular Malaysia between 2002 and 2009," *Journal of Applied Life Sciences International*, vol. 5, no. 2, pp. 2394–1103, 2016.

[26] C. N. Magnan, M. Zeller, M. A. Kayala et al., "High-throughput prediction of protein antigenicity using protein microarray data," *Bioinformatics*, vol. 26, no. 23, pp. 2936–2943, 2010.

[27] J. Parkhill, G. Dougan, K. D. James et al., "Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18," *Nature*, vol. 413, no. 6858, pp. 848–852, 2001.

[28] W. Deng, S.-R. Liou, G. Plunkett III et al., "Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18," *Journal of Bacteriology*, vol. 185, no. 7, pp. 2330–2337, 2003.

[29] S. Y. Ong, C. B. Pratap, X. Wan et al., "Complete genome sequence of Salmonella enterica subsp. Enterica serovar typhi P-stx-12," *Journal of Bacteriology*, vol. 194, no. 8, pp. 2115–2116, 2012.

[30] D. Xu, J. O. Cisar, F. Poly et al., "Genome sequence of *Salmonella enterica* serovar Typhi oral vaccine strain Ty21a," *Genome Announcements*, vol. 1, no. 4, Article ID e00650-13, 2013.

[31] S. M. Harish, K. S. Sim, F. Mohd Nor et al., "Complete genome sequence of *Salmonella enterica* subsp. *enterica* serovar Typhi isolate B/SF/13/03/195 associated with a Typhoid carrier in pasir mas, Kelantan, Malaysia," *Genome Announcements*, vol. 3, no. 6, Article ID e01285-15, 2015.

[32] S. Muhamad Harish, K.-S. Sim, N. Najimudin, and I. Aziah, "Genome sequence of *Salmonella enterica* subsp. *enterica* serovar Typhi isolate PM016/13 from untreated well water associated with a typhoid outbreak in pasir mas, Kelantan, Malaysia," *Genome Announcements*, vol. 3, no. 6, Article ID e01261-15, 2015.

[33] B. D. Cobb and J. M. Clarkson, "A simple procedure for optimising the polymerase chain reaction (PCR) using modified Taguchi methods," *Nucleic Acids Research*, vol. 22, no. 18, pp. 3801–3805, 1994.

[34] C.-C. Chang, C.-C. Chen, S.-C. Wei, H.-H. Lu, Y.-H. Liang, and C.-W. Lin, "Diagnostic devices for isothermal nucleic acid amplification," *Sensors (Switzerland)*, vol. 12, no. 6, pp. 8319–8337, 2012.

[35] C. B. Pratap, G. Kumar, S. K. Patel et al., "Targeting of putative fimbrial gene for detection of *S. typhi* in typhoid fever and chronic typhoid carriers by nested PCR," *Journal of Infection in Developing Countries*, vol. 7, no. 7, pp. 520–527, 2013.

[36] G. J. Yin Ngan, L. M. Ng, R. T. P. Lin, and J. W. P. Teo, "Development of a novel multiplex PCR for the detection and differentiation of *Salmonella enterica* serovars Typhi and Paratyphi A," *Research in Microbiology*, vol. 161, no. 4, pp. 243–248, 2010.

[37] K. E. Holt, N. R. Thomson, J. Wain et al., "Pseudogene accumulation in the evolutionary histories of *Salmonella enterica* serovars Paratyphi A and Typhi," *BMC Genomics*, vol. 10, article 36, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 8569129, 9 pages
http://dx.doi.org/10.1155/2016/8569129

Research Article

Single-Trial Sparse Representation-Based Approach for VEP Extraction

Nannan Yu,[1] Funian Hu,[1] Dexuan Zou,[1] Qisheng Ding,[1] and Hanbing Lu[2]

[1]*School of Electrical Engineering and Automation, Jiangsu Normal University, Xuzhou 221116, China*
[2]*Department of Internal Neurology, Xuzhou Central Hospital, Xuzhou 221116, China*

Correspondence should be addressed to Hanbing Lu; luhanbing111@126.com

Received 7 June 2016; Revised 25 August 2016; Accepted 14 September 2016

Academic Editor: Tun-Wen Pai

Sparse representation is a powerful tool in signal denoising, and visual evoked potentials (VEPs) have been proven to have strong sparsity over an appropriate dictionary. Inspired by this idea, we present in this paper a novel sparse representation-based approach to solving the VEP extraction problem. The extraction process is performed in three stages. First, instead of using the mixed signals containing the electroencephalogram (EEG) and VEPs, we utilise an EEG from a previous trial, which did not contain VEPs, to identify the parameters of the EEG autoregressive (AR) model. Second, instead of the moving average (MA) model, sparse representation is used to model the VEPs in the autoregressive-moving average (ARMA) model. Finally, we calculate the sparse coefficients and derive VEPs by using the AR model. Next, we tested the performance of the proposed algorithm with synthetic and real data, after which we compared the results with that of an AR model with exogenous input modelling and a mixed overcomplete dictionary-based sparse component decomposition method. Utilising the synthetic data, the algorithms are then employed to estimate the latencies of P100 of the VEPs corrupted by added simulated EEG at different signal-to-noise ratio (SNR) values. The validations demonstrate that our method can well preserve the details of the VEPs for latency estimation, even in low SNR environments.

1. Introduction

Evoked potentials (EPs) are bioelectrical signals that are generated by the central nervous system when the latter is stimulated by well-defined external stimuli. Depending on the modality of stimulation, EPs are categorised into auditory evoked potential (AEP), visual evoked potential (VEP), and somatosensory evoked potential (SEP). In clinical environments, these signals are used to reflect the various functions of auditory, optic, and sensory nerve sense-conducting pathways. In this paper, we concentrate on the second type, namely, the VEPs. Generally speaking, there exist three prominent components (N75, P100, and N145) in the VEP signal, whereas the preceding and following segments are almost flat. Of the three components, the P100 wave is the most significant and stable; hence, it is the most important component in clinical applications [1].

