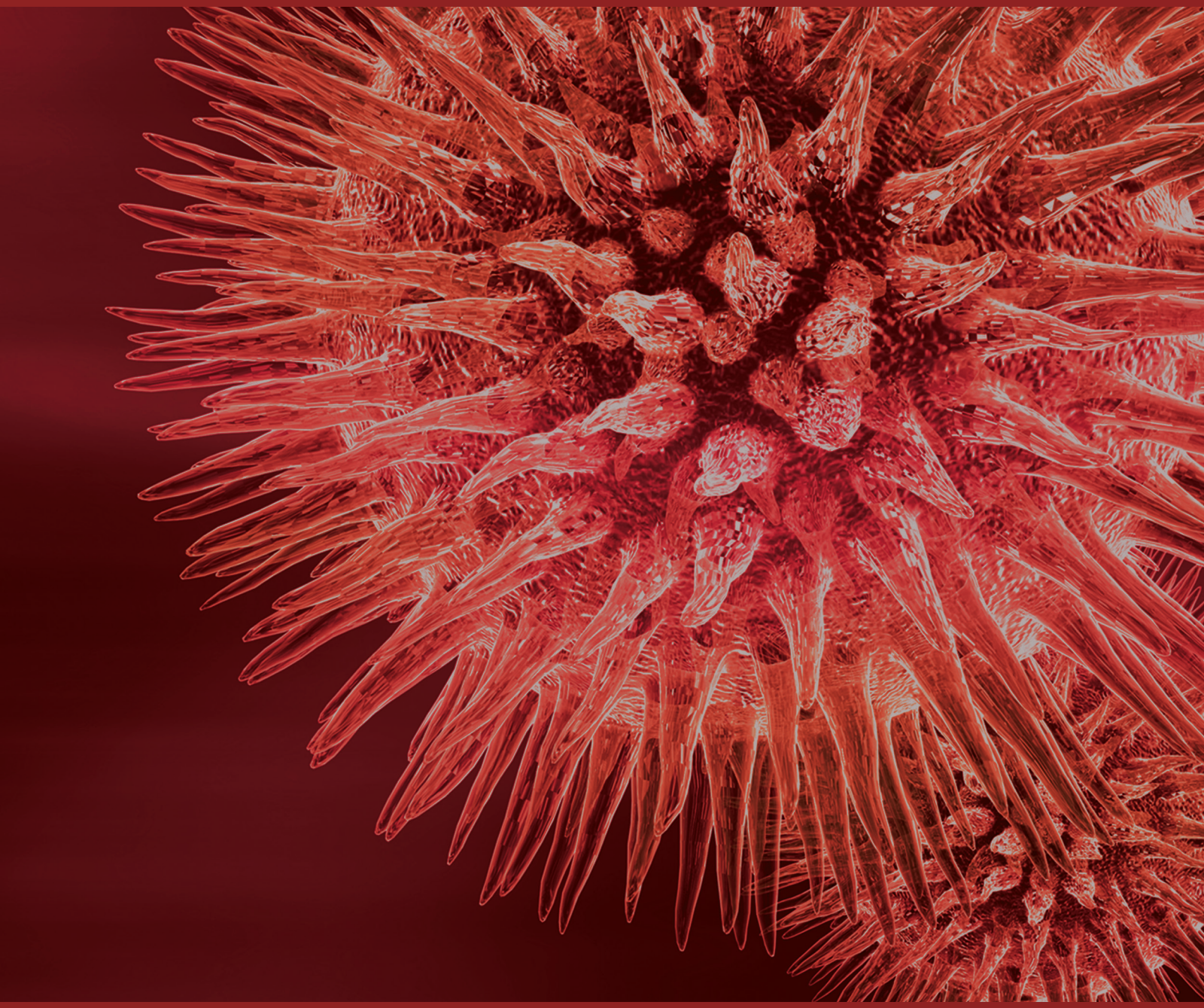
BioMed Research International

Systems Biology Approaches to Mining High Throughput Biological Data

Guest Editors: Fang-Xiang Wu, Min Li, Jishou Ruan, and Feng Luo



Systems Biology Approaches to Mining High Throughput Biological Data

BioMed Research International

Systems Biology Approaches to Mining High Throughput Biological Data

Guest Editors: Fang-Xiang Wu, Min Li, Jishou Ruan, and Feng Luo

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 504362, 2 pages
http://dx.doi.org/10.1155/2015/504362

Editorial

Systems Biology Approaches to Mining High Throughput Biological Data

Fang-Xiang Wu,[1,2] Min Li,[3] Jishou Ruan,[1] and Feng Luo[4]

[1]*School of Mathematical Sciences, Nankai University, Tianjin 300071, China*
[2]*Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK, Canada S7N 5A9*
[3]*School of Information Science and Engineering, Central South University, Changsha 410083, China*
[4]*School of Computing, Clemson University, Clemson, SC 29634, USA*

Correspondence should be addressed to Fang-Xiang Wu; faw341@mail.usask.ca

Received 8 July 2015; Accepted 8 July 2015

With advances in high throughput measurement techniques, large-scale biological data have been and will continuously be produced, for example, gene expression data, protein-protein interaction (PPI) data, tandem mass spectra data, microRNA expression data, lncRNA expression data, and biomolecule-disease association data. Such data contain insightful information for understanding the mechanism of molecular biological systems and have proved useful in diagnosis, treatment, and drug design for genetic disorders or complex diseases. For this focus issue, we have invited the researchers to contribute original research articles which develop or improve systems biology approaches to mining high throughput biological data.

With high throughput data, it is appealing to develop systems biology approaches to understand important biological processes. In the paper "Differential Expression Analysis in RNA-Seq by a Naive Bayes Classifier with Local Normalization," Y. Dou et al. developed a new tool for the identification of differentially expressed genes with RNA-Seq data, named GExposer. This tool introduced a local normalization algorithm to reduce the bias of nonrandomly positioned read depth. The Naive Bayes classifier was employed to integrate fold change, transcript length, and GC-content to identify differentially expressed genes. Results on several independent tests showed that GExposer had better performance than other methods. In the paper "K-Profiles: A Nonlinear Clustering Method for Pattern Detection in High Dimensional Data," K. Wang et al. designed the nonlinear K-profiles clustering method, which can be seen as the nonlinear counterpart of the K-means clustering algorithm. The method had a built-in statistical testing procedure that ensures genes not belonging to any cluster do not impact the estimation of cluster profiles. Results from extensive simulation studies showed that K-profiles clustering outperformed traditional linear K-means algorithm. In addition, K-profile clustering generated biologically meaningful results from a gene expression dataset.

Replicative senescence is of fundamental importance for the process of cellular aging. In the paper "Similarities in Gene Expression Profiles during In Vitro Aging of Primary Human Embryonic Lung and Foreskin Fibroblasts," S. Diekmann et al. elucidated cellular aging process by comparing gene expression changes, measured by RNA-Seq, in fibroblasts originating from two different tissues, embryonic lung (MRC-5) and foreskin (HFF), at five different time points during their transition into senescence. Their results showed that a number of monotonically up- and downregulated genes had a novel strong functional link to aging and senescence related processes.

More and more studies have shown that many complex diseases are contributed jointly by alterations of numerous genes. Genes often coordinate together as a functional biological pathway or network and are highly correlated. In the paper "Module Based Differential Coexpression Analysis Method for Type 2 Diabetes," L. Yuan et al. proposed a gene differential coexpression analysis algorithm and applied it to

a publicly available type 2 diabetes (T2D) expression dataset. Two differential coexpression gene modules about T2D were detected and were expected to be useful for exploring the biological functions of the related genes.

Oral mucosa is a useful material for regeneration therapy with the advantages of its accessibility and versatility regardless of age and gender. In the paper "Gene Signature of Human Oral Mucosa Fibroblasts: Comparison with Dermal Fibroblasts and Induced Pluripotent Stem Cells," K. Miyoshi et al. reported the comparative profiles of the gene signatures of human oral mucosa fibroblasts (hOFs), human dermal fibroblasts (hDFs), and hOF-derived induced pluripotent stem cells (hOF-iPSCs), linking these with biological roles by functional annotation and pathway analyses. Their findings demonstrated that hOFs had unique cellular characteristics in specificity and plasticity. These data may provide useful insight into application of oral fibroblasts for direct reprograming.

Predicting disease genes for a particular genetic disease is very challenging. However, this challenge can be tackled via exploring high throughput data. In the paper "ProSim: A Method for Prioritizing Disease Genes Based on Protein Proximity and Disease Similarity," G. U. Ganegoda et al. proposed a new algorithm called proximity disease similarity algorithm (ProSim), which took use of two types of data: disease similarity data and protein-protein interaction data. The computational results have shown that their proposed method outperformed existing methods.

In order to learn the protein structures and functions via computational methods, it is important to predict the solvent accessibility and the contact number of protein residues from protein sequence. In the paper "AcconPred: Predicting Solvent Accessibility and Contact Number Simultaneously by a Multitask Learning Framework under the Conditional Neural Fields Model," J. Ma and S. Wang presented a method AcconPred for predicting solvent accessibility and contact number simultaneously, which was based on a shared weight multitask learning framework under the CNF (Conditional Neural Fields) model. Trained on a 5729 monomeric soluble globular protein dataset, AcconPred could reach 0.68 three-state accuracy for solvent accessibility and 0.75 correlation for the contact number. Tested on the 105 CASP11 domain dataset for solvent accessibility, AcconPred could reach 0.64 accuracy, which outperformed existing methods.

The Smith-Waterman algorithm is one of the key sequence search algorithms for sequence alignment and has gained popularity due to improved implementations and rapidly increasing compute power. Recently, the Smith-Waterman algorithm has been successfully mapped onto the emerging general-purpose graphics processing units (GPUs). In the paper "Improving the Mapping of Smith-Waterman Sequence Database Searches onto CUDA-Enabled GPUs," L.-T. Huang et al. employed the CUDA-enabled GPU to improve the mapping of Smith-Waterman algorithm, especially for short query sequences. The computational results showed that the proposed method significantly improved Smith-Waterman algorithm on CUDA-enabled GPUs in proper allocation of block and thread numbers.

Protein interaction article classification is a text classification task in the biological domain to determine which articles describe protein-protein interactions. In the paper "Improving Classification of Protein Interaction Articles Using Context Similarity-Based Feature Selection," Y. Chen et al. proposed new context similarity-based feature selection methods. Their performances were evaluated on two protein interaction article collections and compared against the frequency-based methods. The experimental results revealed that the context similarity-based methods performed better in terms of the F1 measure and the dimension reduction rate.

Recent studies suggest that posttranscriptional RNA modifications play a crucial role in regulating gene expression. In practice, a single methylation site can contain multiple RNA methylation residuals, some of which can be regulated by different enzymes and thus differentially methylated between two conditions. However, existing peak-based methods could not effectively differentiate multiple methylation residuals located within a single methylation site. In the paper "Spatially Enhanced Differential RNA Methylation Analysis from Affinity-Based Sequencing Data with Hidden Markov Model," Y.-C. Zhang et al. proposed a hidden Markov model (HMM) based approach to address this issue. The proposed algorithms were tested on both simulated data and real data. Results suggested that their proposed algorithm clearly outperformed existing peak-based approach on simulated systems and could detect differential methylation regions with higher statistical significance on real data, indicating an improved performance.

Pregnane X Receptor (PXR) and drug-metabolizing target genes are involved in most of inductive herb-drug interactions. To predict this kind of herb-drug interactions, the protocol could be simplified to only screen agonists of PXR from herbs because the relations of drugs with their metabolizing enzymes are well studied. In the paper "Screening Ingredients from Herbs against Pregnane X Receptor in the Study of Inductive Herb-Drug Interactions: Combining Pharmacophore and Docking-Based Rank Aggregation," Z. Cui et al. employed a combinational in silico strategy of pharmacophore modelling and docking-based rank aggregation (DRA) to identify PXR's agonists. To validate their method, a curated herb-drug interaction database was built, which recorded 380 herb-drug interactions. The results showed that, among the top 10 herb ingredients from the ranking list, 6 ingredients were reported to involve herb-drug interactions.

In summary, this focus issue has reported the recent progress in systems biology approaches to analyzing high throughput data such as gene expression data, various biomolecular interaction data, and sequencing data. We hope that the readers of this focus issue could get some benefits from these newly developed methods.

Fang-Xiang Wu
Min Li
Jishou Ruan
Feng Luo

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 213750, 11 pages
http://dx.doi.org/10.1155/2015/213750

Research Article

ProSim: A Method for Prioritizing Disease Genes Based on Protein Proximity and Disease Similarity

Gamage Upeksha Ganegoda, Yu Sheng, and Jianxin Wang

School of Information Science and Engineering, Central South University, Changsha 410083, China

Correspondence should be addressed to Yu Sheng; shengyu@csu.edu.cn

Received 15 December 2014; Accepted 16 January 2015

Academic Editor: Fang-Xiang Wu

Predicting disease genes for a particular genetic disease is very challenging in bioinformatics. Based on current research studies, this challenge can be tackled via network-based approaches. Furthermore, it has been highlighted that it is necessary to consider disease similarity along with the protein's proximity to disease genes in a protein-protein interaction (PPI) network in order to improve the accuracy of disease gene prioritization. In this study we propose a new algorithm called proximity disease similarity algorithm (ProSim), which takes both of the aforementioned properties into consideration, to prioritize disease genes. To illustrate the proposed algorithm, we have conducted six case studies, namely, prostate cancer, Alzheimer's disease, diabetes mellitus type 2, breast cancer, colorectal cancer, and lung cancer. We employed leave-one-out cross validation, mean enrichment, tenfold cross validation, and ROC curves to evaluate our proposed method and other existing methods. The results show that our proposed method outperforms existing methods such as PRINCE, RWR, and DADA.

1. Introduction

Disease gene prioritization aims to suggest potential implications of genes in disease susceptibility. Also, it is important to know genes that are related to a particular disease in order to treat it. Hence identifying the genes related to a specific disease is one of the major challenges in the field of bioinformatics. To do so it is vital to consider biological details such as biological functions, patterns of expression in different conditions, and interactions with other genes. Furthermore, it is important to know functional annotations of candidate genes to a disease or phenotype under investigation as there are close relationships between the biological aspects and related diseases.

In medical research it is necessary to understand the genetic background of diseases with major implications in order to diagnose, treat, and develop drug for these diseases. Linkage analysis and association studies are some of the traditional gene-mapping approaches that have demonstrated remarkable success in this field [1]. Family-based linkage analysis is able to correlate diseases with specific genomic regions. Experimental examination of causative mutations in genomic regions is expensive and laborious as it consists of hundreds of genes. Thus to handle this challenge, traditional approaches are more time-consuming and costly while computational approaches are often considered to offer more efficient and effective alternatives. Therefore computational approaches have been developed to prioritize candidate genes for a particular disease. Different approaches have used various data sources such as gene expression [2, 3], sequence similarity of genes, DNA methylation [4], tissue-specific information [5], functional similarity and annotations [2, 6], and protein-protein interactions (PPIs) [7, 8] in determining the strength of association between genes and diseases as well as associations between diseases and protein complexes [9]. Network-based prioritization methods [10] are based on the observation that genes related to similar diseases tend to lie close to one another in the PPI network [11].

Furthermore, some other researchers have considered phenotype similarity in terms of gene closeness to prioritize disease genes. Depending on these studies, the correlation between phenotype similarity and gene closeness, defined

by a concordance score, is a strong and robust predictor of disease genes [12]. Meanwhile, some researchers used tissue-specific gene expression data along with PPI networks to prioritize disease genes as many disorders are involved a disruption of the "molecular fabric" of different, healthy tissues [12]. In addition, some used support vector machine recursive feature elimination (SVM-RFE) method for gene selection in different cancer tissues by incorporating a minimum-redundancy maximum-relevancy (MRMR) filter [13]. In current studies, phenotypic similarity between the diseases of interest to other diseases for which causal genes are known has been used to prioritize candidate genes [3, 14]. Simultaneously, some researchers grouped diseases into separate disease families to facilitate the prioritization task [15]. Topological properties of PPI networks are also used to understand genetic diseases [3, 8, 16] and essential proteins [17, 18]. Gonçalves et al. combined full topology scores which were computed by using local clustering on graphs or diffusion kernels over confidence weighted gene association networks by integrating evidence from heterogeneous sources, in order to prioritize disease genes [19]. Furthermore, special local clustering has been used to identify genes associated with Alzheimer's disease. With the use of special local clustering algorithm it is able to group genes together with similar expression patterns and identify significantly varied gene expression values as isolated points [20].

This study proposes a new algorithm called proximity disease similarity algorithm (ProSim) which combines protein proximity in PPI networks and disease similarity into a single mathematical formula used to prioritize candidate genes. According to the previous work, protein proximity shows that genes close to the true disease genes tend to be disease genes as well in PPI networks. Meanwhile, disease similarity provides details of how query disease is related to other diseases with regard to phenotypic characteristics. This is expected to increase the power of prioritizing candidate genes to a relevant query disease. The proposed algorithm is evaluated on six case studies and its performance is compared with other existing methods. The results show that the proposed method is superior to existing methods.

2. Materials and Method

2.1. Materials. PPI networks: positive PPI network is downloaded from the Human Protein Reference Database (HPRD). It consists of 9673 nodes with 39240 edges. Negative PPI network is downloaded from the Negatome database, which consists of 1828 nodes with 2171 edges.

Tissue-specific gene expression data are downloaded from Gene Expression Omnibus (GEO) in the National Center for Biotechnology Information (NCBI) website (GEO accession number GSE 7307).

2.2. Methods. The proposed method consists of five main steps described in detail in the following five subsections. The first subsection describes the feature extraction process in which three main features are used to evaluate the effectiveness of the PPIs in positive and negative PPI databases.

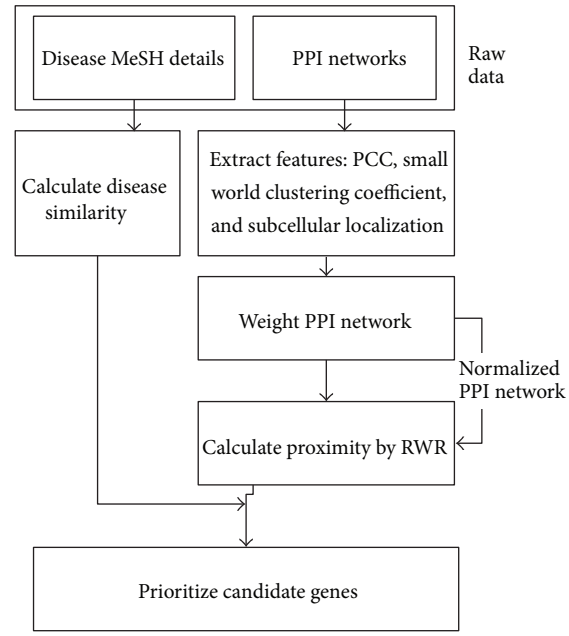


Figure 1: Flow chart of ProSim method.

Next a logistic regression function in these three features is described. The third subsection explains how random walk with restart method is used to calculate the topological similarity of the proteins in the PPI network. The fourth subsection describes how disease similarity is calculated. The fifth subsection explains how all these details are then used to prioritize candidate proteins [21]. The entire process is illustrated as a flow chart in Figure 1.

2.2.1. Feature Extraction. As mentioned in Materials, two kinds of PPI networks are employed in this study: positive and negative PPI networks from which three types of features have been extracted as described in the sequel.

(1) Small World Clustering Coefficient. Small world networks have high clustering coefficients. The cliquishness of the neighborhood around an individual edge should therefore be an indication of how well this edge fits the pattern of a small world network. In order to calculate the clustering coefficient of proteins v and u, the following equation was used:

$$C_{vu} = -\log \sum_{i=|N(v)\cap N(u)|}^{\min\{|N(v)|,|N(u)|\}} \frac{\binom{|N(v)|}{i}\binom{N-|N(v)|}{|N(u)-i|}}{\binom{N}{|N(u)|}}, \quad (1)$$

where C_{vu} denotes the small world clustering coefficient, $N(v)$ and $N(u)$ denote the sets of proteins that directly interact with proteins v and u, respectively, and N is the total number of proteins in the network. By using the above formula, a small world clustering coefficient is calculated for each pair of proteins in both positive and negative PPI networks.

(2) Pearson Correlation Coefficient. Pearson correlation coefficient (PCC) is used to calculate coexpression measurements

for corresponding genes derived from multiple sets of tissue-specific microarray experiments [22]. PCC values are used since coregulated genes are more likely to interact with each other compared with other genes [23, 24]. Gene PCCs are mapped to corresponding proteins. In this study, the correlation coefficient quantifies the similarity of expression between two genes and shows whether the corresponding proteins interact or not [11, 25]. Let v and u be two m-dimensional vectors representing expression profiles of two genes which correspond to proteins v and u, respectively. $\overline{v}$ and $\overline{u}$ are the mean of v and u, while σ_v and σ_u are the standard deviations of v and u, respectively. PCC is calculated for both positive and negative PPI networks by using

$$\rho_{vu} = \frac{(1/m)\sum_{i=1}^{m} v_i u_i - \overline{vu}}{\sigma_v \sigma_u}. \quad (2)$$

(3) Protein Subcellular Localization. The last feature is a protein subcellular localization data of interacting partners. Protein subcellular localization data of interacting partners is represented by either zero or one [26]. To identify the subcellular location of each protein Hum-PLoc technique was used [27]. Hum-PLoc is a predictor developed by Shen and Chou that is able to deal with the multiplex problems of the human protein system. As such, the coverage scope for human proteins is extended from 4 to 12 location sites. The subcellular location sites are cytoplasm, mitochondria, nucleus, plasma membrane, centriole, cytoskeleton, endoplasmic reticulum, extracell, Golgi apparatus, lysosome, microsome, and peroxisome. In order to get the subcellular location, all the protein sequences should be obtained first. To get the sequences of proteins in positive and negative PPI networks proteins are mapped from the Swiss-Prot database. Thereafter, protein sequences are used with Hum-PLoc predictor to identify the subcellular location of that particular protein. Huh et al. [28] note that not every subcellular location has a biological significance with others. They publish a list showing which subcellular locations have a biological significance with each other. Based on the list, every pair of proteins in PPI networks is assigned a value of 1 if listed and 0 otherwise. The value is purely based on the fact of biological significance of each subcellular location without considering whether protein interaction actually exists or not.

All these three features are then used in a logistic regression function to calculate a reliability score for each protein in the network [26].

2.2.2. Logistic Regression Function. A logistic regression function is employed to calculate the weight of each interaction in the PPI network. To train the logistic regression model we have used HPRD as the golden standard positive PPI network and Negatome database as the negative PPI network. According to the logistic distribution, the probability of true interaction T_{vu} given the three input features $X = (X_1, X_2, X_3)$ is calculated based on the formula as follows:

$$\Pr(T_{vu} \mid X) = \frac{1}{1 + \exp\left(-\beta_0 - \sum_{i=1}^{3} \beta_i X_i\right)}. \quad (3)$$

According to the general logistic regression function it is better if the positive and negative datasets are balanced in order to obtain the best values for regression constants. For this purpose we have considered only a portion of the positive PPI network whose size will be equal to that of the negative PPI network. For the simplicity we have selected first 2000 interactions from the HPRD and almost the same amount of interactions from the negative PPI network as well.

2.2.3. Random Walk with Restart to Calculate the Proximity. For a given disease q, in order to calculate the proximity between proteins, two sets of genes, seed set S and candidate set C, are used. A seed set S consists of genes known to be associated with the query disease q. A candidate set C specifies one or more genes, "potentially" associated with the disease q, created by excluding the seed genes from the PPI network. The rest of genes related to the given PPI are then considered as the candidate set.

Random walk with restart method is used to compute the proximity of candidate genes to relevant seed genes. Random walk with restart method is actually a generalization of Google's well-known page-rank algorithm [29]. After getting the proximity values for each protein in the network, the values are then sorted in a descending order. These proximity values are used at the final stage of the process, to prioritize the candidate proteins.