VEP signals have time-locked (quasiperiodic) characteristics and are always accompanied by ongoing electroencephalogram (EEG) signals. Moreover, the signal-to-noise ratio (SNR) of VEP records is usually low (−5 to −10 dB). Ensemble averaging (EA) is the most widely used method for estimating VEP against a noisy background. However, EA cannot be used to detect latency and amplitude variations from one trial to another; thus, single-trial analysis is better suited for investigations into the dynamics of brain activation. The single-trial VEP estimation is very meaningful in cognitive science research and clinical applications, such as brain-computer interfacing and intraoperative monitoring [2].

Many single-trial EP estimation methods have been proposed over the past two decades. These methods can be divided into two categories, namely, denoising methods and separation methods. The denoising methods assume that the measurement of the VEP is corrupted by noise and that the main source of noise is the EEG. Many conventional denoising methods have been applied, such as the Wiener filter [3], Kalman filter [4], and ARX [5]. Among these methods, ARX is widely recognised and has previously been applied to monitor the depth of anaesthesia during surgery. In ARX, the EEG

can be viewed as an autoregressive (AR) model driven by white noise, and the EP can be modelled by an ARMA filter with a known signal accurately. The known signal is typically the average of the reference EPs (AREP). The orders and parameters of the AR and ARMA models can be estimated by utilising various optimisation techniques, such as the final prediction error (FPE) [6] and the least-squares (LS) method [7]. The EPs can then be reconstructed by ARMA filtering with the AREP. Recently, Cerutti et al. [6] found that EP extraction using ARX modelling is only capable of extracting latency EP variations in relatively high SNRs and that it is completely invalid because the latency varies greatly compared with the AREP from systemic experiments. The separation methods separate the VEP and EEG signals by modelling them based on their characteristics, such as wavelet transformation and sparse representation.

Meanwhile, Causevic et al. [8] and Martazi et al. [9] used wavelet transformation to separate the EP and EEG signals. Sparse coding is a powerful tool in analysing non-stationary signals, and it has shown significant success in signal denoising and separation. Xu and Yao [10] proposed the mixed overcomplete dictionary-based sparse component decomposition method (MOSCA), which decomposes the EP and EEG signals in the wavelet dictionary and discrete cosine transform (DCT) dictionary, respectively. However, given that EEG is not considered white noise and that many components of EP and EEG look alike in a single trial, their components are represented by the wrong dictionaries and their corresponding coefficients. Therefore, MOSCA cannot separate the EP and EEG signals sufficiently [11, 12].

In this paper, we present a novel sparse representation-based approach to solving the VEP extraction problem. Instead, of the mixed signals from the EEG and EP, we utilised an EEG in a previous trial that did not contain VEP to identify the parameters of the EEG AR model. Then, we used sparse representation in the ARMA model, instead of MA, to simulate the VEP. The sparse coefficients can be calculated by an optimisation method. Finally, the VEP can be derived from the AR model. Experiments carried out on synthetic and real data confirm the superior performance of our method. The rest of the paper is organised as follows. Section 2 provides the details of our single-trial estimation algorithm. Section 3 contains experimental results obtained from the proposed method and a comparison with ARX and MOSCA. Section 4 provides the conclusions.

2. Method

Let the VEP signal $p(k) \in R^{N\times 1}$ to be estimated be corrupted by noise from ongoing background activities. The main source of noise is the spontaneous EEG $e(k) \in R^{N\times 1}$. The measurement $s(k) \in R^{N\times 1}$ is given by

$$s(k) = p(k) + e(k). \tag{1}$$

We need to design a method that can remove the noise from $s(k)$, getting as close as possible to the original EP signal $p(k)$ [13].

2.1. The VEP Signal. In ARX, VEP $p(k)$ is derived by filtering the reference $u(k) \in R^{N\times 1}$, which is chosen to be the average of a sufficient number of trials and can represent the general form of the evoked response under analysis, by the ARMA model parameters; that is,

$$\widehat{P}(z) = \left(\frac{B\left(z^{-1}\right)}{A\left(z^{-1}\right)}\right) U(z), \tag{2}$$

where $\widehat{P}(z)$ and $U(z)$ are the z-transform of $\widehat{p}(k)$ and $u(k)$ and $A(z^{-1}) = 1 - \sum_{i=1}^{n} a_i z^{-i}$, $B(z^{-1}) = z^{-d} \sum_{j=0}^{m-1} b_j z^{-j}$. Sparse coding is a powerful tool for the analysis of nonstationary signals; it has achieved significant success in signal denoising and separation. Compared with ARMA, sparse coding is more flexible and uses the dictionary and the corresponding coefficient to represent signals. VEP has been proven to have strong sparsity over an appropriate dictionary in our previous paper [12]. Thus, in the current paper, we use sparse coding to represent the single-trial VEP instead of the MA model in ARMA. Therefore, formula (2) can be rewritten as

$$\widehat{p}(k) = \frac{B\left(z^{-1}\right) u(k)}{A\left(z^{-1}\right)} = \frac{G\theta}{A\left(z^{-1}\right)}, \tag{3}$$

where $G \in R^{N\times M}$ and $\theta \in R^{M\times 1}$ are the dictionary and sparse coefficient of $B(z^{-1})u(k)$, respectively. The transfer function $B(z^{-1})/A(z^{-1})$ merely represents a mechanism to incorporate deterministic VEP $p(k)$ variations into the reference signal $u(k)$, rather than a physiologically meaningful process.

2.2. Dictionary Construction. Inspired by the modelling method in [14], we proposed a dictionary construction method for the EP signal, as reported in our previous paper. This method assumes that the atoms in the dictionary can be extracted from a reference signal and that the single-trial EP can be decomposed sparsely by the dictionary. Many previous experiments have demonstrated this result.

The reference signal $u(k)$ consists of a superposition of M components expressed as

$$u(k) = \sum_{m=1}^{M} a_m s_m(k). \tag{4}$$

$u(k)$ can be acquired by AREP. $s_m(k)$ can be extracted from $u(k)$ using a certain filtering window function, such as Hamming window and Blackman window. The central location and width of the window are determined by the location of point of peak (and valley) amplitude and peak (and valley) width of the mth component. The dictionary can be represented by

$$D = \left(S_1 \;\; S_2 \;\; \cdots \;\; S_M\right), \tag{5}$$

where $S_m \in R^{N\times 2d}$, and

$$S_m^T = \begin{pmatrix} s_m(d) & \cdots & \cdots & s_m(N) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_m(2) & s_m(3) & \cdots & \cdots & s_m(N) & 0 \\ s_m(1) & s_m(2) & s_m(3) & \cdots & \cdots & s_m(N) \\ 0 & s_m(1) & s_m(2) & s_m(3) & \cdots & s_m(N-1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & s_m(1) & \cdots & s_m(N-d) \end{pmatrix}. \quad (6)$$

Then, $u(k)$ can be represented by the dictionary D and the coefficient θ_1, as

$$u(k) = D\theta_1. \quad (7)$$

In this paper, we aim to construct the dictionary $B(z^{-1})u(k)$. The transfer function $B(z^{-1})$ represents a mechanism, which incorporates deterministic single-trial EP variations into the reference signal, rather than a physiologically meaningful process. From this, it follows that

$$B\left(z^{-1}\right)u(k) = \sum_{l=0}^{m+d-1} b_l u(k-l), \quad (8)$$

where m and d are usually small positive integers. Given that $u(k)$ is sparse on dictionary D, $B(z^{-1})u(k)$ is also sparse on dictionary D. Thus, in this paper, $G = D$.

2.3. EEG Signal. Similar to ARX, in this paper, the EEG $e(k)$ is viewed as an AR model driven by white noise $w(k)$; that is,

$$\hat{e}(k) = \frac{1}{A\left(z^{-1}\right)}w(k). \quad (9)$$

The parameters can be estimated using the least-squares method. We assume that the statistical characteristics of the EEG in the successive trials are similar, as has been reported in many papers [7, 8]. Thus, in the current paper, instead of the mixed signal of EEG and EP, we utilise the EEG from a previous trial, to estimate the parameters of AR model. This EEG does not contain EP.

2.4. Single-Trial Extraction. Substituting formulas (3) and (5) into formula (1), we get

$$\hat{s}(k) = \hat{p}(k) + \hat{e}(k) = \frac{G\theta}{A\left(z^{-1}\right)} + \frac{1}{A\left(z^{-1}\right)}w(k). \quad (10)$$

Then,

$$A\left(z^{-1}\right)\hat{s}(k) = G\theta + w(k). \quad (11)$$

Let $x(k) = A(z^{-1})\hat{s}(k)$; then, formula (11) can be simplified as

$$x(k) = G\theta + w(k). \quad (12)$$

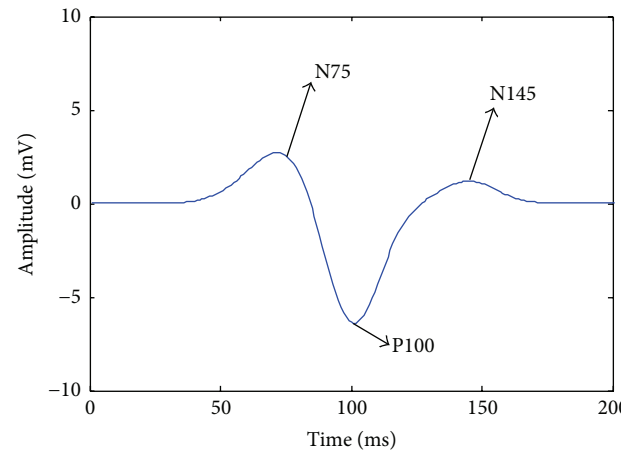


Figure 1: Simulated EP indicating three components: N75, P100, and N145.

Hence,

$$\hat{\theta} = \arg\min_{\theta} \|\theta\|_0 \quad \text{s.t.} \quad \|x(k) - G\cdot\theta\|_2 \le \varepsilon_0, \quad (13)$$

where ε_0 is determined by the variance of the EEG. Formula (10) can be solved by using optimisation methods, such as basis pursuit (BP) [15], orthonormal matching pursuit (OMP) [16], and Lasso [17].

The single-trial VEP can then be reconstructed by using

$$\hat{p}(k) = \frac{G\hat{\theta}}{A\left(z^{-1}\right)}. \quad (14)$$

3. Experimental Results

3.1. Analysis of the Simulations. Computer simulation is conducted to verify the performance of our proposed VEP signal extraction method. Depending on the characteristics of the VEP, the simulated VEP is constructed with three components and is expressed as

$$p(k) = 3\exp\left(-\frac{(k-75)^2}{20^2}\right) - 7\exp\left(-\frac{(k-(100+m))^2}{15^2}\right) + 1.2\exp\left(-\frac{(k-145)^2}{15^2}\right). \quad (15)$$

The three Gaussian functions represent a prominent VEP with similar morphological characteristics to those of the negative (N75), positive (P100), and negative (N145) peaks of a real VEP, respectively. The simulated VEP is shown in Figure 1.

The background EEG that is superimposed on the EP signal is simulated by an AR process [18], which is given by

$$e(k) = 1.5084e(k-1) - 0.1587e(k-2) - 0.3109e(k-3) - 0.0510e(k-4) + w(k), \quad (16)$$

where $w(t)$ is the Gaussian white noise. The simulated VEP is shown in Figure 1.