2.2.4. Calculation of Disease Similarity. Disease similarity is calculated based on a metric proposed by van Driel et al. [14], who used the medical subject headings vocabulary (MeSH) to extract terms from Online Mendelian Inheritance in Man (OMIM) to identify similar diseases. Each MeSH entry is a collection of terms with synonyms and plurals, called a concept. MeSH provides a standardized way to retrieve information that uses different terminologies to refer to the same concepts. van Driel et al. tested the prediction power of different ranges of similarity values by calculating the correlation between the similarity of two diseases and the functional relatedness of their causal genes. According to their research findings, similarity values in the range $[0, 0.3]$ were not informative while genes with similarities in the range $[0.6, 1]$ showed significant functional similarity. A logistic function L shown in (4) is used to calculate a probability that two diseases are related:

$$L(x) = \frac{1}{1 + e^{(cx+d)}}. \quad (4)$$

The values for the parameters c and d were set as $c = -15$ and $d = \log(9999)$. L was then used to compute the prior knowledge to a particular disease, denoted by Y, as shown in

$$Y(q) = L(S(p, q)), \quad (5)$$

where q is the query disease and $S(q, p)$ is the similarity between diseases q and p. This equation is slightly different from the way of prioritization and complex elucidation algorithm (PRINCE) [30] to calculate the disease similarity. Here disease similarity is calculated as a global representation.

Table 1: Values of β for each folder.

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
1058.461	1059.242	1058.926	1033.53	8125.654	1058.279	1058.381	1058.601	1069.272	1029.123
−0.04366	−0.03784	0.00557	1.690454	367.802	−0.02873	−0.02004	0.005051	−0.47721	−0.71004
0.875513	0.878244	0.877808	0.748819	58.37504	0.876113	0.877078	0.878938	0.873331	0.669624
−1166.52	−1167.42	−1167.09	−1140.27	−9689.94	−1166.34	−1166.48	−1166.75	−1178.31	−1132.55

The original PRINCE algorithm showed how each protein in the PPI network was associated with different diseases. Thus if disease p was more similar to disease q, disease p was selected.

2.2.5. Final Stage of the Process. At the final stage a prioritization score was calculated. In the proposed approach PRINCE algorithm was modified by incorporating both proximity and the disease similarity values as shown in

$$F(v) = \alpha \left[\sum_{u \in N(v)} F(u)\, w'(v,u) \right] + (1-\alpha)\left(Y(q) + \text{Pro}(v)\right), \tag{6}$$

where $F(v)$ reflects the relevance of protein v to disease q. The prioritization function consists of three inputs: given a PPI network $G = (V, E, w)$, where V represents the set of proteins, E is the set of interactions, and w denotes the weight of each interaction, a normalized form of adjacency matrix w, denoted by w', is derived and used as one of the inputs. The adjacency matrix is constructed by using the reliability score obtained from the logistic regression function discussed in Section 2.2.2. The matrix has been normalized with the weight of an edge by the degree of its end-points. The second input gives the prior information of disease similarity in relation to the query disease q. The calculation of disease similarity is explained in Section 2.2.4. The third input value is the proximity of proteins related to the seed genes. Proximity is calculated by using the random walk with restart method. In the proposed equation, w' is a $|V| \times |V|$ matrix, while F, Y, and Pro are displayed as vectors of size $|V|$. To improve the execution speed and ensure a convergence of (6), a propagation based approach similar to work reported by Zhou et al. [31] is applied. The resultant equation (7) is thus guaranteed to converge after *enough* iteration:

$$F^t = \alpha W' F^{t-1} + (1-\alpha)\left(Y(q) + \text{Pro}(v)\right). \tag{7}$$

From (7), it can be seen that if a node has prior information, it will propagate the information to its neighbors. The process continues until the value converges or the maximum iteration value, T, is reached. In this study T is set to 100 and values of α varied in $(0, 1)$.

3. Results and Discussion

This section details experiment results for the proposed method and also provides comparisons with existing methods. In particular, the proposed method has been evaluated via six diseases: breast cancer (MIM: 114480), colorectal cancer (MIM: 114500), lung cancer (MIM: 211980), prostate cancer (MIM: 176807), Alzheimer's disease (MIM: 104300), and diabetes mellitus type 2 (MIM: 125853). Performance comparison was done against the original PRINCE algorithm [30], random walk with restart (RWR) [24], and degree-aware disease algorithm (DADA) [29].

As indicated in Section 2.2.1, the first part of the experiment was to extract features of concern. Based on (1), the small world clustering coefficient of each protein interaction was calculated. Experiment results showed that most of the coefficients lay between −0.6 and −0.7. These high clustering coefficients of small world networks indicated that neighbors of a given vertex are more likely to have edges between them than expected. Next, PCC was calculated for each PPI. From the results PCC value was greater than 0.5 for most of the potential candidate genes selected for a specific disease while negative values for the same indicated that those proteins were less significant to a given query disease. Lastly, a protein subcellular localization data of interacting partners was determined, in order to find the biological significance of the subcellular localization. Results indicated that a large amount of proteins was biologically significant in a given PPI network.

These three features were then used as inputs of a logistic regression function (3). Based on a tenfold cross validation method, ten sets of values of β were obtained by the logistic regression function, shown as fold i, $i = 1, 2, \ldots, 10$, in Table 1. Under each fold, 4 values corresponding to β_0 to β_4 are shown. The average value of β was then taken as the value used to calculate the probability of true interaction T_{vu} given the three input features $X = (X_1, X_2, X_3)$, which are 1760.946913, 36.81855608, 6.593051189, and −2014.167331 for β_0 to β_4, respectively.

Proximity values show proteins that are closer to a specific disease. Therefore if a proximity value is high, it indicates that the given protein has a high chance of being related to a given disease. Random walk with restartwas used to give proximity values. For example, Figure 2 shows proximity values of different proteins for prostate cancer disease. It illustrates around 25% of proteins in the network have high proximity value. As a result these proteins have a high prominent factor to be a disease related gene for a particular disease.

Disease similarity measures how similar a query disease was to other diseases. If the disease similarity value was less than 0.3 it meant that the diseases in question were not significantly relevant. When the value was higher than 0.6, a high connection existed between the given diseases.

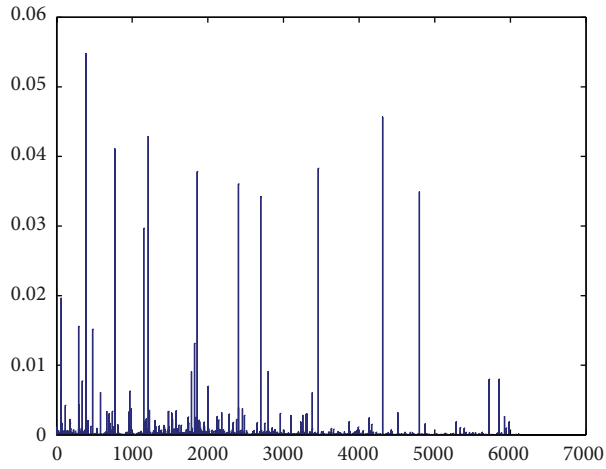


Figure 2: Proximity values for different proteins to prostate cancer.

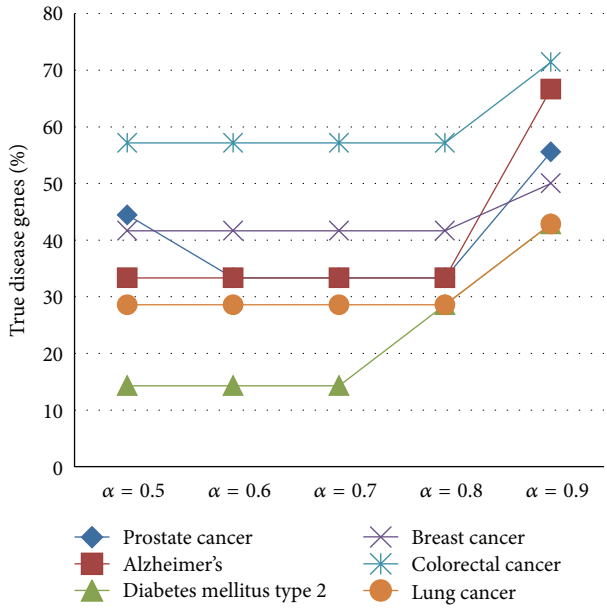


Figure 3: Alpha value prediction.

The final part of the experiment combined the proximity values with disease similarity values to calculate a prioritization score. Testing was carried out based on the prediction capacity of disease genes of the final equation in order to find out the best combination of proximity values with disease similarity. According to the results, by including the proximity value as an explicit value, the prediction capability will be increased. Furthermore, to tune the value of α used in (6), several values were experimented with for all the six diseases and the results are depicted in Figure 3. From Figure 3 the best results were obtained when $\alpha = 0.9$.

The proposed method is compared with three other methods, namely, PRINCE, RWR, and DADA. By comparing the top ten genes ranked by each method one can conclude that the proposed method is able to predict more known and unknown disease genes than existing methods. Table 2 shows the top ten genes predicted for each method and for each case study. Figure 4 shows another representation of how much each method is able to predict known and unknown disease genes within the top ten disease genes.

To provide more performance comparisons of the proposed method to existing methods, a leave-one-out cross validation procedure was used. With each cross validation trial, a single seed gene related to the query disease was removed and then each method is evaluated on its success of identifying and ranking the removed seed gene. To replicate the case of prioritizing the proteins encoded by genes inside a linkage interval, an interval of size 100 was used similarly to the work by Köhler et al. [15]. But in this research, it will calculate the percentage of true disease genes identified within the top 100 genes. Hence the threshold value used to rank the candidate genes was set to 100.

As shown in Table 3 the proposed algorithm performed better than the other methods. The proposed algorithm identified true disease genes at 80%, 71%, 69%, 66%, 57%, and 50% for breast cancer, prostate cancer, Alzheimer's disease, colorectal cancer, diabetes mellitus, and lung cancer, respectively, significantly higher than all the rest; however, it was comparable to other methods on diabetes. Because the gene expression details used for the calculation of PCC are not affect to cause diabetes disease. Therefore it has given a negative impact on the final result. Hence this has given a direction to improve in the future work.

Further performance evaluation of the algorithm was based on sensitivity and specificity measures and used to draw ROC curves shown in Figure 5 for breast cancer, Alzheimer's disease, colorectal cancer, diabetes mellitus type 2, lung cancer, and prostate cancer, respectively. Sensitivity is defined as the percentage of true disease genes that are ranked above a specified threshold while specificity is defined as percentage of all nonrelated disease genes that are ranked below a specified threshold.

Lastly, a mean enrichment value [15] was calculated for each method and used for performance comparison purposes. In general, the mean enrichment formula is enrichment = 50/(rank). Based on ranking values, by using the leave-one-out cross validation process, it was possible to identify the rank of true disease genes for each method. The results are shown in Table 4. From the results, the proposed algorithm performed better than the other algorithms.

Tenfold cross validation is carried out to illustrate the performance of the proposed algorithm with the combination of positive and negative PPI networks as well as with positive PPI network. At each cross validation trial onefold PPIs were removed from the total PPI network and the rest of PPIs were used for the prioritization process. By calculation, onefold is one-tenth of total PPIs in the total network. Figure 6 shows the results of the testing. These results show that the proposed algorithm is effective in identifying the correct disease related genes from all the positive and negative PPIs. Comparatively, the proposed method is able to predict disease genes with the positive PPI network, as well as the total positive and negative PPI network in an effective way.

The original PRINCE algorithm does not include subcellular localization data as a feature in calculating a reliability of

Table 2: Top ten genes predicted for ProSim, PRINCE, RWR, and DADA.

Breast cancer				Colorectal cancer			
ProSim	PRINCE	RWR	DADA	ProSim	PRINCE	RWR	DADA
ATM	NBN	PIK3R1	PIK3R1	TP53	BRCA1	TP53	TP53
PPP1R13L	TIE1	HTT	FNTA	JUN	MSH2	AKT1	AURKA
FOSL2	BRCA1	CD44	ATM	PIK3R1	MLH1	AKT	RHOA
ERBB4	MSH2	CALM1	HTT	HTT	EP300	IL3	RET
BRIP1	PATZ1	JUN	SHOC2	NRAS	AKT1	NRAS	NTRK1
HIP1	DGCR2	MAPK3	RASSF2	PSEN1	BUB1B	BRCA2	CCND1
UBE2K	ATM	TPX2	CD44	CTNNB1	OGG1	BUB1B	BUB1B
RALGDS	MRE11A	SVIL	CALM1	CBL	RNF139	LYN	BRCA2
FOXG1	H2AFX	RAF1	JUN	AKT1	APC	AURKA	EP300
PLK1	TERF2	PSEN1	RAF1	VAV1	EXO1	HNF1A	GABRB1
Lung cancer				Diabetes mellitus			
ProSim	PRINCE	RWR	DADA	ProSim	PRINCE	RWR	DADA
PPP1R13L	TP53	PPP1R13L	IMPDH2	TP53	RHOA	AKT1	AKT1
ERBB4	VRK1	MAP3K8	APLP2	JUN	MAFA	PLN	INS
PLK1	TP53RK	IGF1R	PLK1	PIK3R1	BSCL2	PSEN1	PSEN1
BRD7	ERBB4	BRD7	ICMT	HTT	SLC2A2	LYN	JUN
UBE2K	CDKN2A	UBE2K	GSTM4	HNF1A	HNF4A	HNF1A	HNF1A
EGFR	TP53INP1	HUWE1	RAD17	PSEN1	INS	PCBD1	TP53
TAPBP	TAG	TAPBP	KLF4	LYN	IAPP	CASP8	ALB
UHMK1	EGFR	UHMK1	CASP8	CBL	NEUROD1	PIK3R1	NEUROD1
MAP3K8	RRM2	DGCR2	PIAS1	AKT1	GATA5	HTT	CEBPA
HIP1	CUL9	HIP1	FOSL2	VAV1	MAP3K13	EP300	RAC3
Prostate cancer				Alzheimer's disease			
ProSim	PRINCE	RWR	DADA	ProSim	PRINCE	RWR	DADA
TP53	NBN	TP53	TP53	TP53	HSD17B10	AKT1	AKT1
JUN	BRIP1	RNASEL	RNASEL	JUN	BACE2	PAX2	SFRP2
HTT	JUN	PAX2	AR	HTT	TP53	CASP8	PAX2
CD44	BRCA1	ERBB2	ABCE1	APBB2	JUN	PSEN1	CASP8
BARD1	HTT	CD19	AKT1	PSEN1	HADHB	APBB3	WT1
CD82	RAD50	BACE2	ERBB2	LYN	NAE1	BLMH	PSEN1
ERBB2	TP53	CASP8	CASP8	VAV1	APLP2	MAPK8	RAC3
REL	MRE11A	TSG101	BUB1B	CCND1	BLMH	WT1	RB1
SLPI	ATM	ATM	CCNE1	CASP8	APBB1	SFRP2	MAPK8
JUNB	BRCA2	CCND1	STAT5A	KIT	F12	RAC3	RHOA

Table 3: Fraction of true disease genes.

	ProSim	PRINCE	RWR	DADA
Breast cancer	**80%**	77%	72%	75%
Alzheimer's	**69%**	58%	55%	53%
Colorectal cancer	**66%**	63%	61%	62%
Diabetes mellitus type 2	57%	58%	53%	**69%**
Lung cancer	**50%**	48%	42%	45%
Prostate cancer	**71%**	61%	53%	60%

Table 4: Mean enrichment.

	ProSim	PRINCE	DADA	RWR
Breast cancer	**0.5565**	0.4896	0.393	0.2852
Alzheimer's	**0.1634**	0.1552	0.087	0.151
Colorectal cancer	**6.3233**	0.5926	3.6511	2.8521
Diabetes mellitus type 2	0.1998	0.246	**0.2414**	0.1488
Lung cancer	**0.2544**	0.1757	0.1625	0.1531
Prostate cancer	**6.2578**	3.0871	5.8649	2.386

the PPIs which could have impacted negatively on PRINCE in prioritizing of the candidate genes from the results of ProSim. Unlike the original PRINCE algorithm which included disease similarity unique to a specific PPI network, ProSim gives a global representation on how diseases relate to each other. In ProSim disease similarity is included as to how query disease relates to other disease. Hence, it will give more focus on the query disease than the original PRINCE algorithm, in

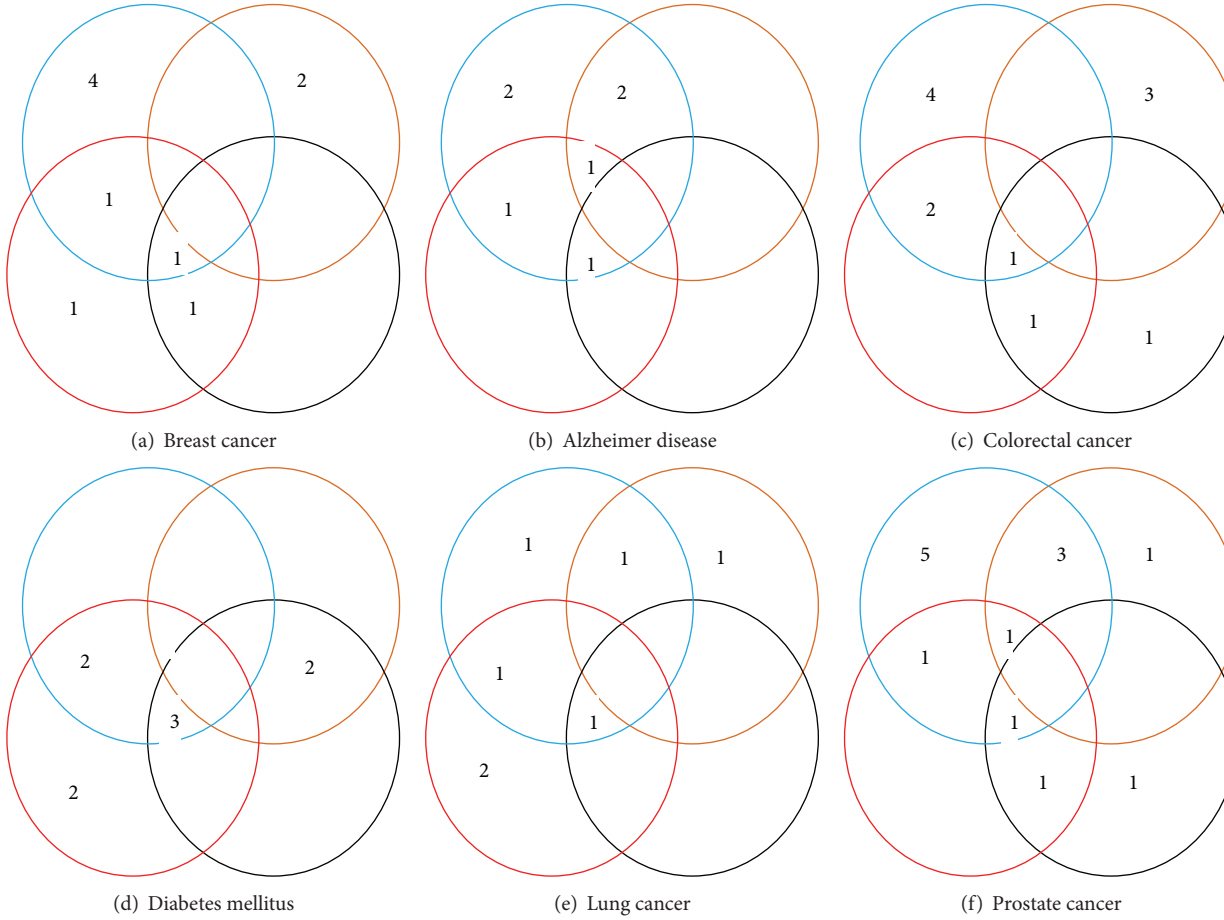


FIGURE 4: Comparison of top ten genes as a Venn diagram. Blue, orange, red, and black circles represent ProSim, PRINCE, RWR, and DADA methods, respectively. (a) Breast cancer. (b) Alzheimer disease. (c) Colorectal cancer. (d) Diabetes mellitus. (e) Lung cancer. (f) Prostate cancer.

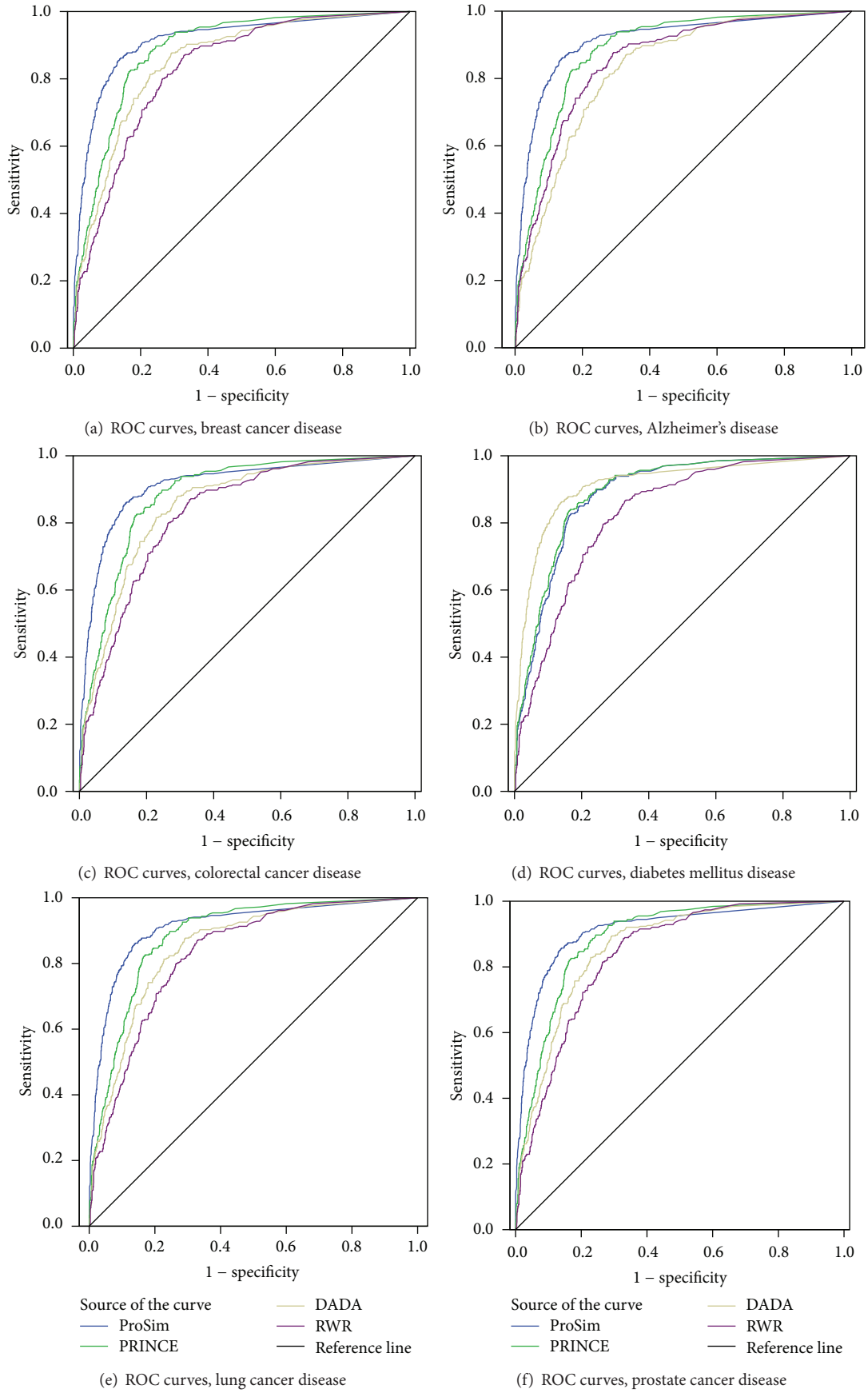
which disease similarity is included for each protein in the PPI network. On the other hand random walk with restart method does not include disease similarity, which could have had a negative influence on their final result on detecting effective candidate genes for a specific disease.