In this paper, we assume that the AR parameters of spontaneous EEG in two consecutive trials are extremely similar. In order to validate this assumption, three consecutive trials of spontaneous EEG signals $e_i(k)$ $(i = 1, 2, 3)$ are chosen randomly for the experiment. We compute their least-squares AR model with an approach. We set

$$\begin{aligned} A_1\left(z^{-1}\right) &= 1 - 3.106z^{-1} + 4.291z^{-2} - 3.465z^{-3} + 1.683z^{-4} - 0.389z^{-5}, \\ A_2\left(z^{-1}\right) &= 1 - 3.138z^{-1} + 4.404z^{-2} - 3.621z^{-3} + 1.759z^{-4} - 0.399z^{-5}, \\ A_3\left(z^{-1}\right) &= 1 - 3.143z^{-1} + 4.504z^{-2} - 3.887z^{-3} + 1.643z^{-4} - 0.410z^{-5}. \end{aligned} \quad (17)$$

Then, each $A_i(z^{-1})$ is used to transform the EEG signal in the other trial, so these parameters are, respectively, changed as

$$\begin{aligned} A_1'\left(z^{-1}\right) &= A_3\left(z^{-1}\right), \\ A_2'\left(z^{-1}\right) &= A_1\left(z^{-1}\right), \\ A_3'\left(z^{-1}\right) &= A_2\left(z^{-1}\right). \end{aligned} \quad (18)$$

The three EEG signals are transformed by $A_i'(z^{-1})$; that is, $W_i(z) = A_i'(z^{-1})E_i(z)$, where $W_i(z)$ and $E_i(z)$ are the z-transform of $w_i(k)$ and $e_i(k)$, respectively. In Figures 2 and 3, we, respectively, provide the frequency content and independence of $w_i(k)$. As can be seen in the figures, compared with $e_1(k)$, the energy of each frequency band of $w_i(k)$ is more uniform, and the autocorrelation coefficients of $w_i(k)$ are lower.

Let $A_i \in R^{6\times 1}$ represent the parameter vectors of $A_i(z^{-1})$ and $d_{ij} = [\sum_{l=1}^{6}((A_{il} - A_{jl})/A_{jl})^2]^{1/2}$ represent the difference between A_i and A_j. From formula (13), we obtain $d_{12} = 0.0743$ and $d_{13} = 0.1626$. In order to test the robustness of the proposed method in case inaccurate estimations of the AR coefficients are obtained, we use $e_1(k)$ to estimate the AR parameter and $e_2(k)$ and $e_3(k)$ to generate the measurements $s_2(k)$ and $s_3(k)$, respectively. Then, the extracted VEP2 and VEP3 from $s_2(k)$ and $s_3(k)$ with our proposed method are shown in Figure 4. With our method, the VEP2 and VEP3 are extracted from $s_2(k)$ and $s_3(k)$. As shown in the figure, when SNR = −5 dB, both VEP2 and VEP3 show results that approach the simulated VEP.

Table 1: The SNR of VEP2 and VEP3 extracted by our method.

SNR (dB)	VEP2		VEP3	
	Mean	Standard deviation	Mean	Standard deviation
−5	9.98	0.07	9.47	0.08
−10	5.48	0.24	5.24	0.17

Table 1 shows the mean and standard deviation of SNR obtained from 100 extracted VEP2 or VEP3. As shown in the table, with the same SNR, the SNR values of VEP2 and VEP3 are similar, although d_{13} is two times larger than d_{12}.

During estimation, the observed SNR values may change over time due to the nonstationary characteristics of the EEG. Therefore, in this experiment, the performance of our method is examined under various SNR conditions. The EEGs are generated with formula (16).

As shown in Figure 5, although the estimation performance degrades with decreasing SNR, the prominent morphological characteristics (N75, P100, and N145) are preserved in all the SNR values.

The average values of SNR obtained with our method, MOSCA [10], and ARX [6] are shown in Figure 6. The results are acquired by identifying the average of 100 trials for each piece of data. In our method, the dictionary of VEP is constructed by using formula (15) where $m = 0$. Similarly, in ARX, the reference VEP is generated by formula (15) where $m = 0$. In this experiment, we change the latency of P100 by setting $m = 5$ and $m = 10$. We can see from this figure that our method consistently demonstrates the greatest improvement in all methods. Compared with sparse coding, ARX ($m = 5$) shows superior performance at low initial SNRs. However, when the latencies change greatly ($m = 10$), ARX method degrades seriously. We can also see that the latency change has hardly any impact on the estimation performance of our method.

To increase the objectivity of the evaluation, for each SNR and m, we generate data from 50 trials and then estimate the latencies of P100. As shown in Table 2, we change m from −10 to 10 and the SNR from −10 dB to 0 dB and estimate the single-trial VEP signal. Results show that, with the decrease of SNR, all standard deviations also increase. The RMSE value depends primarily on the SNR, rather than on the variations of latency (m), thereby indicating that our method is appropriate for tracking the latency variations when SNR ≥ −10 dB.

3.2. Analysis of the Real VEP. To further evaluate the performance of our method, we collected VEPs from three pairs of eyes of three human subjects during pattern reversal VEP experiments. The basic data obtained from the three subjects are shown in Table 3.

An example from the 50 trials is selected randomly from the original recorded VEPs of subject 2's right eye. Figure 7 shows the corresponding average VEP.

We extract the VEPs with our method, MOSCA, and ARX, and the results are shown in Figure 8. Clearly, the three components N75, P100, and N145 of VEPs extracted with our

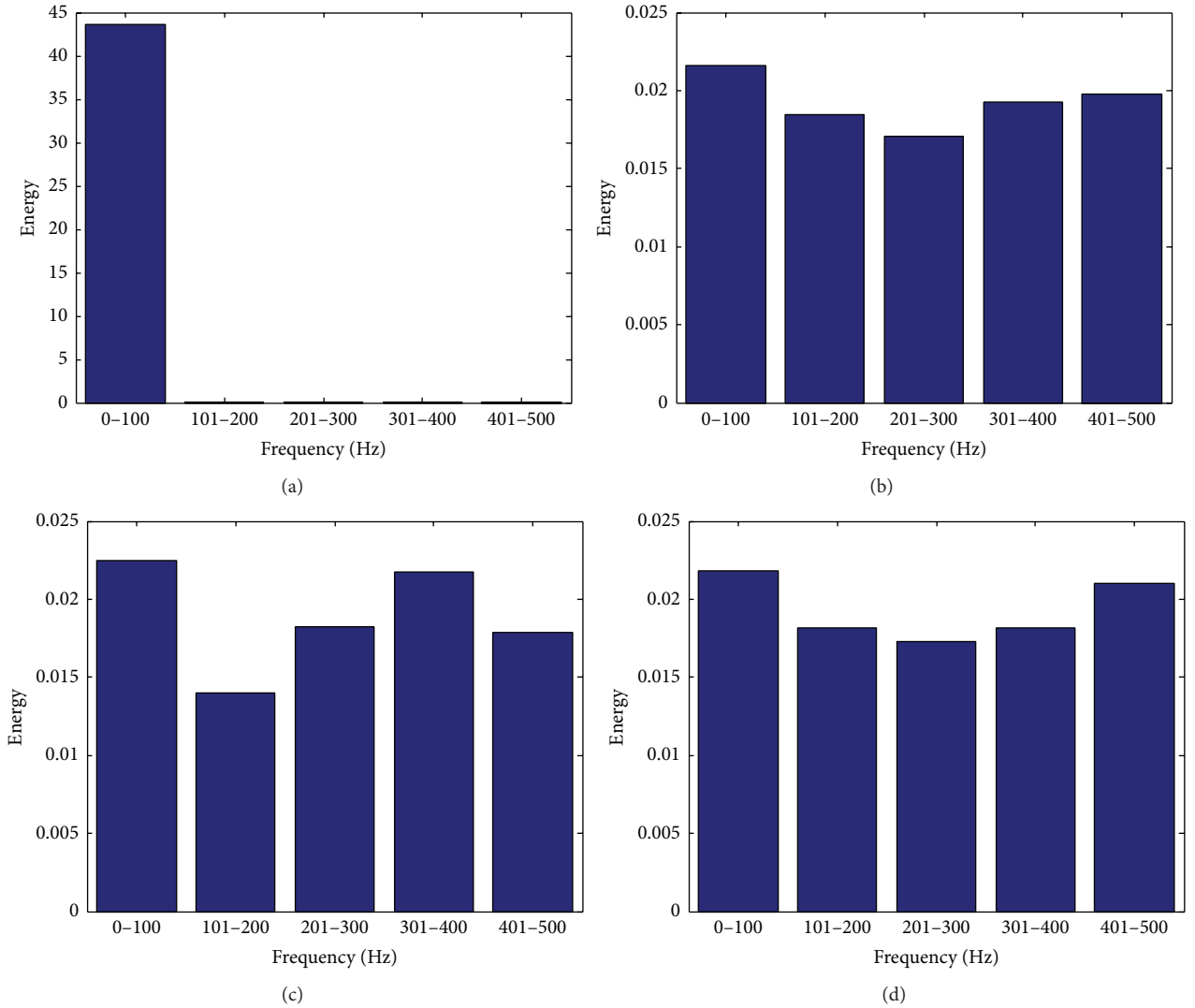


Figure 2: The frequency band energy spectrum. (a) EEG $e_1(k)$; (b) $w_1(k)$; (c) $w_2(k)$; (d) $w_3(k)$.

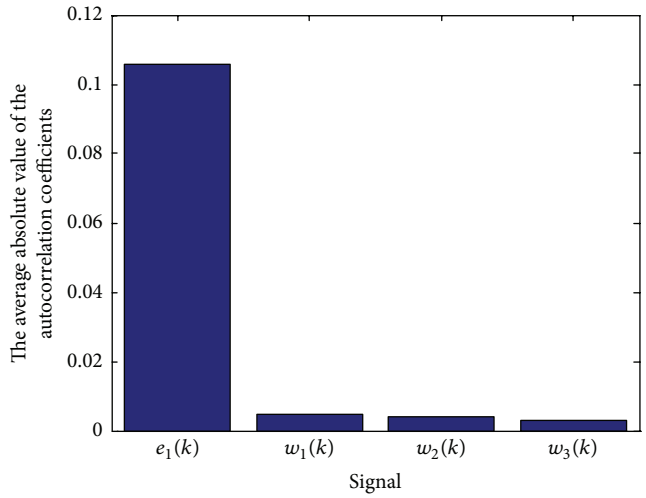


Figure 3: The average absolute value of the autocorrelation coefficients.

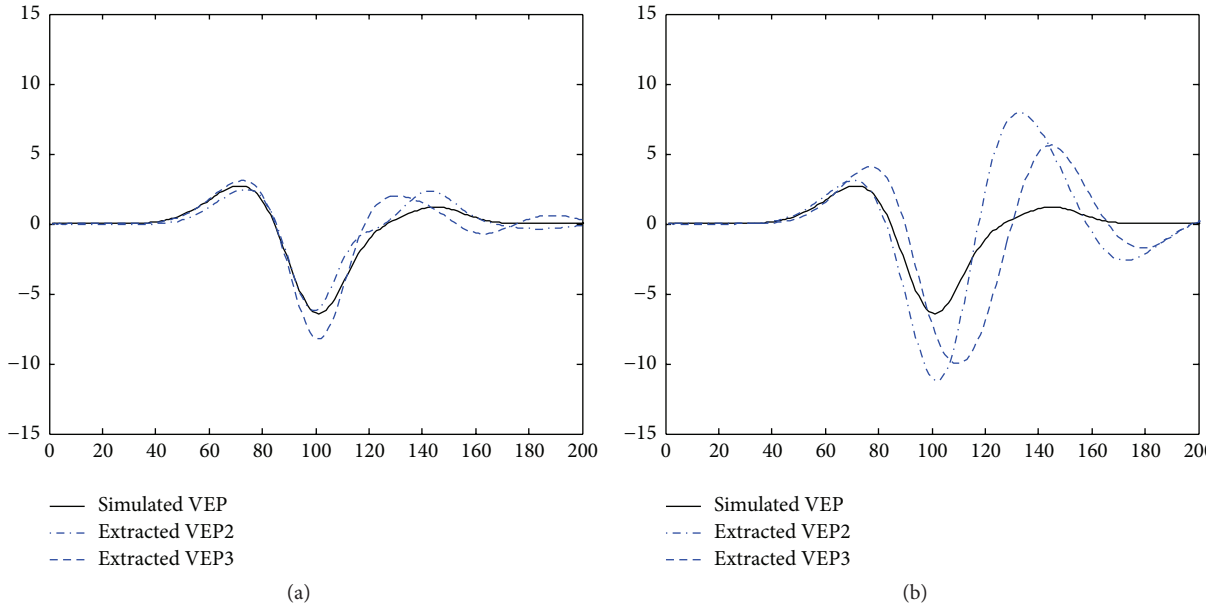


Figure 4: The VEP2 and VEP3 extracted with different SNR values. (a) SNR = −5 dB; (b) SNR = −10 dB.