Besides using different techniques to evaluate the performance of ProSim, it was important to identify the relevance of high ranked candidate genes to a given query disease. Thus, further evidence was sought from other online databases and scientific publications. By cross-checking predicted genes with other online databases and scientific publications it was found that HTT, SLPI, JUN, REL, and CD44 genes are tumor suppressors involved in several types of cancer, which were not predicted by the original PRINCE algorithm [32–36] yet identified by ProSim. ProSim also identified JUNB, MDM2 genes, which are used for therapy in prostate cancer [37, 38]. For Alzheimer's disease PSEN1 and JNK genes [39, 40] were ranked higher by the proposed algorithm. Finally, for diabetes mellitus type 2 disease, the proposed method ranked PIK3R1 and JUN genes [41] high. One important finding was that TP53 gene [35, 42] was ranked high except breast cancer and lung cancer as a common disease gene related to other diseases. When it comes to breast cancer, PPP1R13L, FOSL2, and ERBB4 were detected as new disease genes [43–45]. Furthermore, JUN, PIK3R1, and HTT genes were detected as tumor progression genes for colorectal cancer [46–48] and PPP1R13L gene for lung cancer disease [49]. In addition, for lung cancer disease it is able to detect some genes which affect therapy, such as ERBB4 and PLK1 genes [50, 51].

By considering the overall process, subcellular localization data, protein's small world clustering coefficient, and PCC of gene expression values gave a very good combination for calculating reliability of the PPIs. Furthermore experiment results showed that combining protein's proximity and disease similarity concepts resulted in improved performance in identifying and ranking candidate genes for a specific disease.

4. Conclusion

Prioritizing disease related genes is one of the important challenges in the field of bioinformatics. In order to address this challenge different computational methods have been introduced in past years. From the literature review it has been suggested that it is important to incorporate topological similarity with disease similarity in an algorithm for



(a) ROC curves, breast cancer disease

(b) ROC curves, Alzheimer's disease

(c) ROC curves, colorectal cancer disease

(d) ROC curves, diabetes mellitus disease

(e) ROC curves, lung cancer disease

(f) ROC curves, prostate cancer disease

Figure 5: ROC curves: (a) breast cancer, (b) Alzheimer's disease, (c) colorectal cancer, (d) diabetes disease, (e) lung cancer, and (f) prostate cancer.

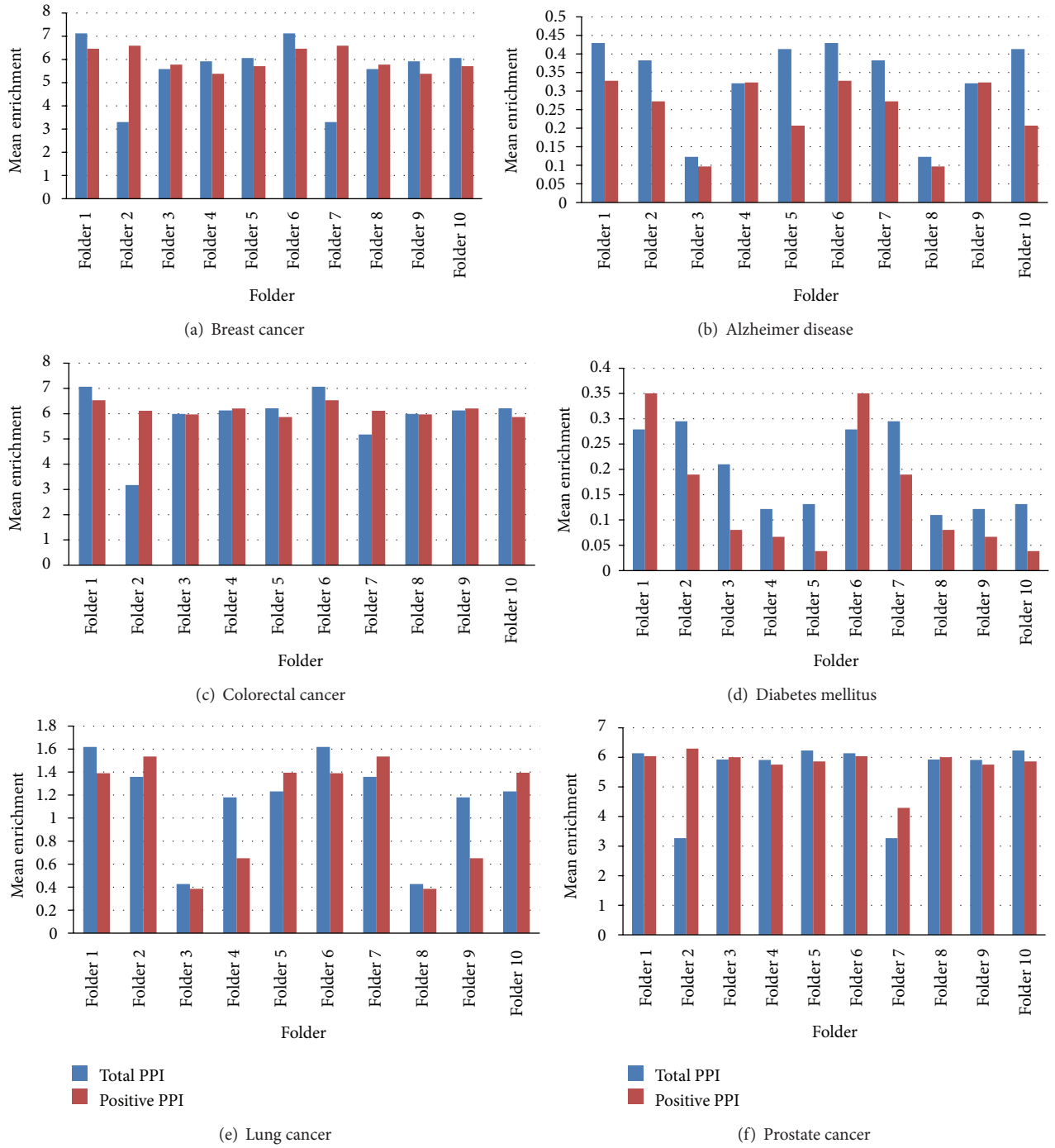


Figure 6: Tenfold cross validation for true disease gene prediction: (a) breast cancer, (b) Alzheimer disease, (c) colorectal cancer, (d) diabetes mellitus, (e) lung cancer, and (f) prostate cancer.

prioritizing genes related to a particular disease. This paper has proposed a new algorithm called ProSim. In this study, topological similarity is calculated using a random walk with restart method while disease similarity is calculated using the method introduced by van Driel. The performance of the proposed method is evaluated by comparing with three other methods, PRINCE, RWR, and DADA. Leave-one-out cross validation, mean enrichment, and ROC curves are the main evaluation techniques. Furthermore, the proposed method is able to predict disease genes effectively from a PPI network which consists of positive and negative PPIs. Last but not least it is able to identify some important candidate genes, previously ranked low by other methods, which include TP53, BRCA1, JUN, and PSEN1. Even though it outperforms existing methods considered, further experiments should be carried out to fine-tune its performance by including

other biological data such as tissue-specific details as well as incorporating other mathematical procedures.

Conflict of Interests

The authors declare that there is no conflict of interests.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant nos. 61232001, 61428209, and 61420106009 and the Program for New Century Excellent Talents in University NCET-12-0547.

References

[1] D. Botstein and N. Risch, "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease," *Nature Genetics*, vol. 33, pp. 228–237, 2003.

[2] S. Aerts, D. Lambrechts, S. Maity et al., "Gene prioritization through genomic data fusion," *Nature Biotechnology*, vol. 24, pp. 537–544, 2006.

[3] K. Lage, E. O. Karlberg, Z. M. Størling et al., "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nature Biotechnology*, vol. 25, no. 3, pp. 309–316, 2007.

[4] M. Li, J. Zhang, Q. Liu, J. Wang, and F.-X. Wu, "Prediction of disease-related genes based on weighted tissue-specific network by using DNA methylation," *BMC Medical Genomics*, vol. 7, article S4, supplement 2, 2014.

[5] G. U. Ganegoda, J. Wang, F. X. Wu, and M. Li, "Prediction of disease genes using tissue-specified gene-gene network," *BMC Systems Biology*, vol. 8, supplement 3, p. S3, 2014.

[6] H. Ge, Z. Liu, G. M. Church, and M. Vidal, "Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*," *Nature Genetics*, vol. 29, no. 4, pp. 482–486, 2001.

[7] A. Grigoriev, "A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*," *Nucleic Acids Research*, vol. 29, no. 17, pp. 3513–3519, 2001.

[8] J. Chen, B. J. Aronow, and A. G. Jegga, "Disease candidate gene identification and prioritization using protein interaction networks," *BMC Bioinformatics*, vol. 10, article 73, 2009.

[9] T. Jacquemin and R. Jiang, "Walking on a tissue-specific disease-protein-complex heterogeneous network for the discovery of disease-related protein complexes," *BioMed Research International*, vol. 2013, Article ID 732650, 12 pages, 2013.

[10] Q. Zou, J. Li, C. Wang, and X. Zeng, "Approaches for recognizing disease genes based on network," *BioMed Research International*, vol. 2014, Article ID 416323, 10 pages, 2014.

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

[12] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Molecular Systems Biology*, vol. 4, article 189, 2008.

[13] 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.

[14] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. M. Leunissen, "A text-mining analysis of the human phenome," *European Journal of Human Genetics*, vol. 14, no. 5, pp. 535–542, 2006.

[15] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *The American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

[16] M. Li, Q. Li, G. U. Ganegoda, J. Wang, F. X. Wu, and Y. Pan, "Prioritization of orphan disease-causing genes using topological feature and GO similarity between proteins in interaction networks," *Science China Life Sciences*, vol. 57, no. 11, pp. 1064–1071, 2014.

[17] W. Peng, J. Wang, Y. Cheng, Y. Lu, F. X. Wu, and Y. Pan, "UDoNC: an algorithm for identifying essential proteins based on protein domains and protein-protein interaction networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, no. 99, 1 page, 2014.

[18] M. Li, Y. Lu, J. X. Wang, F. X. Wu, and Y. Pan, "A topology potential-basedmethod for identifying essential proteins from PPI networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. PP, no. 99, p. 1, 2014.

[19] J. P. Gonçalves, A. P. Francisco, Y. Moreau, and S. C. Madeira, "Interactogeneous: disease gene prioritization using heterogeneous networks and full topology scores," *PLoS ONE*, vol. 7, no. 11, Article ID e49634, 2012.

[20] C.-Y. Pang, W. Hu, B.-Q. Hu et al., "A special local clustering algorithm for identifying the genes associated with alzheimers disease," *IEEE Transactions on Nanobioscience*, vol. 9, no. 1, pp. 44–50, 2010.

[21] G. U. Ganegoda, J. Wang, F.-X. Wu, and M. Li, "Prioritization of candidate genes based on disease similarity and protein's proximity in PPI networks," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '13)*, pp. 103–108, Shanghai, China, December 2013.

[22] M. Li, R. Zheng, H. Zhang, J. Wang, and Y. Pan, "Effective identification of essential proteins based on priori knowledge, network topology and gene expressions," *Methods*, vol. 67, no. 3, pp. 325–333, 2014.

[23] R. Sharan, S. Suthram, R. M. Kelley et al., "Conserved patterns of protein interaction in multiple species," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 6, pp. 1974–1979, 2005.

[24] S. Erten, G. Bebek, and M. Koyutürk, "VAVIEN: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks," *Journal of Computational Biology*, vol. 18, no. 11, pp. 1561–1574, 2011.

[25] D. Masotti, C. Nardini, S. Rossi et al., "TOM: enhancement and extension of a tool suite for in silico approaches to multigenic hereditary disorders," *Bioinformatics*, vol. 24, no. 3, pp. 428–429, 2008.

[26] G. Bebek and J. Yang, "PathFinder: mining signal transduction pathway segments from protein-protein interaction networks," *BMC Bioinformatics*, vol. 8, article 335, 2007.

[27] H. B. Shen and K. C. Chou, "A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0," *Analytical Biochemistry*, vol. 394, no. 2, pp. 269–274, 2009.

[28] W.-K. Huh, J. V. Falvo, L. C. Gerke et al., "Global analysis of protein localization in budding yeast," *Nature*, vol. 425, no. 6959, pp. 686–691, 2003.

[29] S. Erten, G. Bebek, R. M. Ewing, and M. Koyutürk, "DADA: degree-aware algorithms for network-based disease gene prioritization," *BioData Mining*, vol. 4, no. 1, article 19, 2011.

[30] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS Computational Biology*, vol. 6, no. 1, Article ID e1000641, 2010.

[31] D. Zhou, O. Bousquet, T. Navin Lal, and B. Scholkopf, "Learning with local and global consistency," in *Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, 2004.

[32] C. D. Chen and C. L. Sawyers, "NF-κB activates prostate-specific antigen expression and is upregulated in androgen-independent prostate cancer," *Molecular and Cellular Biology*, vol. 22, no. 8, pp. 2862–2870, 2002.

[33] D. S. Rao, T. S. Hyun, P. D. Kumar et al., "Huntingtin-interacting protein 1 is overexpressed in prostate and colon cancer and is critical for cellular survival," *The Journal of Clinical Investigation*, vol. 110, no. 3, pp. 351–360, 2002.

[34] M. Thompson, J. Lapointe, Y.-L. Choi et al., "Identification of candidate prostate cancer genes through comparative expression-profiling of seminal vesicle," *The Prostate*, vol. 68, no. 11, pp. 1248–1256, 2008.

[35] T. H. Ecke, H. H. Schlechte, K. Schiemenz et al., "TP53 gene mutations in prostate cancer progression," *International Journal of Cancer Research and Treatments*, vol. 30, no. 5, pp. 1579–1586, 2010.

[36] H. Wan, M. Wu, S. Yu, W. Qiang, and T. Liu, "Mutation and expression of nm23-H1 and CD44 gene in prostate cancers," *Zhonghua Wai Ke Za Zhi*, vol. 38, no. 5, pp. 382–384, 2000.

[37] N. Konishi, K. Shimada, M. Nakamura et al., "Function of JunB in transient amplifying cell senescence and progression of human prostate cancer," *Clinical Cancer Research*, vol. 14, no. 14, pp. 4408–4416, 2008.

[38] R. K. Mandal and R. D. Mittal, "Are cell cycle and apoptosis genes associated with prostate cancer risk in North Indian population?" *Urologic Oncology: Seminars and Original Investigations*, vol. 30, no. 5, pp. 555–561, 2012.

[39] A. Antonell, M. Balasa, R. Oliva et al., "A novel *PSEN1* gene mutation (L235R) associated with familial early-onset Alzheimer's disease," *Neuroscience Letters*, vol. 496, no. 1, pp. 40–42, 2011.

[40] H. Okazawa and S. Estus, "The JNK/c-Jun cascade and Alzheimer's disease," *American Journal of Alzheimer's Disease and other Dementias*, vol. 17, no. 2, pp. 79–88, 2002.

[41] S. Chen, W. Yan, J. Huang, D. Ge, Z. Yao, and D. Gu, "Association analysis of the variant in the regulatory subunit of phosphoinositide 3-kinase (p85α) with Type 2 diabetes mellitus and hypertension in the Chinese Han population," *Diabetic Medicine*, vol. 22, no. 6, pp. 737–743, 2005.

[42] L. Qu, B. He, Y. Pan et al., "Association between polymorphisms in *RAPGEF1*, *TP53*, *NRF1* and type 2 diabetes in Chinese Han population," *Diabetes Research and Clinical Practice*, vol. 91, no. 2, pp. 171–176, 2011.

[43] B. A. Nexø, U. Vogel, A. Olsen et al., "Linkage disequilibrium mapping of a breast cancer susceptibility locus near *RAI/PPP1R13L/iASPP*," *BMC Medical Genetics*, vol. 9, article 56, 2008.

[44] K. J. Meaburn, P. R. Gudla, S. Khan, S. J. Lockett, and T. Misteli, "Disease-specific gene repositioning in breast cancer," *The Journal of Cell Biology*, vol. 187, no. 6, pp. 801–812, 2009.

[45] H.-C. Kim, J.-Y. Lee, H. Sung et al., "A genome-wide association study identifies a breast cancer risk variant in *ERBB4* at 2q34: results from the Seoul Breast Cancer Study," *Breast Cancer Research*, vol. 14, no. 2, article R56, 2012.

[46] H. Wang, M. Birkenbach, and J. Hart, "Expression of Jun family members in human colorectal adenocarcinoma," *Carcinogenesis*, vol. 21, no. 7, pp. 1313–1317, 2000.

[47] L. Li, S. J. Plummer, C. L. Thompson, T. C. Tucker, and G. Casey, "Association between phosphatidylinositol 3-kinase regulatory subunit p85α *Met326Ile* genetic polymorphism and colon cancer risk," *Clinical Cancer Research*, vol. 14, no. 3, pp. 633–637, 2008.

[48] S. Savas, A. Hyde, S. N. Stuckless, P. Parfrey, H. B. Younghusband, and R. Green, "Serotonin transporter gene (SLC6A4) variations are associated with poor survival in colorectal cancer patients," *PLoS ONE*, vol. 7, no. 7, Article ID e38953, 2012.

[49] J. Yin, L. Guo, C. Wang et al., "Effects of PPP1R13L and CD3EAP variants on lung cancer susceptibility among nonsmoking Chinese women," *Gene*, vol. 524, no. 2, pp. 228–231, 2013.

[50] A. Starr, J. Greif, A. Vexler et al., "ErbB4 increases the proliferation potential of human lung cancer cells and its blockage can be used as a target for anti-cancer therapy," *International Journal of Cancer*, vol. 119, no. 2, pp. 269–274, 2006.

[51] B. Spänkuch-Schmitt, J. Bereiter-Hahn, M. Kaufmann, and K. Strebhardt, "Effect of RNA silencing of polo-like kinase-1 (PLK1) on apoptosis and spindle formation in human cancer cells," *Journal of the National Cancer Institute*, vol. 94, no. 24, pp. 1863–1877, 2002.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 789516, 9 pages
http://dx.doi.org/10.1155/2015/789516

Research Article

Differential Expression Analysis in RNA-Seq by a Naive Bayes Classifier with Local Normalization

Yongchao Dou,[1] Xiaomei Guo,[2] Lingling Yuan,[2] David R. Holding,[2,3] and Chi Zhang[1,3]

[1]*School of Biological Sciences, University of Nebraska, Lincoln, NE 68588, USA*
[2]*Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE 68588, USA*
[3]*Center for Plant Science and Innovation, University of Nebraska, Lincoln, NE 68588, USA*

Correspondence should be addressed to Chi Zhang; czhang5@unl.edu

Received 6 November 2014; Accepted 15 March 2015

Academic Editor: Ji-shou Ruan

To improve the applicability of RNA-seq technology, a large number of RNA-seq data analysis methods and correction algorithms have been developed. Although these new methods and algorithms have steadily improved transcriptome analysis, greater prediction accuracy is needed to better guide experimental designs with computational results. In this study, a new tool for the identification of differentially expressed genes with RNA-seq data, named GExposer, was developed. This tool introduces a local normalization algorithm to reduce the bias of nonrandomly positioned read depth. The naive Bayes classifier is employed to integrate fold change, transcript length, and GC content to identify differentially expressed genes. Results on several independent tests show that GExposer has better performance than other methods. The combination of the local normalization algorithm and naive Bayes classifier with three attributes can achieve better results; both false positive rates and false negative rates are reduced. However, only a small portion of genes is affected by the local normalization and GC content correction.

1. Introduction

RNA-Seq is a technology based on next-generation sequencing to determine transcript abundance, transcriptional structure of genes, and posttranscriptional modifications. It is essential to accurately construct genome-wide gene expression profiles in order to interpret the functional elements of the genome, molecular constituents of cells, development of organisms, and mechanism of diseases [1]. RNA-seq has many advantages over microarray such as high resolution, low background noise, no requirement on prior knowledge of reference sequences, and the ability to distinguish isoforms and allelic expression [1]. RNA-seq data are typically generated from a library of cDNA fragments made from a population of mRNAs. Then cDNAs are sequenced *en masse* with or without amplification. There are two steps in analyzing the RNA-seq reads. The obtained short reads are first aligned to a reference genome or transcriptome, and, in the second step, for a given gene, the numbers of reads are compared between two different samples. The number of short reads mapped onto one gene is the count that is taken as a measure of the expression level of the gene. Many different types of analyses can be applied to the results of short-read alignment, including single nucleotide polymorphism discovery, alternative transcript identification, and gene expression profiling.

Because of the importance of RNA-seq, many methods have been developed to analyze aligned RNA-seq data to identify differentially expressed (DE) genes over the last four years. They include edgeR [2], DESeq [3], Cuffdiff [4], baySeq [5], TSPM [6], NBPSeq [7], BitSeq [8], POME [9], NOISeq [10], Gfold [11], and MRFSeq [12]. EdgeR [2], the first statistical method developed for digital gene expression data, is a parametric statistical method, which is based on a negative binomial model (an overdispersed Poisson model) [13]. DESeq [3] is also a parametric statistical method based on the negative binomial model. When estimating variances, DESeq and edgeR both employ gene information but edgeR estimates the gene-wise variance or dispersion by conditional maximum likelihood conditioning on the total count for that gene [14]. Cuffdiff [4], a part of the Cufflinks package developed for the identification of differentially expressed

genes and revealing differential splicing events, uses a similar normalization method as DESeq and specifically addresses the uncertainties of read counts caused by ambiguous reads from different but similar isoforms. The baySeq [5], another parametric statistical method using a negative binomial model, takes a Bayesian approach which assumes that non-differentially expressed genes should possess the same prior distribution on the underlying parameters across conditions, while differentially expressed genes should possess variant parameters for prior distributions. NBPSeq [7] is based on an overparameterized version of the negative binomial distribution that is called an NBP model. BitSeq [8] is a recently developed method, which estimates the distribution of transcript levels based on a probabilistic model of the read generation process and is simulated with a Markov chain Monte Carlo (MCMC) algorithm. BitSeq estimates the variance in the transcript expression based on a hierarchical log-normal model and determines the probability of differential expression by Bayesian model averaging. POME is another recently developed algorithm for gene expression analysis with RNA-seq, which uses Poisson mixed-effects model to characterize base-level read coverage within each transcript [9]. NOISeq [10] is a nonparametric statistical method, and several different normalization methods for the raw read counts are implemented with NOISeq, including RPKM (reads per kilobase of exon model per million mapped reads) [15], TMM [16], and UQUA [17]. Gfold is designed for samples without replicates, and significantly differentially expressed genes are determined based on the posterior distribution of their log fold changes [11]. MRFSeq [12] combines a Markov random field (MRF) model and the gene coexpression data to predict differential gene expression. Recently, a quantile normalization method has been developed to remove technical variability in RNA-seq data [18].

The transcript abundance of genes causes bias in detecting differential expression [19]. Nonuniform read coverage as a result of experimental protocols and bias caused by local sequence context also exists and some correction methods have been developed. The biases from the GC content can be corrected by base-level correction methods, such as the random hexamer bias correction method [20] and multiple additive regression trees (MART) [21]. Additional gene-level methods [22, 23] are developed to detect GC content biases and dinucleotide frequencies based on aggregated read counts for each gene and to remove the GC content bias trend across genes. Other types of GC content correction algorithms have also been developed [24, 25]. It has been reported that even after global normalization, longer transcripts are more likely to be called as differentially abundant compared to the shorter ones using t-tests [26]. Gao et al. developed an algorithm for transcript length normalization based on Poisson models; each gene's test statistics were adjusted using the square root of the transcript length followed by testing for gene set using the Wilcoxon rank-sum test [27]. Both positional [28] and sequence-specific [20, 29] biases are identified in sequenced fragments. Positional bias refers to a local effect in which fragments are preferentially located towards either the beginning or the end of transcripts, and sequence-specific bias is a global effect where the sequence surrounding the beginning or the end of the potential fragment affects its likelihood of being selected for sequencing. These biases can affect expression estimates [21], and Roberts et al. designed algorithms to correct these biases [23].