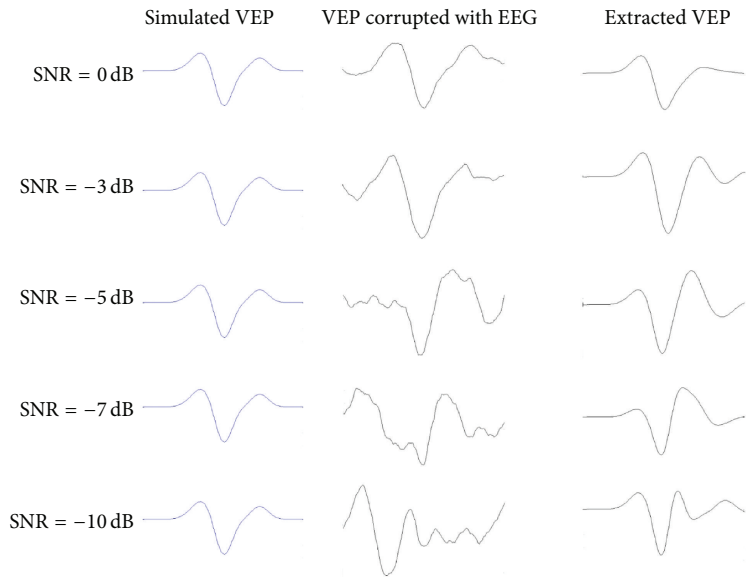


Figure 5: Single-trial VEP extracted by our method with different SNR values.

Table 2: Latency of P100 extracted by our method.

m	SNR (dB)							
	0		−3		−5		−10	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
−10	90.8 ms	1.4 ms	90.4 ms	3.2 ms	89.2 ms	5.8 ms	91.1 ms	6.2 ms
−5	94.9 ms	1.4 ms	95.5 ms	3.9 ms	94.5 ms	5.9 ms	95.7 ms	6.6 ms
0	100.8 ms	1.8 ms	100.7 ms	3.4 ms	102.8 ms	4.2 ms	101.3 ms	6.2 ms
5	104.7 ms	1.5 ms	105.2 ms	3.8 ms	103.4 ms	5.7 ms	104.7 ms	6.4 ms
10	110.7 ms	1.9 ms	110.2 ms	3.9 ms	112.4 ms	6.1 ms	112.4 ms	7.4 ms

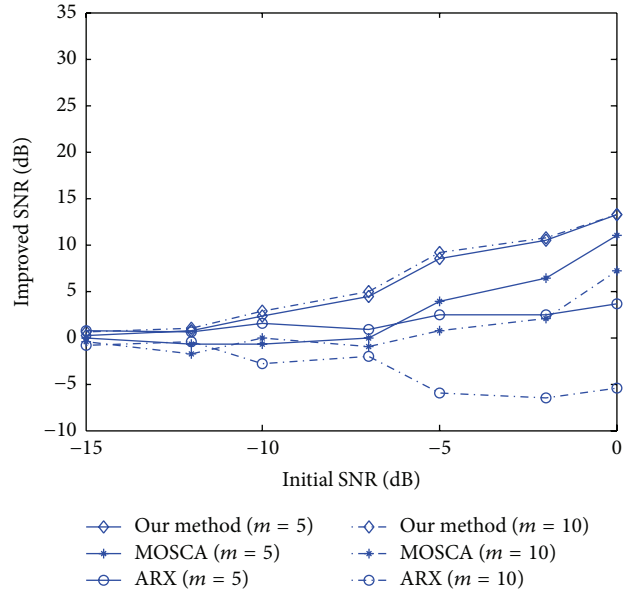


Figure 6: The improved SNR of VEP with different latencies of P100 using our method, MOSCA, and ARX.

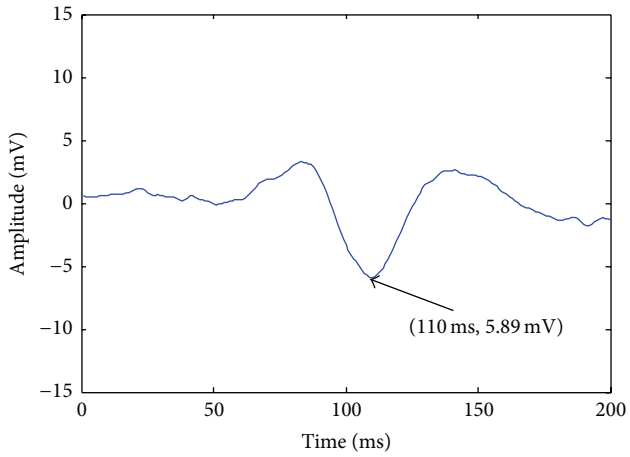


Figure 7: The average VEP.

Table 3: The basic data of 3 subjects.

Subject	Age	Sex	Vision
1	25	F	Normal
2	24	M	Normal
3	25	F	Normal

method are prominent, and the signals estimated using our method are more similar to the ensemble average signal.

Then, our method is used to estimate the amplitudes and latencies of P100 in 200 trials. As shown in Figure 9, the variations in amplitudes and latencies are significant, whereas most amplitudes are between −7 and −4 and most latencies are between 100 and 115. These results have good agreement with those observed in practice.