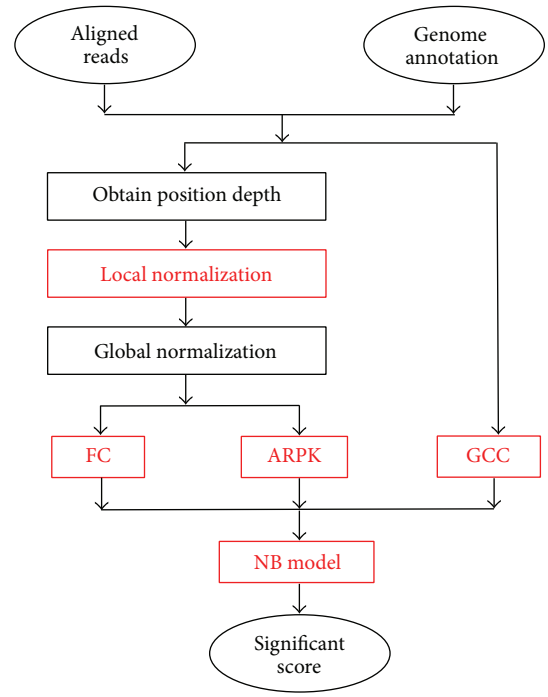


Figure 1: Flowchart of GExposer.

In our research, we have identified a new bias of mapped reads, the unevenly positioned reads depth, and designed a local normalization algorithm to correct this bias. Based on that, we have developed an analysis method using a naive Bayes (NB) classifier to determine DE genes. This method, called GExposer, uses three attributes: fold change (FC), averaged reads per kilobase (ARPK), and relative GC content (GCC). GExposer has performed the best or among the best when compared with other statistical methods and tested on multiple data sets beyond the one the model was constructed on. The software tool is available at http://sysbio.unl.edu/.

2. Materials and Methods

Figure 1 shows the flow chart of GExposer. The inputs of this tool are aligned reads and their genome annotations. All data were preprocessed with the same protocol. The raw short reads underwent the quality control with FASTX-Toolkit package [30], and all low quality reads, that is, average scores < 20, were removed. Any short read mapping tool can be used to align RNA-seq reads to a reference genome. In this study, all RNA-seq reads were mapped by Bowtie [31] allowing up to two mismatches, and the reads mapped to multiple locations were discarded. Numbers of reads in genes were counted by the HTSeq-count tool using corresponding gene annotations, and the "union" resolution mode was used

[32]. The distribution of read depths in each transcript can be determined according to the gene annotation as well. With a Poisson model, all read depths in one transcript are adjusted, and abnormal high or low depths are modified. This step is referred to as the local normalization. The local normalization can reduce the noise from nonspecific resources and the variation among replicates. Therefore, the variation among replicates is not considered as an attribute in the following NB classifier. Then, the global normalization algorithm, TMM, designed in DESeq package [3] is used to normalize reads depths between samples, and three attributes of the gene reads are extracted: FC, ARPK, and relative GC content. Generally, FC shows the relative difference between two samples and ARPK is the expression level of a transcript normalized by the transcript length. Previous studies also suggested that integrating GCC can help find DE genes [25], and, hence, GCC is included. Finally, a naive Bayes (NB) classifier is applied onto these features to score a given gene. The NB classifier is a simple probabilistic classifier based on applying Bayes theorem with naive independence assumptions. Generally, a probability model for a classifier is a conditional model $p(C \mid f_1, f_2, \ldots, f_n)$, where C is the class variable and f_i are attribute variables (e.g., GExposer has three attributes). Using Bayes' theorem, this conditional model can be transformed to

$$p(C \mid f_1, f_2, \ldots, f_n) = \frac{p(C)\, p(f_1, f_2, \ldots, f_n \mid C)}{p(f_1, f_2, \ldots, f_n)}. \quad (1)$$

Since a naive Bayes classifier has the conditional independence assumption, which assumes that, for a given class variable, the value of a particular attribute is unrelated to any other attributes, the conditional distribution over the class variable C becomes

$$p(C \mid f_1, f_2, \ldots, f_n) = \frac{p(C) \prod_{i=1}^{n} p(f_i \mid C)}{p(f_1, f_2, \ldots, f_n)}. \quad (2)$$

All model parameters, including class priors and attribute probability distributions, can be estimated from the training set with the method of maximum likelihood. An advantage of an NB classifier is that it only requires a small amount of training data to estimate the parameters, such as means and variances of the variables. The software package used for this work is the *R* package, e1071, which computes the conditional a posteriori probabilities of a categorical class variable using the Bayes rule. For each gene, its feature vector has three values for the three attributes, which will be described in the following sections in detail. For the special requirement, we modified the calculation of FC, ARPK, and GCC from their common definitions in bioinformatics to a normalized version to make sure their ranges are in $[0, 1]$. For the training step, the function of "naiveBayes" in the package e1071 is used to train a statistical model based on a training set, in which each gene has three attributes and known status. For the prediction step, the function of "prediction" can calculate the conditional a posteriori probability for a gene based on its feature vector with the three values. The conditional a posteriori probability is returned as a score, which is the probability of the given gene belonging to the class. The training data set and the *R* code are available for downloading from our website http://sysbio.unl.edu/.

2.1. Attributes. Three attributes are calculated with the following equations. The attribute of FC is calculated as

$$\mathrm{FC} = \frac{1}{1 + \left|\log_2 (M_1/M_2)\right|}, \quad (3)$$

where M_i $(i = 1, 2)$ are the mean values of the numbers of aligned reads for a given transcript in all replicates of sample S_i $(i = 1, 2)$. Here, M_i $(i = 1, 2)$ are values after both the local and global normalizations. The attribute of ARPK for a transcript is calculated as

$$\mathrm{ARPK} = \frac{1}{1 + \left|\log_2 \left((1000 \times (M_1 + M_2)) / (2 \times \mathrm{TL})\right)\right|}, \quad (4)$$

where M_i $(i = 1, 2)$ have the same definition as in FC calculation and TL is the transcript length. The attribute of GCC can be calculated as

$$\mathrm{GCC} = \frac{1}{1 + c/C}, \quad (5)$$

where c is the average GCC of reads mapped to a given transcript in both samples and C is the average GCC of all mapped reads. The GCC of one read is the ratio of the total number of guanine and cytosine to the length of the read. For each gene, the NB model will give a score in $[0, 1]$ by its three features. The higher the score, the higher possibility a gene is differentially expressed. Here, we also assume that the features of a given gene are independent of its specific biological background.

2.2. Local Normalization. Generally, reads are positioned randomly along every transcript in RNA-seq [33]. For a given transcript T with coordinates (t_1, t_n) on a chromosome, the number of reads starting from position t_i $(i = 1, 2, \ldots, n)$ is defined as r_i. The total number of reads mapped to the transcript is $R = \sum r_i$. As the assumption that r_i is an accumulation of random events, r_i can be modeled as

$$r_i \sim p(\lambda_T), \quad (6)$$

where $p(\lambda_T)$ is a Poisson distribution with parameter λ_T. The unbiased estimate of λ_T is

$$\lambda_T = \frac{1}{n} \sum_{i=1}^{n} r_i. \quad (7)$$

Confidence limits are the lower and upper boundaries of a confidence interval. With the Poisson distribution, we can find the upper confidence limits (UCL) and lower confidence limits (LCL) for λ_T, such as at 97.5% and 2.5% confidence level, respectively, in this paper. With the UCL and LCL, the number of reads for a given position, r_i, can be corrected as

$$r_i^0 = \begin{cases} \mathrm{UCL}, & \text{if } r_i > \mathrm{UCL}, \\ \mathrm{LCL}, & \text{if } r_i < \mathrm{LCL}, \\ r_i, & \text{othewise.} \end{cases} \quad (8)$$

Table 1: Summary of all data sets used in this paper.

Data set	DE genes	NDE genes	SRA accession number
MAQC UHRR and HBRR	1966	3388	SRA010153.1
Colorectal cancer	13	0	SRX026158 and SRX026158
Maize leaf	6	9	SRA012297

With this adjustment, the total number of reads mapped to the transcript is

$$R^0 = \sum_{i=1}^{n} r_i^0. \quad (9)$$

It has been reported that the assumption of a Poisson distribution is too restrictive to predict more accurate variations among data from different replicates [2, 3], but a Poisson distribution can fit data in a specific exon from one sample, because they have fewer variations.

2.3. Training and Testing Data Set. We collected training and testing data sets from several different resources, and Table 1 summarized these data sets. The training data set contains two RNA-seq data sets with 35 base-pair-long reads obtained using Illumina's Genome Analyzer II high-throughput sequencing system [17], and they correspond to data obtained by the microarray quality control (MAQC) project [34]. The accession number of these RNA-seq data in SRA is SRA010153.1. The two RNA sample types used were a universal human reference RNA (UHRR) from Stratagene and a human brain reference RNA (HBRR) from Ambion. There are seven lanes for each sample with about 40 million reads. After processing the RNA-seq data, all RNA-seq reads were aligned against the human genome (GRCh37.68). The data set has about 997 RT-PCR data for validation of RNA-seq analysis results [35], and the genes with mean reads number fewer than 5 in both samples are not considered. Based on their expression ($\log_2$ fold change), the genes were grouped into three sets: DE, no-call, and non-DE (NDE), with the $\log_2$(fold-change) being >1.5 [0.5, 1.5] and <0.5, respectively. The expression $\log_2$(fold-change) for RT-PCR samples was calculated by the ΔΔCT method [36]. This way, we compiled 389, 178, and 235 genes in the categories of DE, no-call, and NDE, respectively. For the same RNA-seq data, corresponding microarray experiments were conducted by MAQC with Affymetrix Human Genome U133 Plus 2.0 arrays (GEO: GSE5350). Microarray data were preprocessed with RMA [37] and analyzed with limma package [38]. For the results of the microarray data analysis, genes having absolute $\log_2$(fold-change) ≥1.5 and P values $<10^{-3}$ were considered as DE; genes were NDE if their absolute $\log_2$(fold-change) <0.5, and the rest were no-call genes. Finally, there are 1756, 2340, and 3372 genes for DE, no-call, NDE, respectively. The classified genes by both PCR and microarray were combined together to be used as the training data set for the NB model. A gene was considered as DE (or NDE) if at least one method, either PCR or microarray assay, confirmed it as DE (or NDE), and finally there are 1966 DE and 3388 NDE genes in this training set. The NB model trained by this data set is also applied to other species, including plants, for testing. More details about other test data sets are described in the following sections.

3. Results and Discussion

3.1. Results for the Training Data Set. The NB model was trained using 1966 DE and 3388 NDE genes in the training data set. DE genes were considered as positive while NDE ones were considered as negative. For validation, the leave-one-out cross-validation was used to score them. To assess the performance of GExposer, the current most popular methods such as edgeR (2.6.7) [2], DESeq (1.12.0) [3], Cuffdiff (2.1.1) [4], NOISeq (2.0) [10], and Gfold (1.0.7) [11] were applied to the same RNA-seq data sets for comparison. The default setups were used for other methods as well. Following the work of Tarazona et al. [10], P values created by the other methods, except for NOIseq, were used as the scores for ranking genes. NOIseq outputs one score for each gene to quantify the expression level. Since different parameters are used by different methods to select DE genes, it is difficult to select a cutoff that can produce comparable analysis and fair comparison for all methods. In this study, we compared the area under receiver operating characteristic curve (AUC) values of all methods, which can avoid the difficulty of selecting a comparable cutoff of P values for all methods. This evaluation method has been used to other RNA-seq data analysis tools before [17, 39]. A receiver operating characteristic (ROC) curve represents a dependency of sensitivity and (1 − specificity), which is plotted with true positives rate versus false positive rate at various threshold settings. To change the threshold setting, the number of the predicted DE genes was increased in steps of one gene. Figure 2 shows ROC curves of all methods for the same data set. The AUC values of all methods are shown in Table 2. GExposer achieved the highest AUC value (0.9255). To test the ability of each method to successfully identify DE (true positive) or NDE (true negative) from a noisy pool, no-call genes are treated as true negative (NDE) or true positive (DE) genes, respectively. For no-call genes, the model trained by all DE and NDE genes was used to score them with an NB classifier. All AUC values are also shown in Table 2. For both cases, GExposer achieved the highest AUC values.

3.2. Independence to the Number of Replicates. To study the dependence on the number of replicates, six RNA-seq analysis methods were applied on some subsets of the training data set, from one lane to seven lanes. Results of these methods are shown in Table 3. Results of all methods are relatively stable with more than one lane, but there is a sharp drop in AUC values for Cuffdiff and DESeq when

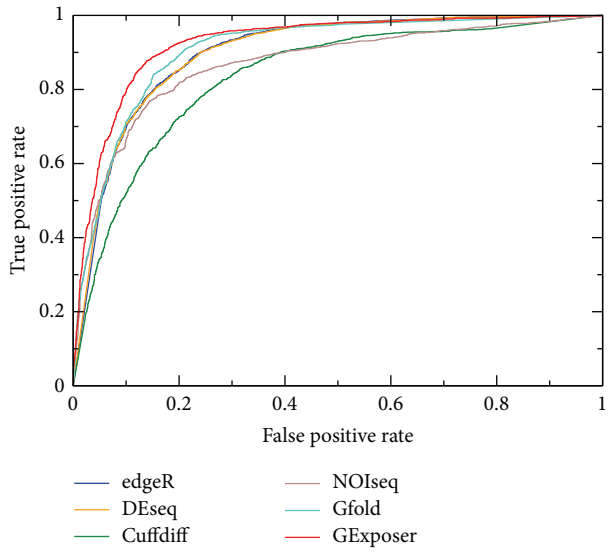


Figure 2: ROC curves of different methods tested on the training data set.

Table 2: AUC values of six methods on the training data set with the leave-one-out cross-validation.

Method		False positive test No-call as NDE	False negative test No-call as DE
edgeR	0.8997	0.8567	0.7945
DESeq	0.9002	0.8602	0.7909
Cuffdiff	0.8347	0.7740	0.7610
NOISeq	0.8679	0.8460	0.7267
Gfold	0.9079	0.8886	0.7790
GExposer	0.9255	0.9030	0.8054

only one lane is available. If there is no replicate, the AUC value for DESeq is no more than 0.67 and for Cuffdiff is no more than 0.60. GExposer constantly has the largest AUC values for different numbers of replicates. Moreover, although Gfold is designed for data without replicate, GExposer still outperformed it even when only using one lane. This test implies that the local normalization reduces the variation among replicates, and, hence, the performance of GExposer is not affected by replicates.

3.3. Results from Human Colorectal Cancer Data Set. An RNA-seq data set for human colorectal cancer generated by Griffith et al. [40] was used to perform an independent test. This data set is from the same species as the training set, but from a different tissue. In this data set, 84 bp paired-end reads were sequenced, and there are eight lanes in total for colorectal cancer cell line MIP/5FU and 15 lanes for cell line MIP101 [40]. In this data set, the top 50 differential or alternative expression events were tested and 13 genes were confirmed as DE by experiments. For the human colorectal cancer data set, all 23 lanes with 84 bp length reads were applied to different RNA-seq analysis tools, and the 13 known DE genes were ranked using these tools. The orders of these genes ranked by different methods are shown in Table 4. One can find that 7 out of the 13 confirmed genes are ranked in the top 50 genes by GExposer. The numbers for other methods are no more than six except for Gfold. In particular, Gene TSPAN12 is ranked in top 50 only by GExposer and Cuffdiff. In this data set, some genes, such as MR1, have long transcript length with many exons, and an isoform with a small portion of exons is expressed. GExposer could not rank this kind of genes to top position because of the transcript length correction. Using the length of a specific isoform, instead of the total length of all exons for a given gene, could be considered to improve GExposer's accuracy of identification of differentially expressed isoform.

3.4. Performance of GExposer on Maize RNA-Seq. In contrast to other RNA-seq analysis tools, GExposer needs a training set, which was a set of RNA-seq data from human tissues for this work. Naturally, one may raise the question whether the performance of GExposer could have potential correlation with the species that is used to generate the RNA-seq data. To test whether the human-data trained GExposer works well on different species, the method was applied to plant (maize) RNA-seq data. The maize RNA-seq data set that compares bundle sheath and mesophyll cells were obtained from the laser-capture microdissected (LCM) samples from the tip of the maize leaf which incorporates two biological replicates [41]. The RNA-seq data were obtained from the NCBI short read archive (SRA) under accession number SRA012297. The expression levels of 40 genes (only 37 genes were found in the current version of B73 gene annotation) were measured by RT-PCR and were also grouped into three sets: DE (6), no-call (22), and NDE (9), as $\log_2$(fold-change) is >1.5 [0.5, 1.5] and <0.5, respectively, for the comparison between cells at a maturing zone (+4 cm above the leaf two ligule) and mature zone (tip, +1 cm below the leaf three tip). All RNA-seq analysis tools were applied onto this RNA-seq data set, and these 37 genes were ranked by different methods. The distributions of top 10 genes ranked by different methods are shown in Table 5. In top 10 genes, GExposer has the largest number of DE genes (5 genes), while DESeq has the least (3 genes). None of these methods rank NDE genes in top 10, except for Gfold.

3.5. Performance of Local Normalization. GExposer was also applied onto the maize RNA-seq data to test the performance of the local normalization method. The RNA-seq data were generated to study differentially expressed genes in quality protein maize (QPM) endosperm tissue [42]. To simplify the results, the RNA-seq data from the two genotypes, W64A *o2* mutant and K0326Y QPM, were used because it is the most important comparison for the QPM study. Each sample has about 20 million 50 bp long reads made up of kernels pooled from five biological replicates. Reads for maize were aligned against the reference genome (ZmB73-RefgenV2), for the pair-wise comparison between the genotypes, W64 *o2* and K0326Y QPM. There is a portion of genes that are assigned with different expression levels when using or not using the local normalization method. We selected seven such genes that have been assigned different fold changes

Table 3: AUC values of six methods on the training data set with different number of replicates.

Method	1	2	3	4	5	6	7
False positive test, no-call as NDE							
edgeR	0.8535	0.8607	0.8595	0.8591	0.8584	0.8574	0.8567
DESeq	0.6621	0.8626	0.8622	0.8618	0.8613	0.8610	0.8602
Cuffdiff	0.5963	0.7904	0.7772	0.7904	0.7773	0.7748	0.7740
NOIseq	0.8334	0.8392	0.8425	0.8445	0.8452	0.8456	0.8460
Gfold	0.8870	0.8334	0.8871	0.8875	0.8874	0.8886	0.8886
GExposer	0.8968	0.9016	0.9024	0.9024	0.9028	0.9032	0.9030
False negative test, no-call as DE							
edgeR	0.7845	0.7903	0.793	0.7934	0.7936	0.7934	0.7945
DESeq	0.5800	0.7871	0.7895	0.7902	0.7912	0.7905	0.7909
Cuffdiff	0.5894	0.7674	0.7645	0.7674	0.7623	0.7606	0.7610
NOIseq	0.7269	0.7342	0.7335	0.7308	0.7288	0.7276	0.7267
Gfold	0.7905	0.6702	0.7498	0.7670	0.7753	0.7753	0.7790
GExposer	0.7942	0.8015	0.8045	0.8051	0.8053	0.8051	0.8054

Table 4: Ranking of 13 genes by six different methods.

Gene	edgeR	DESeq	cuffdiff	NOIseq	Gfold	GExposer
LAPTM4B	109	109	99	**11**	**27**	**8**
TSPAN12	59	54	18	55	98	**10**
TNNI2	8750	8730	41066	671	241	15861
H19	**7**	**6**	**2**	**1**	**2**	**25**
ZNF185	651	672	56286	7879	81	1008
MR1	144	141	22837	51	**12**	4156
ASRGL1	125	269	8560	101	**10**	854
C12orf59	**13**	**11**	70	**4**	**1**	**2**
KLK6	680	646	1568	**17**	**36**	115
ATOH8	66	70	956	224	**21**	99
FUT3	57	62	1583	**6**	**5**	**9**
KRT20	356	625	462	399	**33**	**18**
OLR1	**48**	**42**	8741	**7**	**4**	**4**

The numbers in bold font correspond to genes that were ranked in top 50.

Table 5: The distributions of top 10 differentially expressed genes ranked by six different methods on maize RNA-seq data.

	DE	No-call	NDE
edgeR	4	6	0
DESeq	3	7	0
Cuffdiff	4	6	0
NOIseq	4	6	0
Gfold	4	5	1
GExposer	5	5	0

Table 6: Results of four maize genes with and without the local normalization.

Method	GRMZM2G002678	GRMZM2G018193	GRMZM2G096719	GRMZM2G388461
$\log_2$(FC) without local normalization	−1.20	1.69	−1.72	4.18
$\log_2$(FC) with local normalization	−0.61	0.40	−0.59	0.43

with and without local normalization and applied RT-PCR experiments to validate the real fold change. Since the gene annotations and sequences of B73 are used to design the primers for these genes in W64 *o2* and K0326Y QPM, only four genes, GRMZM2G002678, GRMZM2G018193, GRMZM2G096719, and GRMZM2G38846, showed results for RT-PCR experiments. The details of the RT-PCR experiment and primers of these four genes are described in Supplementary Data (see Supplementary Material available online at http://dx.doi.org/10.1155/2015/789516). For these four genes, their $\log_2$(fold-change) measured using the local normalization method and not using it is shown in Table 6. The $\log_2$(fold-change) measured without the local normalization step shows higher values, while that measured by GExposer with the local normalization is no more than 0.61. The RT-PCR results support the GExposer analysis with the local normalization. The RT-PCR results of these four genes are shown in Figure 3(a), and these four genes are not differentially expressed in W64 *o2* and K0326Y QPM. To understand what causes the difference, read coverage of three exons of GRMZM2G002678 are shown in Figure 3(b). Many reads aggregate in a very narrow peak in exon 5 for W64 *o2*, but there is no peak in K0326Y QPM. Therefore, this peak artificially inflates the absolute value of fold changes between W64 *o2* and K0326Y QPM. With the use of the local normalization method, the adjusted depth is low (at the level of the red dotted line), and this false positive is therefore removed.

3.6. Difference between True and Simulated Data Sets. In order to further evaluate the effect of nonrandomly positioned reads, the real and simulated RNA-seq data sets were compared. The next-generation sequencing read simulator "ART" [43] was used to simulate RNA-seq reads. To simulate the sequencing, the sequencing read simulator assumes that the reads uniformly and randomly distribute on the transcript

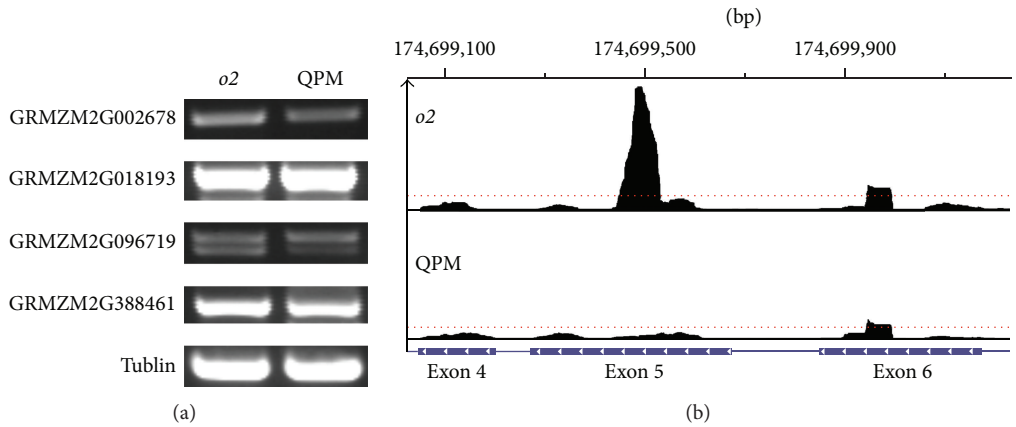


Figure 3: (a) RT-PCR results of four genes in *o2* and QPM lines. (b) Short read distribution in three exons of GRMZM2G002678 in W64 *o2* and QPM, and the red dotted line indicates the adjusted depth by local normalization method.

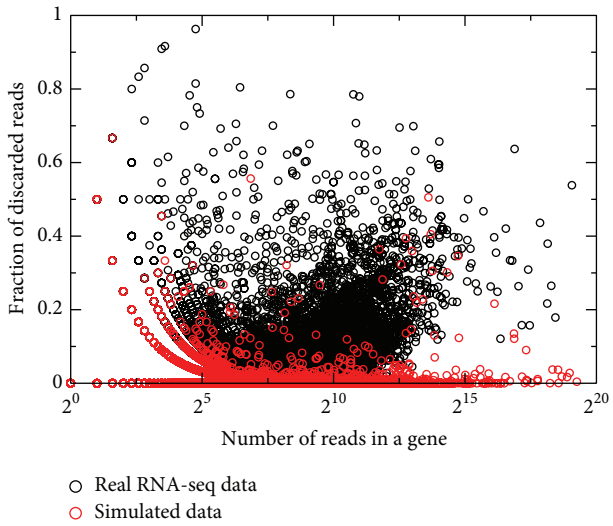


Figure 4: Fractions of discarded reads by local normalization method for both real and simulated RNA-seq data.

[33], and, hence, there is no abnormal peak to be adjusted by the local normalization method. We randomly selected W64 *o2* RNA-seq data from the maize endosperm data set as the template to conduct the simulation. For each transcript, the same number of reads as in the template data set was generated by the simulator. Then, all simulated reads were mapped to the same genome with the same parameters as for the W64 *o2* RNA-seq data. The fractions of the discarded reads by the local normalization method of each transcript are shown in Figure 4, and there is a significant difference between the real and simulated data sets. For the simulated data, only several genes (0.8% of all genes) have more than 10% reads discarded with the local normalization, but thousands of genes (10.3%) for the real RNA-seq data do. It indicates that in reality, sometimes, reads are not uniformly, randomly sequenced on a transcript, and the local normalized method is a necessary step for this kind of cases.

Table 7: Performance of GExposer omitting each attribute.

		False positive test No-call as NDE	False negative test No-call as DE
ΔGCC	0.9028	0.8964	0.8017
ΔARPK	0.8791	0.8656	0.747
ΔFC	0.6315	0.5946	0.6141
GExposer	0.9255	0.903	0.8054

3.7. Assessment of Each Attribute in GExposer. GExposer used three attributes: FC, ARPK, and relative GCC. To understand which one in the three scores plays a more important role, each attribute was removed from the system and the same training and test procedures were conducted to the training data set. The results are shown in Table 7. The absence of any attribute leads to some decrease of the AUC value, but the attribute of FC is the most significant. The largest changes occurred when the FC attribute was removed, whereas removing ARPK and GCC only caused small changes of AUC (0.0464 and 0.0227). It is not surprising that the fold change of read numbers is the major criterion to determine DE genes. The attribute ARPK is related to the expression level of a given gene and, hence, also plays an important role in the DE gene identification. The GCC correction is applied to a very small portion of genes. Therefore, the absence of the correction reduces the AUC value only slightly, although this correction is important for those specific genes.

3.8. Local Normalization and GC Content. The local normalization method mainly focuses on the high peaks of the reads; for a given high peak, its depth will be modified according to the average depth and its standard deviation. This raises one question: do the reads in these peaks have special patterns of nucleotides or GCCs, compared with the sequence background of all reads? If they have a special GCC, for example, some existing correction algorithms for GCC bias [20–23] could be applied to this case, instead of using the local normalization. To answer this question, we calculated the distributions of nucleotides in different types of reads.

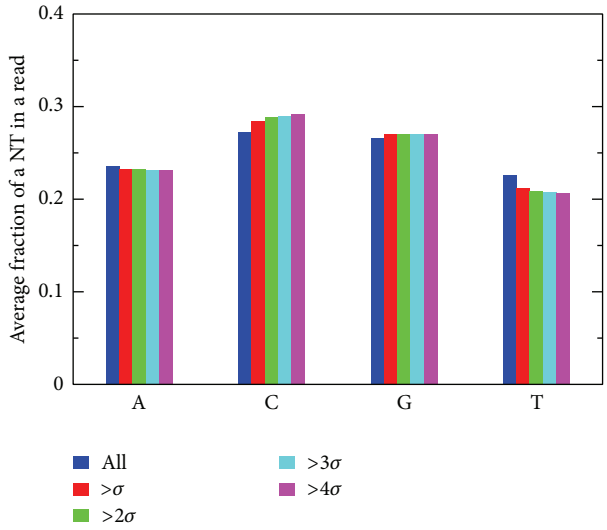


FIGURE 5: The distributions of nucleotides with different depths.

The result on maize W64 *o2* data set from the QPM RNA-seq data is shown in Figure 5. The blue bars are for all aligned reads, and the red, yellow, green, and brown bars are for the reads from peaks with depths of 1, 2, 3, and 4 standard deviations away from the average. The average portion of a type of nucleotide (NT) in one read was calculated. From Figure 5, one can see that, for a certain type of NT, different types of depths have very similar portions. This indicates that the abnormal high peaks of reads have no correlation with the GCC. Therefore, we can conclude that this kind of read abundance does not result from a special pattern of nucleotides.

4. Conclusions

In the study, a new bias in RNA-seq data called nonrandomly positioned reads was identified. Our analysis shows that this bias is different from GCC bias. In order to reduce the bias, a local normalization algorithm has been developed and the false positive rate caused by this bias is reduced, which has been validated by the RT-PCR experiments. Moreover, the combination of three attributes, FC, ARPK, and GCC, can achieve better results; both false positive rates and false negative rates are reduced. However, GCC correction is only applied to a very small portion of genes in a whole genome. The model of GExposer was trained by one data set, and there is great potential for machine learning methods to improve the performance in finding DE genes by combining more training data sets from different species. On the other hand, training data from various species could potentially limit the ability of a naive Bayes classifier to identify DE genes.

Conflict of Interests

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

Acknowledgments

The authors thank Dr. Peng Wang for valuable technical advices. This project was supported by funding under Chi Zhang and David R. Holding's startup funds from University of Nebraska, Lincoln, NE, and the Nebraska Soybean Board Fund.

References

[1] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[2] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

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

[4] C. Trapnell, B. A. Williams, G. Pertea et al., "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

[5] T. J. Hardcastle and K. A. Kelly, "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics*, vol. 11, article 422, 2010.

[6] P. L. Auer and R. W. Doerge, "A two-stage Poisson model for testing RNA-Seq data," *Statistical Applications in Genetics and Molecular Biology*, vol. 10, pp. 1–26, 2011.

[7] Y. M. Di, D. W. Schafer, J. S. Cumbie, and J. H. Chang, "The NBP negative binomial model for assessing differential gene expression from RNA-Seq," *Statistical Applications in Genetics and Molecular Biology*, vol. 10, no. 1, pp. 1–28, 2011.

[8] P. Glaus, A. Honkela, and M. Rattray, "Identifying differentially expressed transcripts from RNA-seq data with biological variation," *Bioinformatics*, vol. 28, no. 13, pp. 1721–1728, 2012.

[9] M. Hu, Y. Zhu, J. M. G. Taylor, J. S. Liu, and Z. S. Qin, "Using poisson mixed-effects model to quantify transcript-level gene expression in RNA-Seq," *Bioinformatics*, vol. 28, no. 1, pp. 63–68, 2012.

[10] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa, "Differential expression in RNA-seq: a matter of depth," *Genome Research*, vol. 21, no. 12, pp. 2213–2223, 2011.

[11] J. Feng, C. A. Meyer, Q. Wang, J. S. Liu, X. S. Liu, and Y. Zhang, "GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data," *Bioinformatics*, vol. 28, no. 21, pp. 2782–2788, 2012.

[12] E.-W. Yang, T. Girke, and T. Jiang, "Differential gene expression analysis using coexpression and RNA-Seq data," *Bioinformatics*, vol. 29, no. 17, pp. 2153–2161, 2013.

[13] M. D. Robinson and G. K. Smyth, "Small-sample estimation of negative binomial dispersion, with applications to SAGE data," *Biostatistics*, vol. 9, no. 2, pp. 321–332, 2008.

[14] G. K. Smyth and A. P. Verbyla, "A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models," *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 58, no. 3, pp. 565–572, 1996.

[15] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[16] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biology*, vol. 11, no. 3, article R25, 2010.

[17] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinformatics*, vol. 11, article 94, 2010.

[18] K. D. Hansen, R. A. Irizarry, and Z. Wu, "Removing technical variability in RNA-seq data using conditional quantile normalization," *Biostatistics*, vol. 13, no. 2, pp. 204–216, 2012.

[19] Z. Wu, B. D. Jenkins, T. A. Rynearson et al., "Empirical bayes analysis of sequencing-based transcriptional profiling without replicates," *BMC Bioinformatics*, vol. 11, article 654, 2010.

[20] K. D. Hansen, S. E. Brenner, and S. Dudoit, "Biases in Illumina transcriptome sequencing caused by random hexamer priming," *Nucleic Acids Research*, vol. 38, no. 12, p. e131, 2010.

[21] J. Li, H. Jiang, and W. H. Wong, "Modeling non-uniformity in short-read rates in RNA-Seq data," *Genome Biology*, vol. 11, no. 5, article R50, 2010.

[22] W. Zheng, L. M. Chung, and H. Zhao, "Bias detection and correction in RNA-Sequencing data," *BMC Bioinformatics*, vol. 12, article 290, 2011.

[23] A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter, "Improving RNA-Seq expression estimates by correcting for fragment bias," *Genome Biology*, vol. 12, no. 3, article R22, 2011.

[24] Y. Benjamini and T. P. Speed, "Summarizing and correcting the GC content bias in high-throughput sequencing," *Nucleic Acids Research*, vol. 40, no. 10, article e72, 2012.

[25] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit, "GC-content normalization for RNA-Seq data," *BMC Bioinformatics*, vol. 12, no. 1, article 480, 2011.

[26] A. Oshlack and M. J. Wakefield, "Transcript length bias in RNA-seq data confounds systems biology," *Biology Direct*, vol. 4, article 14, 2009.

[27] L. Gao, Z. Fang, K. Zhang, D. Zhi, and X. Cui, "Length bias correction for RNA-seq data in gene set analyses," *Bioinformatics*, vol. 27, no. 5, pp. 662–669, 2011.

[28] R. Bohnert and G. Rätsch, "rQuant.web: a tool for RNA-Seq-based transcript quantitation," *Nucleic Acids Research*, vol. 38, no. 2, pp. W348–W351, 2010.

[29] S. Srivastava and L. Chen, "A two-parameter generalized Poisson model to improve the analysis of RNA-seq data," *Nucleic Acids Research*, vol. 38, no. 17, article e170, 2010.

[30] J. Taylor, I. Schenck, D. Blankenberg, and A. Nekrutenko, "Using galaxy to perform large-scale interactive data analyses," in *Current Protocols in Bioinformatics*, chapter 10, unit 10.5, 2007.

[31] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, article R25, 2009.

[32] S. Anders, P. T. Pyl, and W. Huber, "HTSeq—a Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015.

[33] H. Richard, M. H. Schulz, M. Sultan et al., "Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments," *Nucleic Acids Research*, vol. 38, no. 10, article e112, 2010.

[34] L. Shi, L. H. Reid, W. D. Jones et al., "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology*, vol. 24, no. 9, pp. 1151–1161, 2006.

[35] R. D. Canales, Y. Luo, J. C. Willey et al., "Evaluation of DNA microarray results with quantitative gene expression platforms," *Nature Biotechnology*, vol. 24, no. 9, pp. 1115–1122, 2006.

[36] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate—a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 57, no. 1, pp. 289–300, 1995.

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

[38] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, pp. 1–25, 2004.

[39] C. Soneson and M. Delorenzi, "A comparison of methods for differential expression analysis of RNA-seq data," *BMC Bioinformatics*, vol. 14, article 91, 2013.

[40] M. Griffith, O. L. Griffith, J. Mwenifumbo et al., "Alternative expression analysis by RNA sequencing," *Nature Methods*, vol. 7, no. 10, pp. 843–847, 2010.

[41] P. Li, L. Ponnala, N. Gandotra et al., "The developmental dynamics of the maize leaf transcriptome," *Nature Genetics*, vol. 42, no. 12, pp. 1060–1067, 2010.

[42] X. Guo, K. Ronhovde, L. Yuan et al., "Pyrophosphate-dependent fructose-6-phosphate 1-phosphotransferase induction and attenuation of Hsp gene expression during endosperm modification in quality Protein Maize," *Plant Physiology*, vol. 158, no. 2, pp. 917–929, 2012.

[43] W. Huang, L. Li, J. R. Myers, and G. T. Marth, "ART: a next-generation sequencing read simulator," *Bioinformatics*, vol. 28, no. 4, pp. 593–594, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 918954, 10 pages
http://dx.doi.org/10.1155/2015/918954

Research Article

K-Profiles: A Nonlinear Clustering Method for Pattern Detection in High Dimensional Data

Kai Wang,[1] Qing Zhao,[2] Jianwei Lu,[2,3] and Tianwei Yu[4]

[1] *Department of Mathematics and Computer Science, Emory University, Atlanta, GA 30322, USA*
[2] *School of Software Engineering, Tongji University, Shanghai 200092, China*
[3] *The Advanced Institute of Translational Medicine and Department of Gastroenterology, Shanghai Tenth People's Hospital, Tongji University, Shanghai 200092, China*
[4] *Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA*

Correspondence should be addressed to Jianwei Lu; jwlu33@gmail.com and Tianwei Yu; tianwei.yu@emory.edu

Received 5 November 2014; Accepted 18 December 2014

Academic Editor: Fang-Xiang Wu

With modern technologies such as microarray, deep sequencing, and liquid chromatography-mass spectrometry (LC-MS), it is possible to measure the expression levels of thousands of genes/proteins simultaneously to unravel important biological processes. A very first step towards elucidating hidden patterns and understanding the massive data is the application of clustering techniques. Nonlinear relations, which were mostly unutilized in contrast to linear correlations, are prevalent in high-throughput data. In many cases, nonlinear relations can model the biological relationship more precisely and reflect critical patterns in the biological systems. Using the general dependency measure, Distance Based on Conditional Ordered List (DCOL) that we introduced before, we designed the nonlinear *K*-profiles clustering method, which can be seen as the nonlinear counterpart of the *K*-means clustering algorithm. The method has a built-in statistical testing procedure that ensures genes not belonging to any cluster do not impact the estimation of cluster profiles. Results from extensive simulation studies showed that *K*-profiles clustering not only outperformed traditional linear *K*-means algorithm, but also presented significantly better performance over our previous General Dependency Hierarchical Clustering (GDHC) algorithm. We further analyzed a gene expression dataset, on which *K*-profile clustering generated biologically meaningful results.

1. Introduction

In recent years, large amounts of high dimensional data have been generated from high-throughput expression techniques, such as gene expression data using microarray or deep sequencing [1], and metabolomics and proteomics data using liquid chromatography-mass spectrometry (LC-MS) [2]. Mining the hidden patterns inside these data leads to an enhanced understanding of functional genomics, gene regulatory networks, and so forth [3, 4]. However, the complexity of biological networks and the huge number of genes pose great challenges to analyze the big mass of data [5, 6]. Clustering techniques has usually been applied as a first step in the data mining process to analyze hidden structures and reveal interesting patterns in the data [7].

Clustering algorithms have been studied extensively in the last three decades, with many traditional clustering techniques successfully applied or adapted to gene expression data, which led to the discovery of biologically relevant groups of genes or samples [6]. Traditional clustering algorithms usually process data on the full feature space while emerging attention has been paid to subspace clustering. Traditional clustering algorithms, such as *K*-means and expectation maximization (EM) based algorithms, mostly use linear associations or geometric proximity to measure the similarity/distance between data points [8].

When applying traditional clustering algorithms to the domain of bioinformatics, additional challenges are faced due to prevalent existence of nonlinear correlations in the high dimensional space [9]. However, nonlinear correlations are

largely untouched in contrast to the relative mature literature of clustering using linear correlations [5, 10–12]. There are several factors making nonlinear clustering difficult. First, a pair of nonlinearly associated data points may not be close to each other in high-dimensional space. Second, it is difficult to effectively define a cluster profile (i.e., the "center" of a cluster) to summarize a cluster given the existence of nonlinear associations. Third, compared to measures that detect linear correlations, nonlinear association measures lose statistical power more quickly with the increase of random additive noise. Fourth, given the high dimensions, computationally expensive methods, for example, principal curves [13, 14], are hard to be adopted even though they can model nonlinear relationships.

In this paper, we try to address these problems by developing a clustering method that can group data points with both linear and nonlinear associations. We name this method "K-profiles clustering." Our method is based on the previously described nonlinear measure: the Distance Based on Conditional Ordered List (DCOL) [15, 16]. The key concept is to use data point orders in the sample space as the cluster profile. We have previously described a hierarchical clustering scheme named General Dependency Hierarchical Clustering (GDHC). However the computation of GDHC is very intensive. The new K-profiles clustering method is much more efficient, representing a ~20-fold reduction in computing time. Conceptually, it is the nonlinear counterpart of the popular K-means clustering method, while the existing GDHC is the nonlinear counterpart of the traditional hierarchical clustering method. Another key advantage of the K-profiles clustering method is that, by building statistical inference into the iterations, noise genes that do not belong to any cluster will not interfere with the cluster profile estimation, and they are naturally left out of the final results.

2. Methods

2.1. Distance Based on Conditional Ordered List (DCOL). We first consider the definition of Distance Based on Conditional Ordered List (DCOL) in two-dimensional space. Given two random variables X and Y and the corresponding data points $\{(x_i, y_i)\}_{i=1,\dots,n}$, after sorting the points on x-axis to obtain

$$\left\{\left(x_i^*, y_i^*\right) : x_1^* \le x_2^* \le \cdots \le x_n^*\right\} \tag{1}$$

the DCOL is defined as

$$d_{\mathrm{col}}(Y \mid X) = \frac{1}{(n-1)} \sum_{i=2}^{n} \left|y_i - y_{i-1}\right|. \tag{2}$$

Intuitively, when Y is less spread in the order sorted on X, $d_{\mathrm{col}}(Y \mid X)$ is small. We can use $d_{\mathrm{col}}(Y \mid X)$ to measure the spread of conditional distribution $Y \mid X$ in a nonparametric manner [16].

The statistical inference on $d_{\mathrm{col}}(Y \mid X)$ can be conducted using a permutation test. Under the null hypothesis that X and Y are independent of each other, the ordering of the data points based on X is simply a random reordering of Y. Thus we can randomly permute $\{(y_i)\}_{i=1,\dots,n}$ B times and calculate the sum of distances between adjacent Y values in each permutation. Then we can find the mean and standard deviation from the B values sampled from the null distribution. The actual $d_{\mathrm{col}}(Y \mid X)$ can then be compared to the estimated null distribution to obtain the p value. Notice this process does not depend on X. The permutation can be done once for Y and the resulting null distribution parameters apply to any X, which greatly saves computing time.

2.2. Defining a Cluster Profile and Generalizing DCOL to Higher Dimensions. Let $\mathbf{U}$ be a p-dimensional random vector $(X_1, X_2, \dots, X_p)$, where each X_i is a random variable; then an instance of random vector $\mathbf{U}$ can be seen as a point in the p-dimensional space. Assuming instances of random vector $\mathbf{U}$ are sorted in the p-dimensional space, then $d_{\mathrm{col}}(Y \mid \mathbf{U})$ can be computed according to (2) for any random variable Y. Therefore, the key problem is to define the order of a series of p-dimensional points.

When X is one-dimensional, we can easily prove that a list of numbers $(x_1, x_2, \dots, x_n)$ is sorted if and only if $\sum_{i=2}^{n} |x_i - x_{i+1}|$ is minimized. We generalize this to p-dimensional space and define instances $(\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n)$ as sorted if and only if the sum of distances between the adjacent p-dimensional points is minimized. Sorting the points is equivalent to finding the shortest Hamiltonian path through the n points in p dimensions, the solution of which is linked to the Traveling Salesman Problem (TSP) [17]. Many methods exist for solving the TSP [17].

If we consider the p random variables as p genes, we have effectively defined a profile for the cluster made of these p genes. Using this profile, we can compute the $d_{\mathrm{col}}(Y \mid \mathbf{U})$ for any gene Y and determine if the gene is close to this cluster, which serves as the foundation of the K-profile algorithm.

2.3. The K-Profiles Algorithm. In this section, we outline the DCOL-based nonlinear K-profiles clustering algorithm. First, we define the gene expression data matrix $G_{p\times n}$, where n samples are measured for p genes and each cell g_{ij} is the measured expression level of gene i on sample j. Each row represents the expression pattern of a gene while each column represents the expression profile of a specified sample.

The K-profiles clustering process is analogous to the traditional K-means algorithm overall. However there are two key differences: (1) Different from the K-means clustering algorithm, we use the data point ordering (Hamiltonian path) as the cluster profile rather than the mean vector of all data points belonging to this cluster; (2) during the iterations, the association of each point to its closest cluster is judged for statistical significance. Points that are not significantly associated with any cluster cannot contribute to the estimation of the cluster's profile.

Due to the random initialization of clusters, we use a loose p value cutoff at the beginning and decrease it iteration by iteration as the updated cluster profiles become more stable and reflect the authentic clusters more reliably as the clustering process progresses.