4. Conclusions

Single-trial EP estimation is a very useful tool in cognitive science-related studies and clinical applications. Many investigations have been carried out and some amount of success has been achieved. However, only a few practical methods have been proposed. ARX modelling is a classical method that has been applied in clinical practice for several years. However, this method has limitations regarding the tracking of latency variations and is only capable of extracting latency variations of an EP under relatively high SNR values. Meanwhile, sparse coding is a powerful tool in signal denoising, and EPs have been proven to have strong sparsity over an appropriate dictionary. Inspired by this idea, in this paper, we introduce sparse coding into the ARX model and propose a novel single-trial VEP extraction method based on ARX and sparse coding. Compared with ARMA, sparse coding is more flexible. It uses the best matching atoms from the dictionary to represent the EP signal without needing to estimate the number of atoms beforehand. By transforming the electroencephalography signal into white noise, the single-trial EP estimation is transformed into a signal denoising problem for white noise. With the dictionary constructed specially for EPs, the EP signal can be extracted easily with sparse coding. Moreover, since the location of the atom in the dictionary has no influence on the effectiveness of sparse decomposition, variations of the amplitude and latency of EPs have only a minor impact on the performance of the proposed method. The proposed method can thus track EP signal variations. We conducted a series of experiments on synthetic and real data, and the results have been evaluated using waveform observation and several metrics. The validations demonstrate that our method can well preserve the EP details of latency and amplitude estimation simultaneously, even under low SNR conditions.

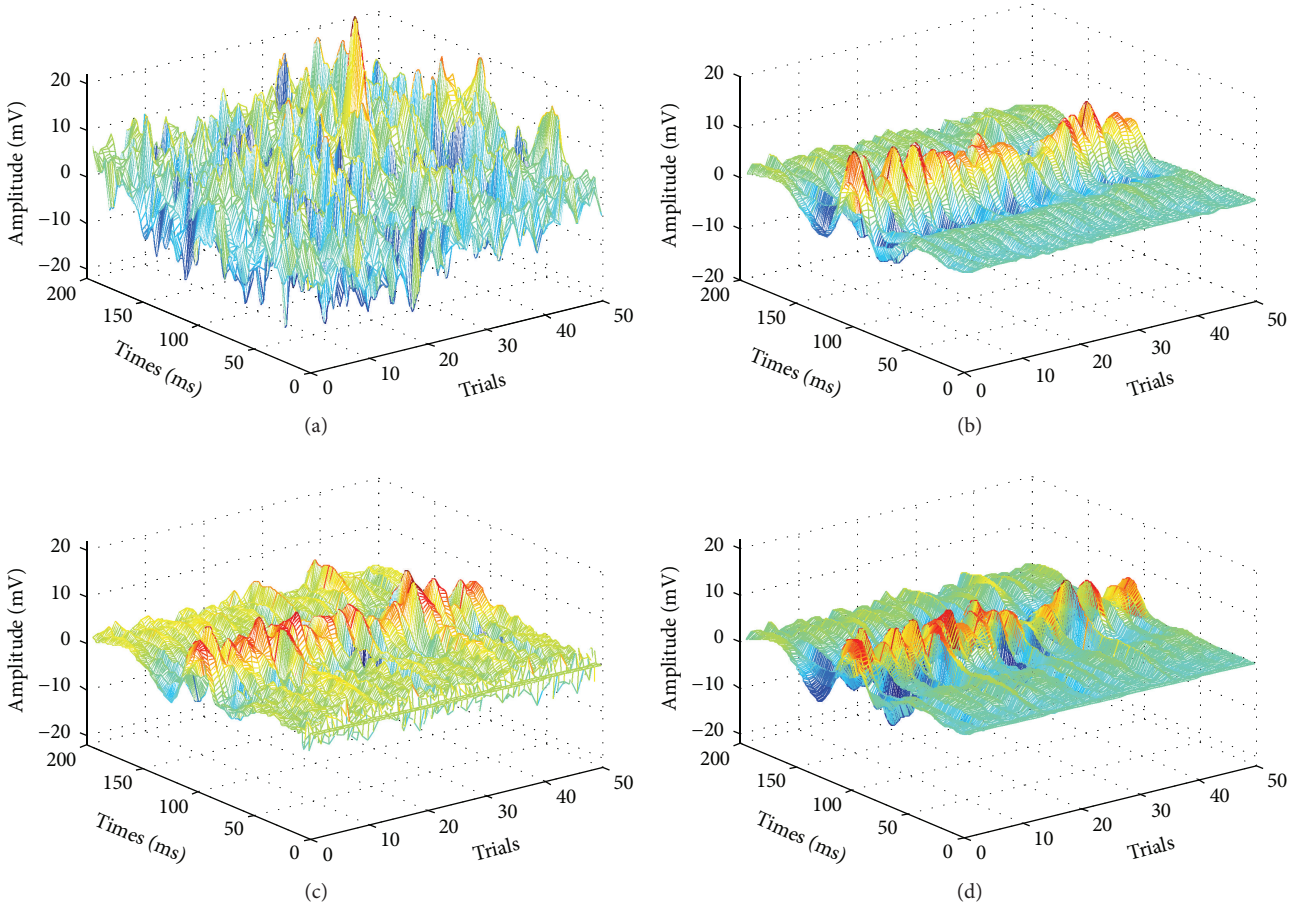


Figure 8: (a) The original 50 VEP signals. (b) The estimated VEP with our method. (c) The estimated VEP with MOSCA. (d) The estimated VEP with ARX.