(a) To start, we compute the null distribution of DCOL distances for each gene (row) and obtain two parameters, mean μ_i and standard deviation σ_i, for each gene

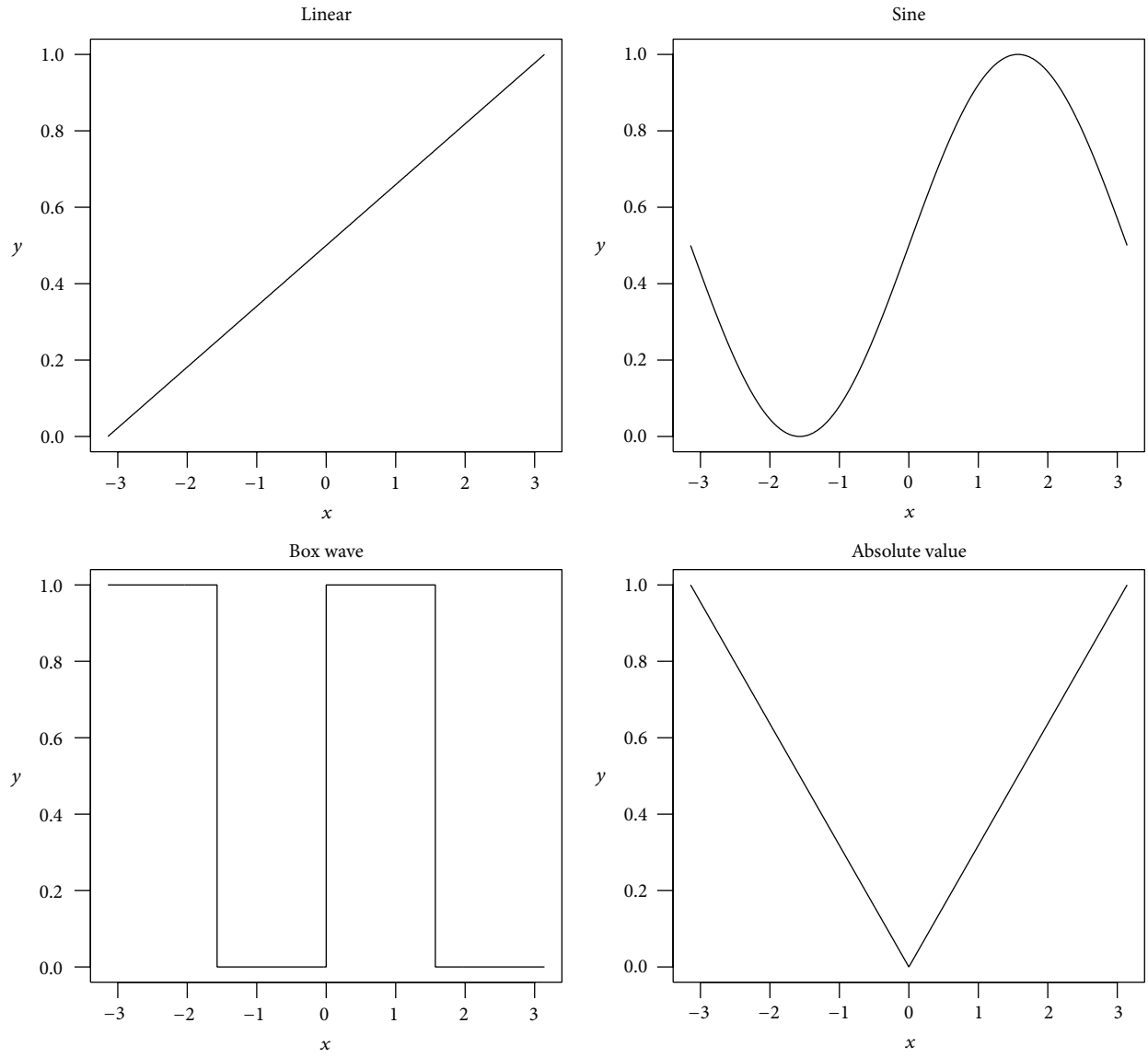


Figure 1: Illustration of the four functions used in simulations.

simultaneously by permuting columns of the matrix 500 times. The gene-specific null distribution parameters are used to compute the p values of the DCOL whenever assigning a gene to the closest cluster.

(b) Initialize K clusters by generating K random orders as cluster profiles; set p value cutoff to upper bound.

(c) For each row vector, compute its DCOL distance to each cluster according to corresponding cluster profile $d_{\text{col}}(X_i \mid \mathbf{U}_k)$, where X_i is the ith gene and $\mathbf{U}_k$ is the kth cluster. Assign it to the closest cluster if the DCOL is statistically significant in terms of p value. In this step, we are implicitly computing K p values for each gene and taking the minimum. Thus we need to adjust the p value cutoff to address the multiple testing issue. We assume each cluster profile is independent of the others. Then it follows that, for each gene, the K p values are independent. Under the null hypothesis that the gene is not associated with any of the clusters, all the p values are *i.i.d.* samples from the standard uniform distribution. Thus the nominal p value cutoff π is transformed to $\pi' = 1 - (1 - \pi)^{1/K}$.

(d) When all gene vectors have been assigned, recalculate the profile of each cluster using a TSP solver.

(e) Repeat steps (c) and (d) until the cluster profiles no longer change or the maximum iteration is reached. We start with a loose p value cutoff. In each iteration we reduce the p value cutoff by a small amount, until the target p value cutoff is reached.

The above procedure is conditioned on a given K, the number of clusters. We used gap statistics for determination of K. Other options such as prediction strength or finding

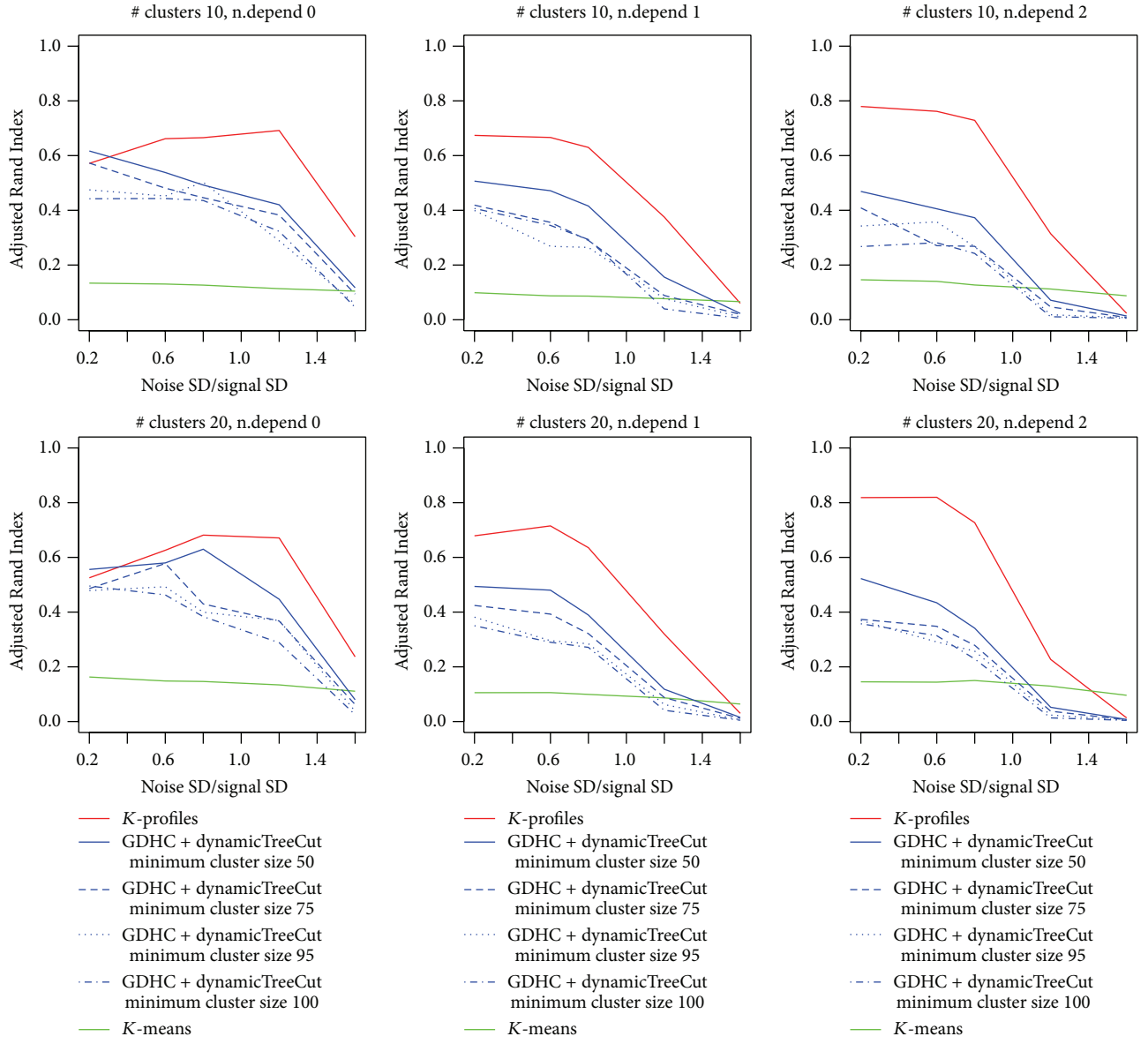


Figure 2: Simulation results with nonlinear data.

the elbow of the variance-cluster number plot are also available. Here we replace the sum of variances by the sum of negative log p values.

2.4. Simulation Study. We generated simulation datasets with 100 samples (columns) and M gene clusters, each containing 100 genes (rows). Another 100 pure noise genes were added to the data. M was set to 10 or 20 in separate simulation scenarios. Within each cluster, we set the genes (rows) to be either linearly or nonlinearly correlated using different link functions, including (1) linear, (2) sine curve, (3) box wave, and (4) absolute value (Figure 1).

Clusters were generated separately using three different mechanisms, namely, (1) the hidden factor data generation approach, (2) 1-dependent approach, and (3) 2-dependent approach.

In the hidden factor approach, for each cluster, we first generated the expression levels of a single controlling factor z by sampling the standard normal distribution. Then for each gene, a function was randomly drawn from the four functions mentioned above (Figure 1). The gene was generated as the function of the hidden controlling factor plus certain level of noise from the normal distribution: $x^{(\text{new})} = f(z) + \varepsilon$.

In the 1-dependent approach, the expressions of genes in a cluster were generated sequentially. The first gene was generated by sampling the standard normal distribution. From the second gene on, we first randomly chose one gene that was already generated and randomly chose one function from the four available functions (Figure 1). We then generated the new gene as the function of the previously generated gene: $x^{(\text{new})} = f(x^{(\text{selected})})$. After the expression of all genes in a cluster was generated, certain level of noise

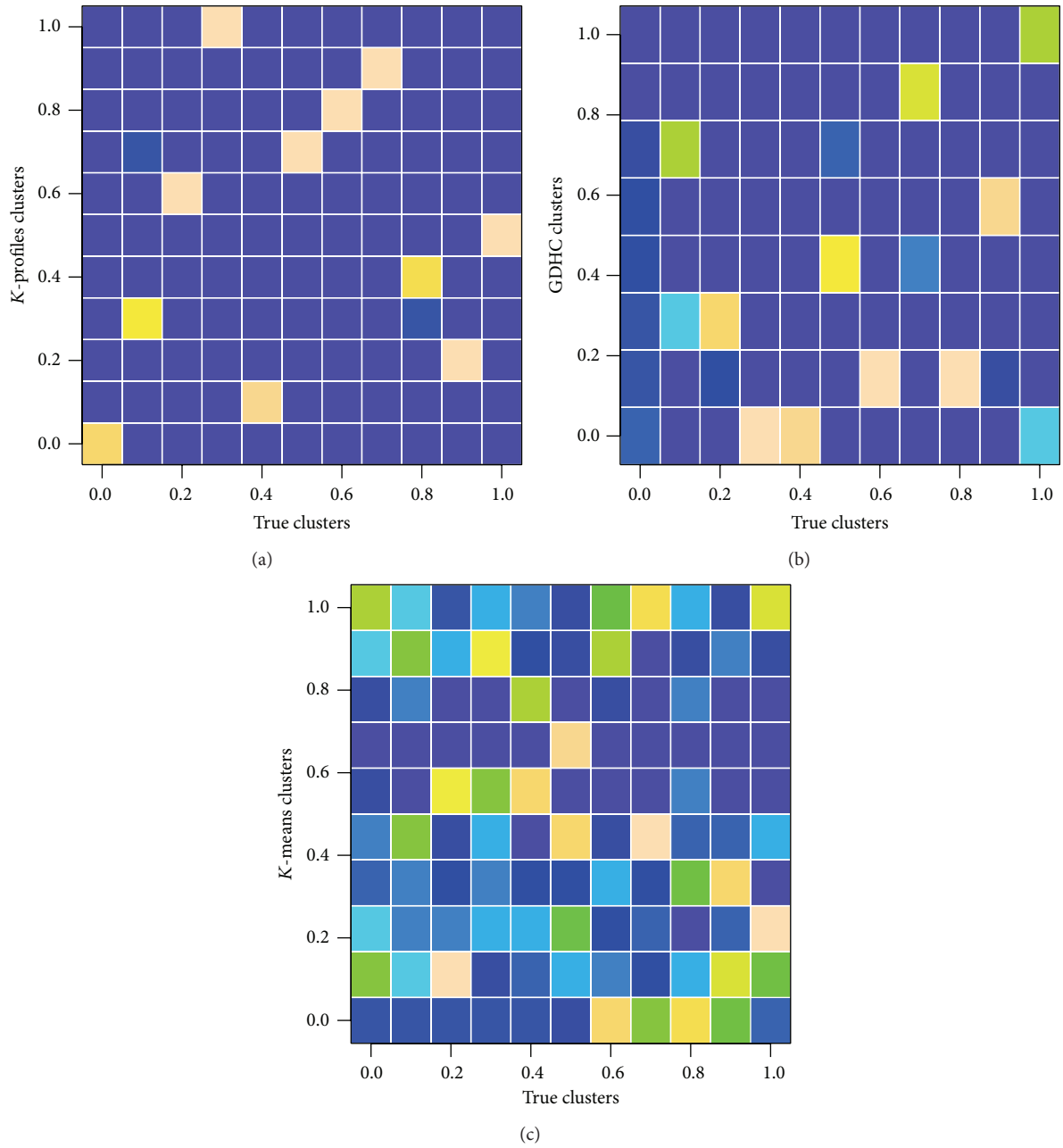


FIGURE 3: An example of confusion matrices shown as images. Cleaner pictures indicate better agreement between true clusters and clustering results. The left-most column of each subplot represents the pure noise gene group. (a) K-profiles clustering result. (b) GDHC result. (c) K-means result.

was generated from the normal distribution and added to the gene expression profiles.

The 2-dependent approach is similar to the 1-dependent approach. The difference is that, for each new gene, two previously generated genes were randomly selected, and two functions were randomly chosen. The new gene was generated as the summation: $x^{(\text{new})} = \beta_1 f(x^{(\text{selected\_1})}) + \beta_2 g(x^{(\text{selected\_2})})$. The β's were sampled from the uniform distribution between -1 and 1. Again certain level of noise was generated from the normal distribution and added to the gene expression profiles.

3. Results and Discussions

3.1. Simulation Results. In the simulation experiments, we compared the K-profiles algorithm with General Dependency Hierarchical Clustering (GDHC) and the traditional K-means clustering algorithm. The GDHC was paired with

the dynamic tree cutting method to cut the trees into clusters [18]. We used the efficient TSP R library to compute the cluster profiles [19]. We adopted the external evaluation metric Adjusted Rand Index (ARI) [20] to compare the clustering results with the true cluster memberships to judge the performance of the methods.

In Figure 2, the average ARI values were plotted against the noise level. Higher ARI values indicate better clustering performance. The figure contains three columns and two rows with each column representing a data generation mechanism and each row representing a different number of clusters. In the left column, data was generated by the hidden factor mechanism, where all features in a true cluster were linearly/nonlinearly linked to a latent factor. In columns 2 and 3, features in each cluster were generated using 1-dependent and 2-dependent mechanism, respectively. In such a generation mechanism, genes generated later depend on previously generated genes in the same cluster [15]. In the meantime, the first row shows results from data with 10 clusters, while the second row shows results from data with 20 clusters.

For GDHC, we used the dynamic tree cutting method [18] to cut each tree. Various values of minimum cluster size were tested. For K-profiles clustering, we started with a p value cutoff of 0.2 and gradually reduced the cutoff to 0.05 with the iterations. We ran each setting (cluster size, data generation scheme) 20 times and plotted the average results in Figure 2. We can see obviously that both K-profiles and GDHC outperformed linear relation-based K-means clustering algorithm significantly in all cluster parameter settings. K-profiles also did a better job than GDHC in recovering the true clusters. We allowed four minimum cluster size levels in the dynamic tree cutting, 50%, 75%, 95%, and 100%, of the true cluster size. Generally the 50% setting performed the best.

Figure 3 shows the confusion matrices of an example clustering result as images. We can see the composition of the reported clusters by the three different clustering algorithms. Cleaner images indicate better agreement between true clusters and the detected clusters. When looking into all three confusion matrices, we can see that in each reported cluster our proposed method discovered a dominant group with only a little impurity. However, in traditional K-means clustering, the reported clusters were mostly composed of several small groups, which rendered it of little use when the data contains much nonlinear relations. GDHC performed much better than K-means with 4 reported clusters (rows) composed mostly of elements from the same true clusters. Clearly, the new K-profiles clustering method achieved the best performance in the simulations.

The K-profiles and GDHC clustering methods were both based on DCOL, which detects both nonlinear and linear relationships, although it has lower power to detect linear relationship compared to correlation coefficient. Next we studied how the methods behave when the true relationships are all linear. We used the same hidden factor data generation scheme but allowed only linear relations in the data generation, which means all genes in the same cluster were linearly related to the same hidden factor. We simulated data with 10 clusters, each containing 100 genes, plus an additional 100

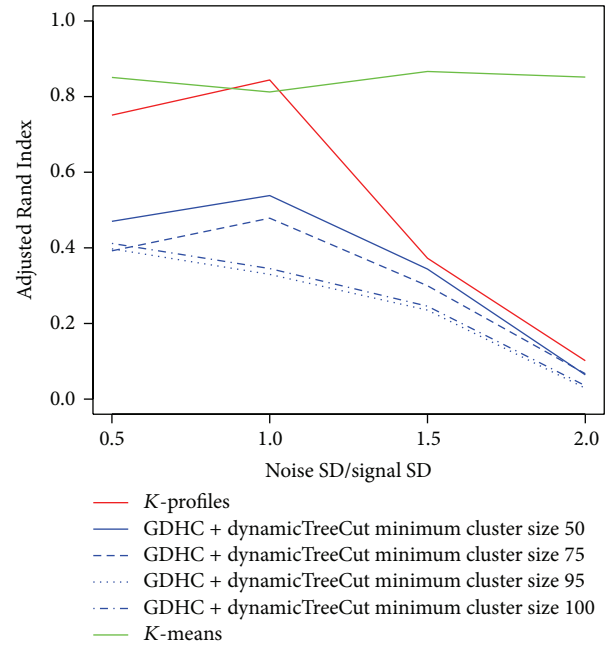


Figure 4: Simulation results from data with linear associations only.

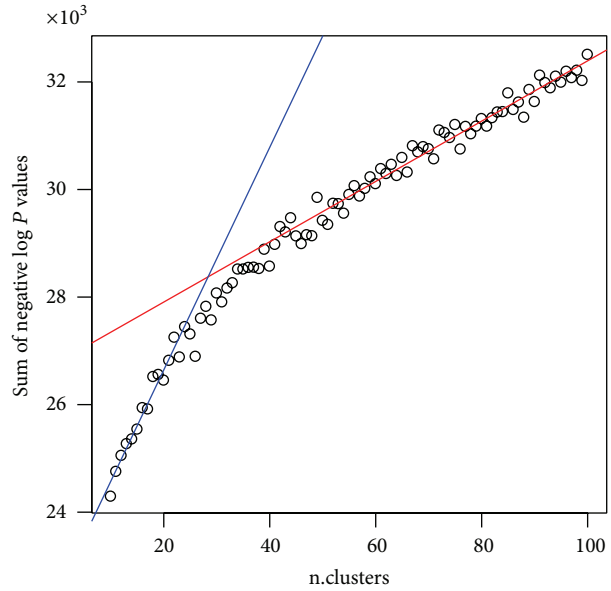


Figure 5: Selecting the number of clusters for the Spellman dataset by plotting sum of negative log p values against the number of clusters.

pure noise genes. K-profiles achieved similar performance to K-means when the noise was at low to moderate levels (Figure 4). This is likely due to the fact that K-means does not involve statistical testing to exclude noise genes from the clusters.

Besides being a more effective nonlinear clustering method, the K-profiles method is also more efficient compared to GDHC. On a data matrix with 2000 rows and 100 columns, the average computing time of K-profiles was ~30

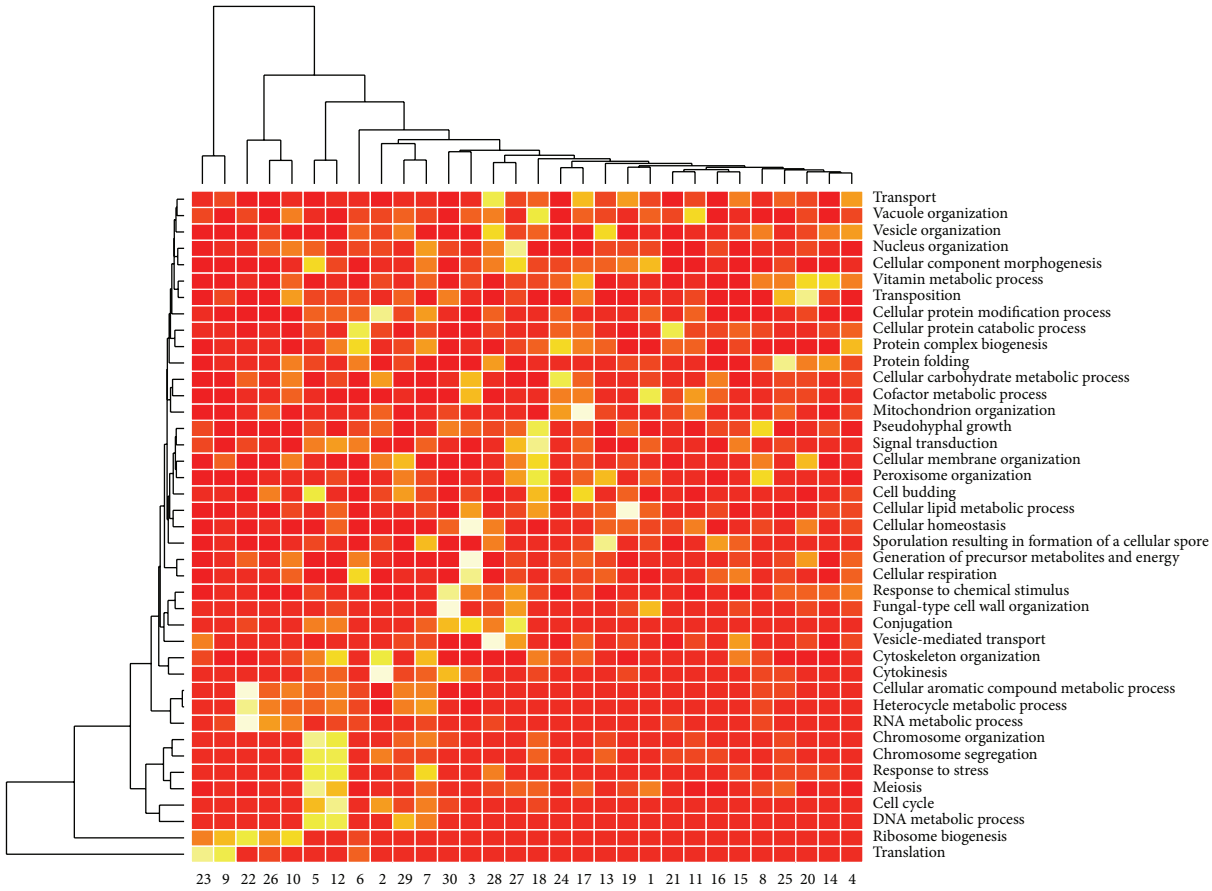


FIGURE 6: Significance levels of GO slims terms. Brighter colors indicate significance using the hypergeometric test for overrepresentation analysis.

seconds on a laptop with i7-3537U CPU and 6Gb memory, while the GDHC used ~600 seconds.

3.2. Real Data Analysis. We conducted data analysis on the Spellman yeast cell cycle data, which consists of four time series synchronized by different chemical reagents, each covering roughly two cell cycles [21]. One of the time series, the cdc15 data, contains a strong oscillating signal [22]. We removed the cdc15 dataset and used the data of the three remaining time series. The data matrix consists of 49 samples (columns) and 6178 genes (rows).

We applied the K-profiles clustering method using a series of K values. With each K value, we retained the final p value p_i of every gene. We then took the negative sum of log p values $\sum_i -\log(p_i)$ at every K and plotted the value against K. An elbow was observed at around 30 (Figure 5). Thus we chose $K = 30$ for subsequent analyses.

Among the 6178 genes under study, 4874 were clustered into 30 clusters. The minimum cluster size was 59, and the maximum cluster size was 328. We then judged the performance of the methods using functional annotations. For this purpose, we resorted to Gene Ontology [23]. We used a set of GO terms that categorize genes into broad functional categories, the GO slim terms from the Saccharomyces Genome Database (SGD) [24]. Some of the GO slim terms are too broad; we limited our analysis to terms with 2000 annotated genes or less. We found that almost all the clusters are associated with certain GO slim terms using the hypergeometric test [25] for overrepresentation (Figure 6).