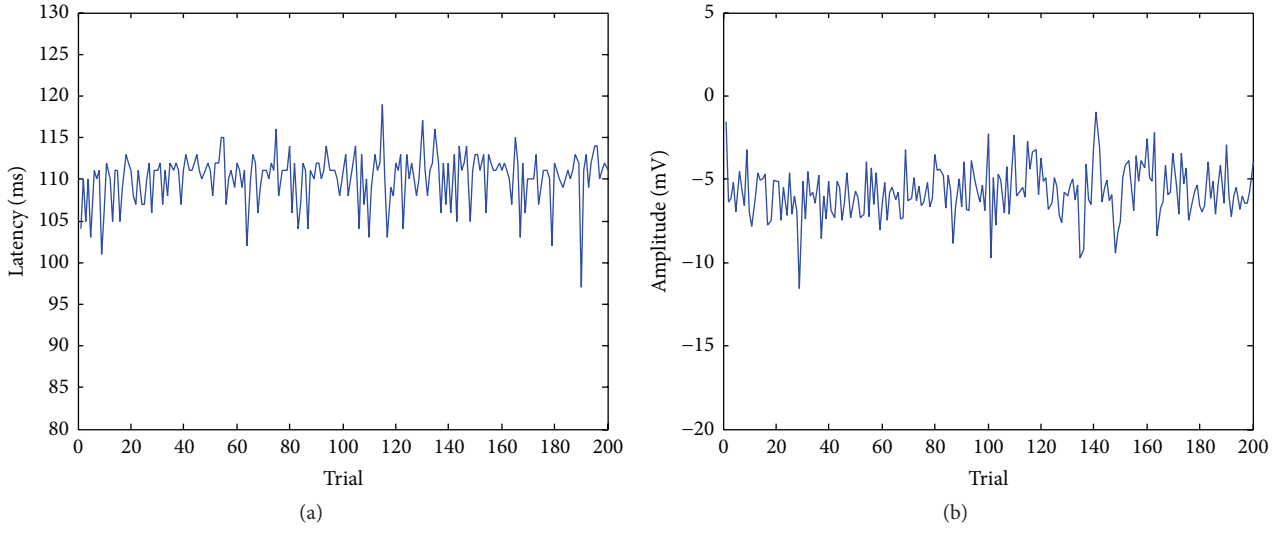


Figure 9: (a) The estimation of amplitudes of P100 of 200 trials. The mean is 110.16 ms and the standard deviation is 3.07 ms. (b) The estimation of latencies of P100 of 200 trials. The mean is −5.73 mv and the standard deviation is 0.54 mv.

Competing Interests

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

Acknowledgments

This work was supported by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant no. 13KJB51-0010), the Natural Science Foundation of Jiangsu province, China (Grant no. BK20130230), and the Natural Science Foundation of China (Grants nos. 61401181 and 61403174).

References

[1] M. H. Costa, "Estimation of the noise autocorrelation function in auditory evoked potential applications," *Biomedical Signal Processing and Control*, vol. 7, no. 5, pp. 542–548, 2012.

[2] C. Reynolds, B. A. Osuagwu, and A. Vuckovic, "Influence of motor imagination on cortical activation during functional electrical stimulation," *Clinical Neurophysiology*, vol. 126, no. 7, pp. 1360–1369, 2015.

[3] J. S. Paul, A. R. Luft, D. F. Hanley, and N. V. Thakor, "Coherence-weighted Wiener filtering of somatosensory evoked potentials," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 12, pp. 1483–1488, 2001.

[4] S. D. Georgiadis, P. O. Ranta-Aho, M. P. Tarvainen, and P. A. Karjalainen, "Single-trial dynamical estimation of event-related potentials: a kalman filter-based approach," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 8, pp. 1397–1406, 2005.

[5] Z. Wang and A. W. Roe, "Trial-to-trial noise cancellation of cortical field potentials in awake macaques by autoregression model with exogenous input (ARX)," *Journal of Neuroscience Methods*, vol. 194, no. 2, pp. 266–273, 2011.

[6] S. Cerutti, G. Baselli, D. Liberati, and G. Pavesi, "Single sweep analysis of visual evoked potentials through a model of parametric identification," *Biological Cybernetics*, vol. 56, no. 2-3, pp. 111–120, 1987.

[7] H. Kumru, D. Soler, J. Vidal, J. M. Tormos, A. Pascual-Leone, and J. Valls-Sole, "Evoked potentials and quantitative thermal testing in spinal cord injury patients with chronic neuropathic pain," *Clinical Neurophysiology*, vol. 123, no. 3, pp. 598–604, 2012.

[8] E. Causevic, R. E. Morley, M. V. Wickerhauser, and A. E. Jacquin, "Fast wavelet estimation of weak biosignals," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 6, pp. 1021–1032, 2005.

[9] S. A. Martazi, S. Qazi, L. S. Stergioulas et al., "Wavelet filtering of the P300 component in event-related potentials," in *Proceedings of the 28th IEEE EMBS Annual International Conference on Engineering in Medicine and Biology Society*, vol. 1, pp. 1719–1722, New York, NY, USA, 2006.

[10] P. Xu and D. Yao, "Development and evaluation of the sparse decomposition method with mixed over-complete dictionary for evoked potential estimation," *Computers in Biology and Medicine*, vol. 37, no. 12, pp. 1731–1740, 2007.

[11] A. C. De Silva, N. C. Sinclair, and D. T. J. Liley, "Limitations in the rapid extraction of evoked potentials using parametric modeling," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1462–1471, 2012.

[12] N. N. Yu, H. K. Liu, X. Y. Wang, and H. Lu, "A joint sparse representation-based method for double-trial evoked potentials estimation," *Computers in Biology and Medicine*, vol. 43, no. 12, pp. 2071–2078, 2013.

[13] J. L. K. Kramer, J. Haefeli, A. Curt, and J. D. Steeves, "Increased baseline temperature improves the acquisition of contact heat evoked potentials after spinal cord injury," *Clinical Neurophysiology*, vol. 123, no. 3, pp. 582–589, 2012.

[14] D. H. Lange, H. Pratt, and G. F. Inbar, "Modeling and estimation of single evoked brain potential components," *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 9, pp. 791–799, 1997.

[15] O. Shental, "Sparse representation of white gaussian noise with application to L0-norm decoding in noisy compressed sensing," http://arxiv.org/abs/1104.2215.

[16] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[17] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers*, vol. 1, pp. 40–44, IEEE, Pacific Grove, Calif, USA, November 1993.

[18] N. Kamel, M. Z. Yusoff, and A. F. M. Hani, "Single-trial subspace-based approach for VEP extraction," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 5, pp. 1383–1393, 2011.