From Figure 6, we see clearly that several clusters, including clusters 2, 5, 7, and 12, are highly associated with cell cycle related processes, which are clustered in the lower 1/3 region of the plot (Figure 6). We then plotted the heatmaps of the expressions of the genes in these clusters, which indeed showed strong periodical behavior. An example, cluster 2, is presented in Figure 7. We notice the genes in this cluster were mostly periodic genes, yet they exhibit different phase shifts. Such genes may not be clustered together using traditional methods based on linear associations.

The GO slim terms are broad functional categories and do not offer enough detail. We further analyzed the data using a set of 430 selected representative GO terms. The approaches to select these terms were previously described in [26, 27]. Essentially the selected terms were relatively specific, yet they were still of reasonable size. We conducted hypergeometric test for overrepresentation of these GO

Table 1: Biological pathways significantly associated with clusters 2, 5, 7, and 12.

Cluster	# genes	GO Biological Process ID[#]	P value[*]	Name of GO term
2	228	GO:0051301	$1.03E-07$	Cell division
		GO:0006468	0.0001307	Protein phosphorylation
		GO:0010696	0.00163665	Positive regulation of spindle pole body separation
		GO:0030473	0.00584256	Nuclear migration along microtubule
		GO:0005977	0.00628021	Glycogen metabolic process
5	116	GO:0006301	$5.94E-06$	Postreplication repair
		GO:0043570	$1.87E-05$	Maintenance of DNA repeat elements
		GO:0006272	$4.90E-05$	Leading strand elongation
		GO:0000070	0.00043025	Mitotic sister chromatid segregation
		GO:0009263	0.00067342	Deoxyribonucleotide biosynthetic process
		GO:0006298	0.00074914	Mismatch repair
		GO:0007131	0.00077629	Reciprocal meiotic recombination
		GO:0045132	0.00300391	Meiotic chromosome segregation
		GO:0006284	0.0034725	Base-excision repair
		GO:0006273	0.0041114	Lagging strand elongation
		GO:0006348	0.00415626	Chromatin silencing at telomere
		GO:0009200	0.00485315	Deoxyribonucleoside triphosphate metabolic process
		GO:0051301	0.00750912	Cell division
7	69	GO:0006334	$4.57E-12$	Nucleosome assembly
		GO:0030473	$6.32E-05$	Nuclear migration along microtubule
		GO:0030148	0.00299059	Sphingolipid biosynthetic process
		GO:0000032	0.00650292	Cell wall mannoprotein biosynthetic process
		GO:0009225	0.00774684	Nucleotide-sugar metabolic process
12	155	GO:0007020	$1.07E-05$	Microtubule nucleation
		GO:0000070	0.0006474	Mitotic sister chromatid segregation
		GO:0006284	0.00078868	Base-excision repair
		GO:0006493	0.00078868	Protein O-linked glycosylation
		GO:0006273	0.00099378	Lagging strand elongation
		GO:0006337	0.00099378	Nucleosome disassembly
		GO:0000724	0.00151593	Double-strand break repair via homologous recombination
		GO:0000086	0.00242563	G2/M transition of mitotic cell cycle
		GO:0006368	0.00243303	Transcription elongation from RNA polymerase II promoter
		GO:0006338	0.0038366	Chromatin remodeling
		GO:0008156	0.00743106	Negative regulation of DNA replication

[#]Total number of GO Biological Process terms under study: 430.
[*]P value threshold: 0.01.

terms in each of the 30 clusters. We found almost all the clusters significantly overrepresent some biological processes. As examples, we show biological processes associated with clusters 2, 5, 7, and 12, which are clearly cell cycle related based on the GO slim analysis (Table 1). Many clusters clearly showed no periodical behavior. They were strongly associated with functional categories such as metabolism and signal transduction. The results are listed online at http://web1.sph.emory.edu/users/tyu8/KPC.

4. Conclusion

In this paper, we described a new nonlinear clustering method named K-profiles clustering. We incorporated statistical inference into the algorithm to remove the impact of noise genes due to their common existence in real world microarray data. The algorithm is efficient due to the quality of the Distance Based on Conditional Ordered List (DCOL). The algorithm outperformed our previous General Dependency Hierarchical Clustering (GDHC) algorithm and the traditional K-means clustering algorithm in our simulation studies. It generated meaningful results in real data analysis. It can be used in the analysis of high-throughput data to detect novel patterns based on nonlinear dependencies.

Conflict of Interests

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

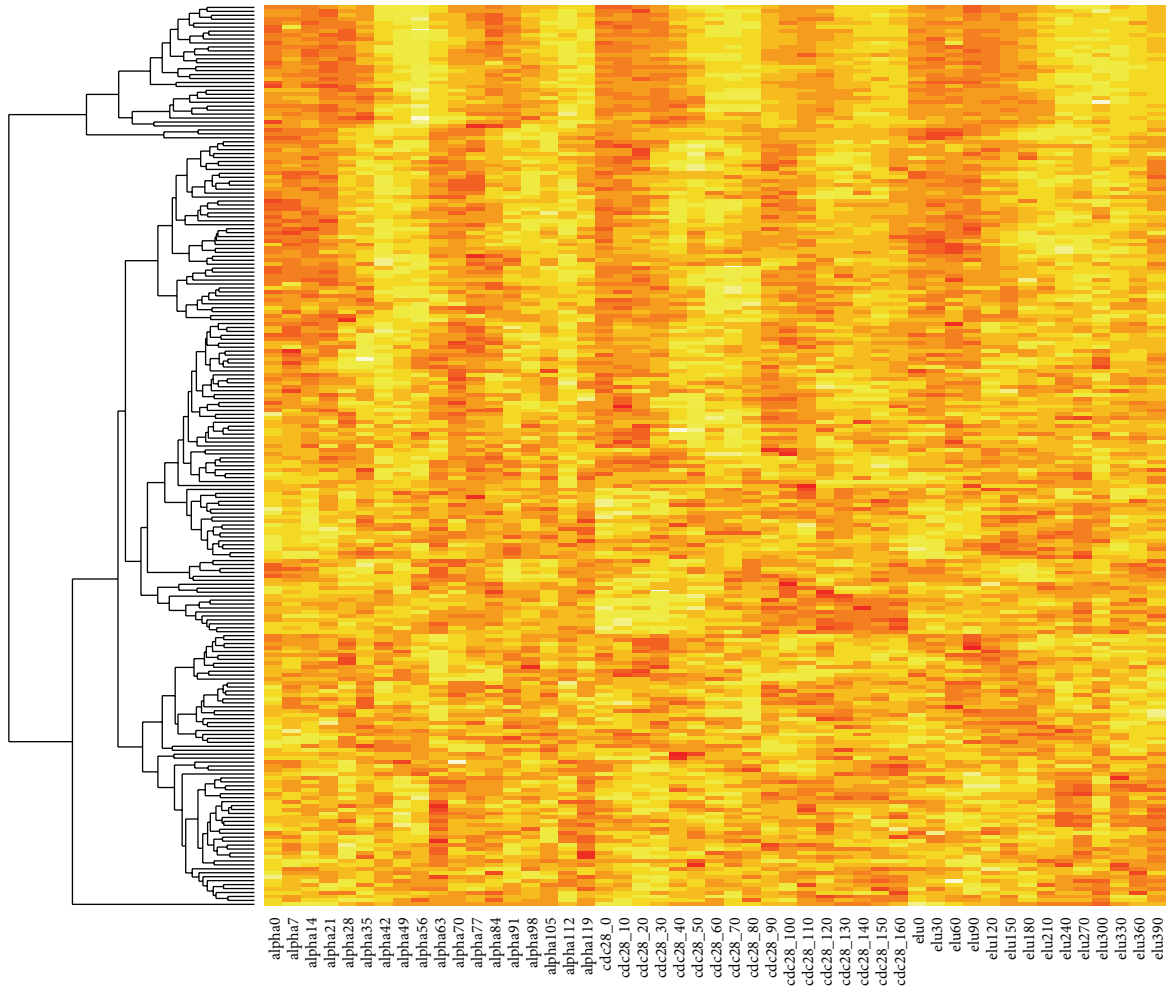


Figure 7: An example cluster with mostly periodically expressed genes.

Acknowledgments

This work was partially supported by NIH Grants P20HL113451 and U19AI090023, 973 Program (no. 2013CB967101) of the Ministry of Science and Technology of China, and Shanghai Science Committee Foundation (13PJ1433200).

References

[1] J. Rung and A. Brazma, "Reuse of public genome-wide gene expression data," *Nature Reviews Genetics*, vol. 14, no. 2, pp. 89–99, 2013.

[2] G. J. Patti, O. Yanes, and G. Siuzdak, "Innovation: metabolomics: the apogee of the omics trilogy," *Nature Reviews Molecular Cell Biology*, vol. 13, no. 4, pp. 263–269, 2012.

[3] T. Yu, "An exploratory data analysis method to reveal modular latent structures in high-throughput data," *BMC Bioinformatics*, vol. 11, article 440, 2010.

[4] Y. Zhao, J. Kang, and T. Yu, "A Bayesian nonparametric mixture model for selecting genes and gene subnetworks," *The Annals of Applied Statistics*, vol. 8, no. 2, pp. 999–1021, 2014.

[5] A. K. H. Tung, X. Xu, and B. C. Ooi, "CURLER: finding and visualizing nonlinear correlation clusters," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 467–478, ACM, New York, NY, USA, June 2005.

[6] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1370–1386, 2004.

[7] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer, New York, NY, USA, 2nd edition, 2009.

[8] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability, SIAM American Statistical Association, Philadelphia, Pa, USA, 2007.

[9] K.-C. Li, C.-T. Liu, W. Sun, S. Yuan, and T. Yu, "A system for enhancing genome-wide coexpression dynamics study," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 44, pp. 15561–15566, 2004.

[10] H. K. Solvang, O. C. Lingjærde, A. Frigessi, A.-L. Børresen-Dale, and V. N. Kristensen, "Linear and non-linear dependencies between copy number aberrations and mRNA expression

reveal distinct molecular pathways in breast cancer," *BMC Bioinformatics*, vol. 12, article 197, 2011.

[11] A. Jian, Z. Zhang, and E. Chang, "Adaptive non-linear clustering in data streams," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06)*, pp. 122–131, Arlington, Va, USA, 2006.

[12] M. Ehler, V. N. Rajapakse, B. R. Zeeberg et al., "Nonlinear gene cluster analysis with labeling for microarray gene expression data in organ development," *BMC Proceedings*, vol. 5, supplement 2, article S3, 2011.

[13] B. Kegl, A. Krzyzak, T. Linder, and K. Zeger, "Learning and design of principal curves," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 3, pp. 281–297, 2000.

[14] P. Delicado and M. Smrekar, "Measuring non-linear dependence for two random variables distributed along a curve," *Statistics & Computing*, vol. 19, no. 3, pp. 255–269, 2009.

[15] T. Yu and H. Peng, "Hierarchical clustering of high-throughput expression data based on general dependences," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 4, pp. 1080–1085, 2013.

[16] T. Yu, H. Peng, and W. Sun, "Incorporating nonlinear relationships in microarray missing value imputation," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 3, pp. 723–731, 2011.

[17] G. Gutin and A. P. Punnen, *The Traveling Salesman Problem and Its Variations*, Combinatorial Optimization, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

[18] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R," *Bioinformatics*, vol. 24, no. 5, pp. 719–720, 2008.

[19] D. Applegate, W. Cook, and A. Rohe, "Chained Lin-Kernighan for large traveling salesman problems," *INFORMS Journal on Computing*, vol. 15, no. 1, pp. 82–92, 2003.

[20] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[21] P. T. Spellman, G. Sherlock, M. Q. Zhang et al., "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[22] K.-C. Li, M. Yan, and S. Yuan, "A simple statistical model for depicting the cdc15-synchronized yeast cell-cycle regulated gene expression data," *Statistica Sinica*, vol. 12, no. 1, pp. 141–158, 2002.

[23] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[24] J. M. Cherry, E. L. Hong, C. Amundsen et al., "Saccharomyces Genome Database: the genomics resource of budding yeast," *Nucleic Acids Research*, vol. 40, no. 1, pp. D700–D705, 2012.

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

[26] T. Yu and Y. Bai, "Improving gene expression data interpretation by finding latent factors that co-regulate gene modules with clinical factors," *BMC Genomics*, vol. 12, article 563, 2011.

[27] T. Yu and Y. Bai, "Capturing changes in gene expression dynamics by gene set differential coordination analysis," *Genomics*, vol. 98, no. 6, pp. 469–477, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 657159, 8 pages
http://dx.doi.org/10.1155/2015/657159

Research Article

Screening Ingredients from Herbs against Pregnane X Receptor in the Study of Inductive Herb-Drug Interactions: Combining Pharmacophore and Docking-Based Rank Aggregation

Zhijie Cui,[1] Hong Kang,[1] Kailin Tang,[1] Qi Liu,[1] Zhiwei Cao,[1,2] and Ruixin Zhu[1,3]

[1]*Department of Bioinformatics, Tongji University, Shanghai 200092, China*
[2]*Shanghai Center for Bioinformation Technology, Shanghai, China*
[3]*School of Pharmacy, Liaoning University of Traditional Chinese Medicine, Dalian, Liaoning, China*

Correspondence should be addressed to Zhiwei Cao; zwcao@tongji.edu.cn and Ruixin Zhu; rxzhu@tongji.edu.cn

Received 4 September 2014; Revised 22 December 2014; Accepted 27 December 2014

Academic Editor: Feng Luo

The issue of herb-drug interactions has been widely reported. Herbal ingredients can activate nuclear receptors and further induce the gene expression alteration of drug-metabolizing enzyme and/or transporter. Therefore, the herb-drug interaction will happen when the herbs and drugs are coadministered. This kind of interaction is called inductive herb-drug interactions. Pregnane X Receptor (PXR) and drug-metabolizing target genes are involved in most of inductive herb-drug interactions. To predict this kind of herb-drug interaction, the protocol could be simplified to only screen agonists of PXR from herbs because the relations of drugs with their metabolizing enzymes are well studied. Here, a combinational in silico strategy of pharmacophore modelling and docking-based rank aggregation (DRA) was employed to identify PXR's agonists. Firstly, 305 ingredients were screened out from 820 ingredients as candidate agonists of PXR with our pharmacophore model. Secondly, DRA was used to rerank the result of pharmacophore filtering. To validate our prediction, a curated herb-drug interaction database was built, which recorded 380 herb-drug interactions. Finally, among the top 10 herb ingredients from the ranking list, 6 ingredients were reported to involve in herb-drug interactions. The accuracy of our method is higher than other traditional methods. The strategy could be extended to studies on other inductive herb-drug interactions.

1. Background

In America, nearly forty percent of adults consume herbs or herbal products regularly every year and this number is still increasing [1]. One-sixth of them take herbal supplements together with prescribed drugs [2]. However, most of them do not realize that they are at the risk of potential adverse herb-drug interactions [3]. In order to avoid the medicine interactions as much as possible, it is urgent to discover the underlying herb-drug interactions.

Herb-drug interactions, as well as drug-drug interactions (DDIs), are generally divided into two categories: pharmacodynamics (PD) interactions and pharmacokinetic (PK) interactions [4]. Many previous studies contributed to the explanation of molecular basis for drug interactions [5, 6]. In the late 1990s, it was found that ligand-activated nuclear receptors can regulate drug metabolism and transporter genes expression [7–9]. Nuclear receptors play an important role in the mechanism of PK interactions [10]. Based on that molecular mechanism (shown in Figure 1), herbal ingredients (agent A) can activate nuclear receptors and regulate metabolizing drugs (agent B) gene expression. Thus, the herbs could alter efficacy and toxicity of coadministered drugs. This process is called inductive herb-drug interaction [7, 11].

Pregnane X Receptor (PXR), as a member of nuclear receptor families, is involved in most of inductive herb-drug interactions through regulating drug-metabolizing gene expression [12, 13]. To predict the inductive drug interaction involving PXR, identifying ligands of PXR and drug-metabolizing enzyme/transporter could be done, respectively. However, because the relations of drugs and their metabolizing enzymes are well known, the key step

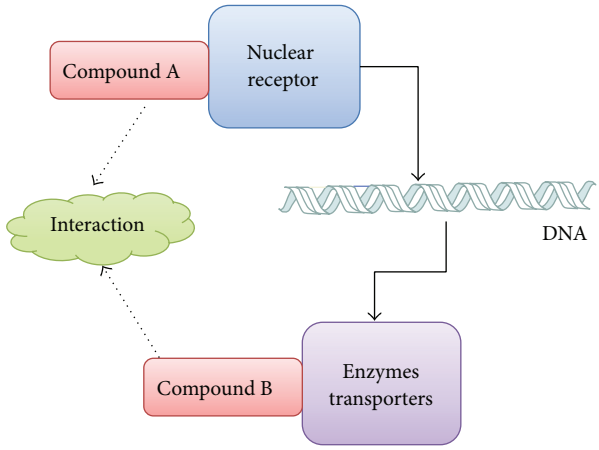


Figure 1: The mode of inductive drug interactions.

of prediction would be simplified to only screen agonists of PXR. Several experimental systems *in vitro* have been developed for identifying agonist of nuclear receptors [14], such as cultured primary human hepatocytes and liver slices, humanized mouse models, transformed hepatocytes or cell lines, reporter gene assays, coactivator recruitment assays, and receptor binding assays [15–17]. But these experimental systems are low-efficiency and high-cost process to screen numerous molecules. Therefore, high-throughput and low-cost method for screening agonists of PXR is needed. Computational technique is just a good complementary for experimental systems.

In the past years, several computational methods have been used for virtual screening PXR's agonists, such as structure-based docking [18–20], ligand-based QSAR [21, 22], machine learning [18, 23], and pharmacophore model [24, 25]. Due to the large and flexible binding site of PXR [26], broad specificity of ligands, and the insufficient activity data [27], a comprehensive in silico strategy with both qualitative and quantitative analysis could be expected. The aim of our study is to propose a combined method of pharmacophore modelling, docking-based rank aggregation (DRA) for screening agonists of PXR. The method can provide aid for predicting inductive herb-drug interactions involving PXR. Also, it is applicable to predicting more herb-drug interactions involving other nuclear receptors.

2. Materials and Methods

2.1. Dataset. The complex crystal structure provides the binding information objectively, which is used for pharmacophore modelling and molecular docking. Three complex structures of PXR were obtained from the Protein Data Bank (PDB http://www.rcsb.org/pdb/home/home.do) [28], including 1NRL [29], 1ILH [26], and 3HVL [30].

266 compounds with EC_{50} values were obtained from the Binding Database (BindingDB http://www.bindingdb.org/bind/index.jsp) [31], which were selected as testing data for the pharmacophore modelling experiment. 71 compounds were labelled as active ligands ($EC_{50} \leq 10\,\mu M$) and 195 compounds were labelled as inactive ligands ($EC_{50} \geq 10\,\mu M$). In these 266 compounds, EC_{50} values of 107 compounds are numeric so that these compounds can be ranked by EC_{50} values (see Supplementary Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/657159). The EC_{50}-based ranking list ($Rank_{EC}$) including 107 compounds was regarded as a reference list in the rank aggregation experiment.

In order to evaluate performance of our method, a dataset of herb-drug interactions was needed. 421 herbs were checked in the PubMed database by text mining method. 90 herbs were found to interact with 230 drugs forming 380 herb-drug interactions. Besides, molecular structures of herbal ingredients should be provided for pharmacophore modelling and molecular docking. Among 421 herbs, 820 ingredients structures were obtained from the PubChem database (http://pubchem.ncbi.nlm.nih.gov/).

2.2. Methods

2.2.1. Pharmacophore Modelling. As shown in Figure 3, three different conformations of SRL12813 were, respectively, extracted from complex crystal structures of PXR (PDB id: 1NRL [29], 1ILH [26], and 3HVL [30]). The red conformation of SRL12813 was extracted from complex 3HVL; the yellow one was extracted from complex 1ILH; the blue one was extracted from complex 1NRL. They were provided as template molecules. Pharmacophore was generated by collecting a common set of template molecules structural features. These structural features are related to the ligand's biological activity and recognition at binding site of receptor. In our model, five pharmacophoric structural features (shown in Figure 2) were fit by all template molecules. The process of pharmacophore modelling was performed in Molecular Operation Environment (MOE) 2008.10.

2.2.2. Docking-Based Rank Aggregation (DRA). Docking-based rank aggregation (DRA) is a two-step process. Firstly, candidate ligands filtered out by the former pharmacophore model were docked into PXR with four different energy scoring functions. The possibility of candidate ligands was ranked according their energy scores. Secondly, four different ranks from four scoring functions were aggregated to obtain a final rank.

The complex crystal structure of PXR (PDB id: 1ILH) was used to define the active site and dock with other molecules. Molecular docking was performed in MOE-Dock 2008.10. The way to place ligand was alpha sphere triangle matching with 4 different scoring functions (ASE Scoring, Affinity dG Scoring, Alpha HB Scoring, and London dG Scoring), respectively. The molecular mechanics force field was used to minimize energy of the system. 0.0001 kcal/(mol·Å) was chosen as the cutoff of the root-mean-squared gradient and maximum iterations was 1000 with their defaulted parameters. Finally, four ranked lists ($Rank_{AS}$, $Rank_{AF}$, $Rank_{AL}$, and $Rank_{LO}$) were calculated by 4 individual scoring functions.

Rank aggregation is a kind of multiview data analysis strategy aiming to fuse ranking results derived from individual views [32]. A final rank with views as comprehensive as

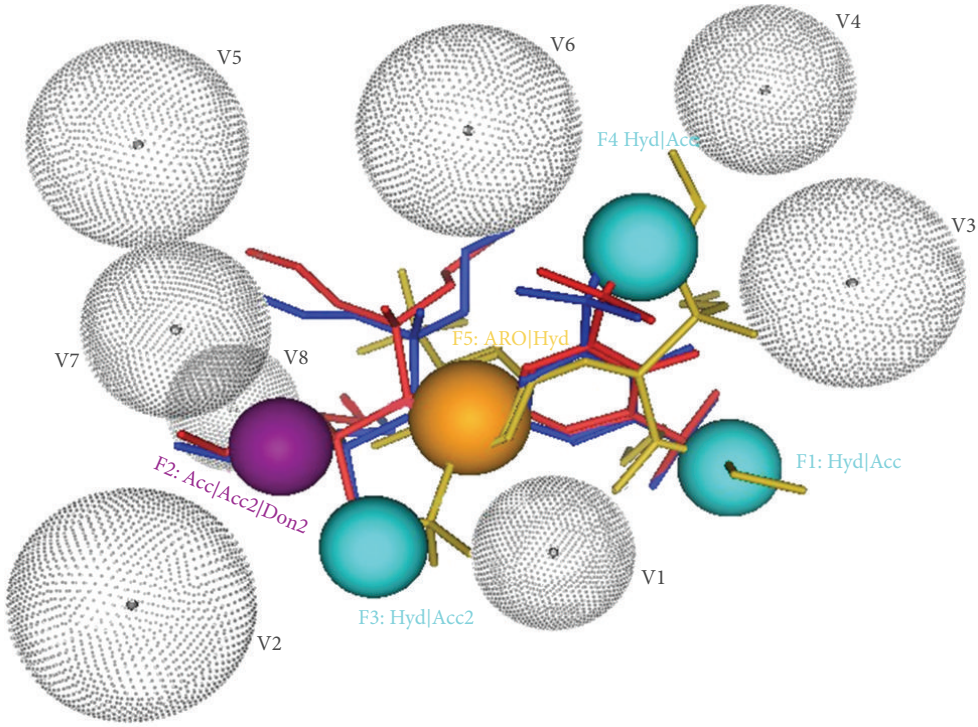


Figure 2: The pharmacophore of PXR (F1: Hyd|Acc; F2: Acc|Acc2|Don2; F3: Hyd|Acc2; F4: Hyd|Acc; F5: ARO|Hyd; V1–V8: excluded volume).

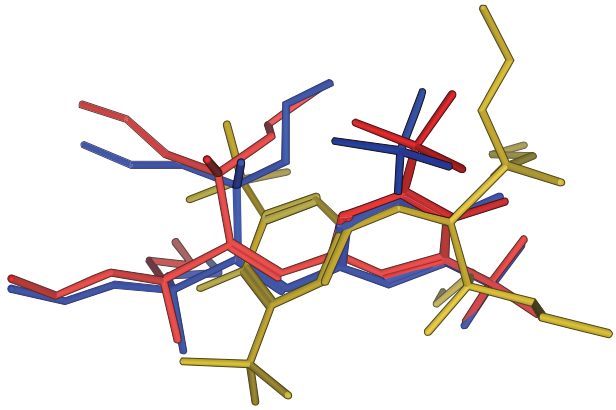


Figure 3: The molecular structure of template by superposing three SRL12813 in three different conformations.

possible, which is expected to better reflect the real rank, is worked out by aggregation of ranks from individual views [33].

Some concepts and details which are used in the process of rank aggregation are introduced below. Spearman's distance [34] is used to definition of distance between two given ranks:

$$S\left(L_i, L_j\right) = \sum_{t \in L_i \cup L_j} \left| r^{L_i}(t) - r^{L_j}(t) \right|. \quad (1)$$

Then, weighted Spearman's footrule distance between L_i and L_j is obtained via the following weighted summation representation:

$$\mathrm{WS}\left(L_i, L_j\right) = \sum_{t \in L_i \cup L_j} \left| M\left(r^{L_i}(t)\right) - M\left(r^{L_j}(t)\right) \right| \times \left| r^{L_i}(t) - r^{L_j}(t) \right|. \quad (2)$$

A function to detect the best list that is as close as possible to all the given ranks is defined as

$$\delta^* = \arg\min \emptyset(\delta), \qquad \emptyset(\delta) = \sum_{i=1}^{m} w_i d\left(\delta, L_i\right), \quad (3)$$

where w_i refers to the list L_i. d is Spearman's footrule distance between the best list δ^* and L_i. The aim of rank aggregation is to discover the e distance between the best list δ^* and L_i. The cross-entropy method was carried out to associate every two lists in our study [35].

To evaluate ranking performance in comparison with the control rank, discounted cumulative gain (DCG), a usual method to measure effectiveness of a web search engine algorithm, is used for evaluating performance of ranking. Two assumptions are acknowledged along with the use of DCG. One was that highly relevant items are more important when having higher ranks. The other is that highly relevant items are more important than irrelevant items. For a particular

rank, the discounted cumulative gain accumulated position p was defined as

$$\mathrm{DCG}_p = \mathrm{rel}_1 + \sum_{i=2}^{p} \frac{\mathrm{rel}_i}{\log_2 i}. \quad (4)$$

The rel_i is the graded relevance of the result at the position i.

Due to the variety of lists in length relying on the query, the best rank would not be achieved if DCG is used along consistently. It was necessary for normalizing the cumulative gain of each rank. The normalized DCG (nDCG) was computed as

$$\mathrm{nDCG}_p = \frac{\mathrm{DCG}_p}{\mathrm{IDCG}_p}. \quad (5)$$

Finally, the average of every nDCG of lists is used to measure the similarity of two ranks. The range of nDCG is on the interval 0 to 1.

Table 1: The value of nDCG to measure distance between ranks.

Rank	nDCG
EC_{50}	1
ABD	0.7149
AB	0.5397
D (London dG)	0.4599
ACD	0.4023
B (Affinity dG)	0.3972
BD	0.3961
AD	0.3947
BCD	0.3743
CD	0.3670
A (ASE)	0.3650
ABCD	0.3639
ABC	0.3609
AC	0.3423
C (Alpha HB)	0.3416
BC	0.3405

3. Results and Discussion

3.1. Pharmacophore Modelling. As a result of pharmacophore model, the true positive rate (sensitivity) is 53.52% (38/71) and the true negative rate (specificity) is 81.54% (159/195). The pharmacophore model of PXR is displayed in Figure 2. Remarkably, besides three different conformations of SRL12813, other agonists of PXR in the Protein Data Bank (PDB) were predicted exactly by our pharmacophore model such as RFP, HYF, PNU, and T0901317.

Two different views on how to superpose template molecules were used to construct the pharmacophore [36]. One is that the superposed conformation is gained by minimum energy [36, 37]. Yet, the other one is that the extracting conformations of ligand from its complex crystal structure are superposed directly [37, 38]. In our study, the latter method was adopted because its good performance was certified by the previous work [38].

3.2. Docking-Based Rank Aggregation (DRA). Firstly, a list including 107 ligands of PXR was sorted by EC_{50} value and was regarded as reference list, named Rank_{EC}. Secondly, 107 compounds were sorted again by calculated energy score from docking results. In molecular docking process, the binding energy score is used to evaluate binding affinity between protein receptors and ligands. It is estimated by individual scoring function. So four ranking lists of 107 compounds were generated depending on four individual scoring functions. As a result, nDCG values of these four lists were very low. It is indicated that these ranking lists from individual scoring functions were far from the reference list, Rank_{EC}. In fact, the calculated energy score is weakly correlated to experimental binding affinity because individual scoring function just one-sided reflects the true binding situation. The low correlation was verified by previous studies [39, 40]. Our result is consistent with the viewpoint (shown in Table 1).

In order to find a ranking list of ligands, which was closer to the reference list, we aggregated ranking lists derived from docking results. The aggregated result showed that Rank_{ABD}, which aggregated Rank_{AS}, Rank_{AF}, and Rank_{LO}, is the best performance in all ranking lists. The nDCG of Rank_{ABD} is 0.7149, almost twice as high as any other lists (shown in Table 1). 107 compounds in every ranking list are shown in Supplementary Table S1.

Through our aggregated lists by docking result (shown in Table 2), two points are noteworthy. (1) The way to estimate the energy of hydrogen bond in Alpha HB Scoring (C) is much similar to that of Affinity dG Scoring (B), because such two scoring functions both are dependent on the favourable rule. In Affinity dG Scoring (B), two hydroxyl groups are assumed to interact in the most favourable way, but they are also discussed in Alpha HB Scoring (C), including non-sp3 donors and acceptors, sp3 donors and acceptors, and metals in the receptor. The potential redundant content rather than complementation between the two scoring functions caused that the nDCG of Rank_{BC} was lowest. (2) It was hypothesized that a large overlap of information occurs between Alpha HB Scoring (C) and ASE Scoring (A) on account that the same content on ligand atom-alpha sphere pairs was used to evaluate the energy. Likewise, the performance of Rank_{AC} is poor. Its nDCG is the third from bottom. So it was supposed that any good aggregated rank from the 4 scoring functions must not include Alpha HB Scoring (C) and Affinity dG Scoring (B) together, or Alpha HB Scoring (C) and ASE Scoring (A). Therefore, Rank_{ABD} is the best aggregated rank. It aggregates those views that are high complementary and and low redundant with each other. This viewpoint was coincident with the previous work [41].

3.3. The Prediction of Herb-Drug Interactions. The inductive herb-drug interactions were predicted through screening agonist of PXR from herbal ingredients. Every ingredient is contained by one or more herbs. An ingredient will be considered to have the potential of inducing herb-drug interaction if a herb, containing the ingredient, is reported in our herb-drug interaction database. The ingredient is

Table 2: The description of four scoring functions.

Index	Scoring function	Description
A	ASE Scoring	The distance between all ligand atom-receptor atom pairs and ligand atom-alpha sphere pairs.
B	Affinity dG Scoring	The enthalpic contribution to the free energy of various interaction including interactions between hydrogen bond donor-acceptor pairs, ionic interactions, metal ligation, hydrophobic interactions, interactions between hydrophobic and polar atoms, and interactions between any two atoms.
C	Alpha HB Scoring	Combination of two measurements between the geometric fit of the ligand to the binding site and hydrogen bonding effects.
D	London dG Scoring	The free energy for binding of ligand including the gain/loss of rotational and translational entropy, the loss of flexibility of the ligand, geometric imperfections of hydrogen bonds and metal ligation, and the desolvation energy of atom.

regarded as the positive sample. The detection rate is used to measure the performance of the computational method, which is the ratio of positive samples in listed rank of screened ingredients.

305 ingredients were picked out from 820 ingredients of 421 herbs by our pharmacophore model. Then, three ranking lists of these 305 ingredients were generated, respectively, by molecular docking from three individual scoring functions (ASE, Affinity dG, London dG). A final list is obtained by aggregating these three lists. In the top 10 percent of the ranking list, the detection rate reached 0.6 (18/30). The whole results of rank aggregation are shown in Supplementary Table S2.

As validity of methodology, the performance of our method was compared with traditional methods. We predict the inductive herb-drug interactions through screening agonist of PXR. Because candidate agonists screened by us are a ranking list, three methods for screening ligand of protein were chosen to compare, such as molecular docking, Partial Least Squares- (PLS-) based QSAR, Principal Component Regression- (PCR-) based QSAR. Likewise, 820 herbal ingredients are screened by different methods. As shown in Figure 4, the detection rate of our method (SELF) is higher than any other methods in different top percent of ranking. Our method indeed improves the performance of predicting herb-drug interactions. The result of ranking lists was shown in Supplementary Table S3.

As a part of screened result, the top 10 ingredients in final ranking list are shown in Table 3. They can be found in 14 herbs and 5 of these herbs were reported to be related to herb-drug interactions (shown in Table 3). Three cases are discussed in detail in the following.

Case 1 (*Sophora flavescens*-theophylline interaction). Sophoraflavoside III and Sophoraflavoside IV are isolated from the roots of *Sophora flavescens* (SF), which is used to treat diseases such as diarrhea, gastrointestinal hemorrhage, and eczema [42]. Theophylline, also known as dimethylxanthine, is a methylxanthine drug for the treatment of respiratory diseases such as chronic obstructive pulmonary disease (COPD) and asthma [43]. The two herbal ingredients were potential agonists of PXR according to our pharmacophore and docking analysis. Theophylline is the substrate of CYP enzymes such as CYP2B [44, 45], CYP3A4, CYP1A2, CYP2E1, CYP1A1, CYP1B1, CYP2C8, CYP2C9, and CYP2D6. And the activated PXR by components of SF can induce the gene expression of these enzymes. Therefore, we predict that SF may change the metabolism of theophylline. The SF-theophylline interaction was evaluated in Ueng et al.'s experiment in 2010 [46]. They demonstrated that SF extracts reduced blood theophylline concentration via accelerating the clearance of theophylline in male Sprague-Dawley rats. Also, they were convinced that the expression of some enzymes metabolizing theophylline was upregulated such as CYP1A2, CYP2B1/2, and CYP3A4, all of which are target genes of PXR. The experimental results supported our predicted results. It is notable that our model not only predicted the SF-theophylline interaction successfully but also explained the potential molecular mechanism of interaction.

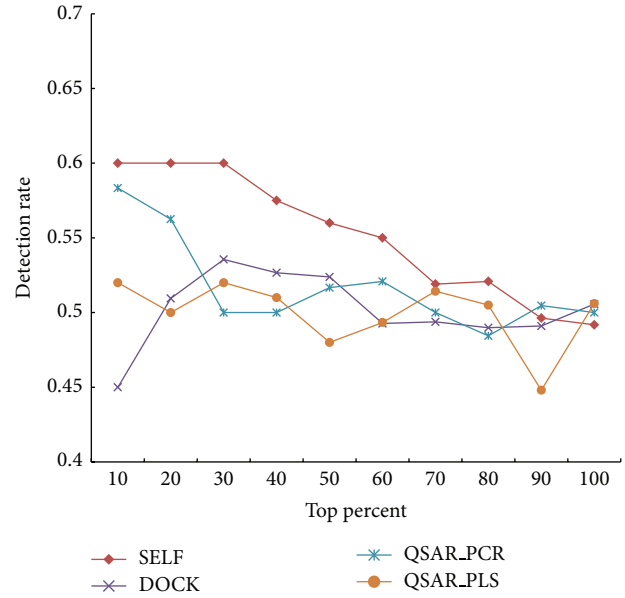


Figure 4: The detection rate in different ranking lists obtained by four methods.

Case 2 (*Sophora flavescens*-nifedipine interaction). This interaction was also observed in Ueng's study. Nifedipine is

Table 3: The top 10 of final rank for candidate agonist of PXR from herbal ingredients.

Rank	Ingredients	Herbs	Reference (Y/N)
1	Sophoraflavoside IV	Sophorae flavescentis	Y
2	Hesperidin	*Sarothamnus scoparius*	Y
		Scrophularia nodosa; *Hyssopus officinalis*; *Tilia* × *europaea*; *Verbascum thapsus*; *Chlorella*	N
3	Sennoside C&D	*Cassia acutifolia*	N
4	Ginsenosides Rgl	*Astragalus membranaceus*	Y
5	Chlorophy II	*Medicago sativa*; *Urtica dioica*	N
6	Solanine	Fritillariae cirrhosae	Y
7	Senegenic acid	*Polygala senega*	N
8	Sophoraflavoside III	Sophorae flavescentis	Y
9	Phellanmurin	*Phellodendron amurense*	Y
10	Torulosic acid	*Juniperus communis*	N

a dihydropyridine calcium channel blocker that primarily blocks L-type calcium channels [47]. A series of genes metabolizing nifedipine are regulated by PXR; SF extract could alter the metabolism of nifedipine by activating PXR. It was also found that the gene of CYP2C11 is upregulated by SF extract in Ueng's study and CYP2C11 is responsible for nifedipine oxidation [48]. But CYP2C11 is not target gene of PXR. So the explanation of SF-nifedipine interaction is outside scope of our model. Multiple interpretations for herb-drug interaction may exist simultaneously. On the one hand, several components in one herb hit multiple targets to influence drugs in different ways, like the influence on nifedipine by SF. On the other hand, different components in one herb sharing the same targets can act on the drug in the same way. Therefore, it is emphasized that our computational model only depends on the PXR-involving mechanistic mode and is incompetent to predict the PXR-independent interaction.

Case 3 (Fritillaria-warfarin interaction). A clinical case of a 61-year-old man indicated that fritillaria lessens anticoagulation of warfarin [49]. The patient takes warfarin therapy regularly with a herbal product called Guilinggao resulting in his easy gum bleeding, epistaxis, and skin bruising. The main component of Guilinggao is fritillaria. Solanine which was one of fritillaria's ingredients was screened out as candidate agonist of PXR. Some of the enzymes which are related to warfarin's metabolism are modulated by PXR including CPY2C9, CYP1A2, CYP2C19, CYP3A4, and CYP2C8. According to our mechanistic mode for herb-drug interaction, fritillaria has an influence on warfarin's metabolism through changing the expression of some metabolic enzymes.

In our results, some ingredients were not reported to be associated with herb-drug interaction. Two potential interpretations are as follows: (1) the ingredient does interact with some drugs, but the interaction is not yet discovered *in vitro* and *in vivo*; (2) as a result of false positive from our pharmacophore model, the ingredients are not agonists of PXR.

4. Conclusions

In this study, a combinational in silico strategy was proposed to predict inductive herb-drug interactions. As a consequence, among 820 ingredients from 421 herbs, a ranking list of 305 ingredients was generated as candidate agonists of PXR. Among the top 10 herb ingredients from the ranking list, 6 ingredients were reported to involve herb-drug interactions. The strategy also could be extended to studies on other inductive herb-drug interactions. Besides, during the process of screening agonists for PXR, our pharmacophore model achieved a good performance across a broad dataset. What is more, the ranking result of traditional molecular docking was improved by rank aggregation. It is suggested that combining merits of scoring functions with less redundancies is a new orientation to optimize scoring functions.

Conflict of Interests

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

Acknowledgments

This work was supported by National Natural Science Foundation of China 30976611 (to RZ) and 31171272 (to WZ) and the Fundamental Research Funds for the Central Universities 2000219083 (to RZ). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper.

References

[1] A. Tachjian, V. Maria, and A. Jahangir, "Use of herbal products and potential interactions in patients with cardiovascular diseases," *Journal of the American College of Cardiology*, vol. 55, no. 6, pp. 515–525, 2010.

[2] D. W. Kaufman, J. P. Kelly, L. Rosenberg, T. E. Anderson, and A. A. Mitchell, "Recent patterns of medication use in the

ambulatory adult population of the United States: the Slone survey," *The Journal of the American Medical Association*, vol. 287, no. 3, pp. 337–344, 2002.

[3] T. M. Bush, K. S. Rayburn, S. W. Holloway et al., "Adverse interactions between herbal and dietary substances and prescription medications: a clinical survey," *Alternative Therapies in Health and Medicine*, vol. 13, no. 2, pp. 30–35, 2007.

[4] C. Palleria, A. Di Paolo, C. Giofre et al., "Pharmacokinetic drug-drug interaction and their implication in clinical management," *Journal of Research in Medical Sciences*, vol. 18, no. 7, pp. 601–610, 2013.

[5] D. M. Jonker, S. A. G. Visser, P. H. van der Graaf, R. A. Voskuyl, and M. Danhof, "Towards a mechanism-based analysis of pharmacodynamic drug-drug interactions in vivo," *Pharmacology & Therapeutics*, vol. 106, no. 1, pp. 1–18, 2005.

[6] X. Bi, M. Gong, and L. Di, "Review on prescription compatibility of shaoyao gancao decoction and reflection on pharmacokinetic compatibility mechanism of traditional chinese medicine prescription based on in vivo drug interaction of main efficacious components," *Evidence-Based Complementary and Alternative Medicine* , vol. 2014, Article ID 208129, 2014.

[7] J. M. Lehmann, D. D. McKee, M. A. Watson, T. M. Willson, J. T. Moore, and S. A. Kliewer, "The human orphan nuclear receptor PXR is activated by compounds that regulate CYP3A4 gene expression and cause drug interactions," *The Journal of Clinical Investigation*, vol. 102, no. 5, pp. 1016–1023, 1998.

[8] M. Baes, T. Gulick, H.-S. Choi, M. G. Martinolu, D. Simha, and D. D. Moore, "A new orphan member of the nuclear hormone receptor superfamily that interacts with a subset of retinoic acid response elements," *Molecular and Cellular Biology*, vol. 14, no. 3, pp. 1544–1552, 1994.

[9] M. Makishima, T. T. Lu, W. Xie et al., "Vitamin D receptor as an intestinal bile acid sensor," *Science*, vol. 296, no. 5571, pp. 1313–1316, 2002.

[10] S. Harmsen, I. Meijerman, J. H. Beijnen, and J. H. M. Schellens, "The role of nuclear receptors in pharmacokinetic drug-drug interactions in oncology," *Cancer Treatment Reviews*, vol. 33, no. 4, pp. 369–380, 2007.

[11] D. R. Abernethy and D. A. Flockhart, "Molecular basis of cardiovascular drug metabolism: implications for predicting clinically important drug interactions," *Circulation*, vol. 101, no. 14, pp. 1749–1753, 2000.

[12] J. L. Staudinger, X. Ding, and K. Lichti, "Pregnane X receptor and natural products: beyond drug-drug interactions," *Expert Opinion on Drug Metabolism & Toxicology*, vol. 2, no. 6, pp. 847–857, 2006.

[13] T. K. H. Chang and D. J. Waxman, "Synthetic drugs and natural products as modulators of constitutive androstane receptor (CAR) and pregnane X receptor (PXR)," *Drug Metabolism Reviews*, vol. 38, no. 1-2, pp. 51–73, 2006.

[14] W. Cao and A.-G. Zhao, "Prescription rules of Chinese herbal medicines in treatment of gastric cancer," *Zhong Xi Yi Jie He Xue Bao*, vol. 7, no. 1, pp. 1–8, 2009.

[15] N. Scheer, J. Ross, A. Rode et al., "A novel panel of mouse models to evaluate the role of human pregnane X receptor and constitutive androstane receptor in drug response," *The Journal of Clinical Investigation*, vol. 118, no. 9, pp. 3228–3239, 2008.

[16] J. B. Mills, K. A. Rose, N. Sadagopan, J. Sahi, and S. M. F. de Morais, "Induction of drug metabolism enzymes and MDR1 using a novel human hepatocyte cell line," *Journal of Pharmacology and Experimental Therapeutics*, vol. 309, no. 1, pp. 303–309, 2004.

[17] B. Goodwin, E. Hodgson, and C. Liddle, "The orphan human pregnane X receptor mediates the transcriptional activation of CYP3A4 by rifampicin through a distal enhancer module," *Molecular Pharmacology*, vol. 56, no. 6, pp. 1329–1339, 1999.

[18] A. Khandelwal, M. D. Krasowski, E. J. Reschly, M. W. Sinz, P. W. Swaan, and S. Ekins, "Machine learning methods and docking for predicting human pregnane X receptor activation," *Chemical Research in Toxicology*, vol. 21, no. 7, pp. 1457–1467, 2008.

[19] S. Kortagere, D. Chekmarev, W. J. Welsh, and S. Ekins, "Hybrid scoring and classification approaches to predict human pregnane X receptor activators," *Pharmaceutical Research*, vol. 26, no. 4, pp. 1001–1011, 2009.

[20] S. Kortagere, M. D. Krasowski, E. J. Reschly, M. Venkatesh, S. Mani, and S. Ekins, "Evaluation of computational docking to identify pregnane X receptor agonists in the ToxCast database," *Environmental Health Perspectives*, vol. 118, no. 10, pp. 1412–1417, 2010.

[21] M. Dybdahl, N. G. Nikolov, E. B. Wedebye, S. O. Jónsdóttir, and J. R. Niemelä, "QSAR model for human pregnane X receptor (PXR) binding: screening of environmental chemicals and correlations with genotoxicity, endocrine disruption and teratogenicity," *Toxicology and Applied Pharmacology*, vol. 262, no. 3, pp. 301–309, 2012.

[22] M. N. Jacobs, "In silico tools to aid risk assessment of endocrine disrupting chemicals," *Toxicology*, vol. 205, no. 1-2, pp. 43–53, 2004.

[23] C. Y. Ung, H. Li, C. W. Yap, and Y. Z. Chen, "In silico prediction of pregnane X receptor activators by machine learning approaches," *Molecular Pharmacology*, vol. 71, no. 1, pp. 158–168, 2007.

[24] S. Ekins and J. A. Erickson, "A pharmacophore for human pregnane X receptor ligands," *Drug Metabolism and Disposition*, vol. 30, no. 1, pp. 96–99, 2002.

[25] G. Lemaire, C. Benod, V. Nahoum et al., "Discovery of a highly active ligand of human pregnane X receptor: a case study from pharmacophore modeling and virtual screening to 'in vivo' biological activity," *Molecular Pharmacology*, vol. 72, no. 3, pp. 572–581, 2007.

[26] R. E. Watkins, G. B. Wisely, L. B. Moore et al., "The human nuclear xenobiotic receptor PXR: structural determinants of directed promiscuity," *Science*, vol. 292, no. 5525, pp. 2329–2333, 2001.

[27] L. B. Moore, D. J. Parks, S. A. Jones et al., "Orphan nuclear receptors constitutive androstane receptor and pregnane X receptor share xenobiotic and steroid ligands," *The Journal of Biological Chemistry*, vol. 275, no. 20, pp. 15122–15127, 2000.

[28] P. W. Rose, A. Prlic, C. Bi et al., "The RCSB Protein Data Bank: views of structural biology for basic and applied research and education," *Nucleic Acids Research*, 2014.

[29] R. E. Watkins, P. R. Davis-Searles, M. H. Lambert, and M. R. Redinbo, "Coactivator binding promotes the specific interaction between ligand and the pregnane X receptor," *Journal of Molecular Biology*, vol. 331, no. 4, pp. 815–828, 2003.

[30] W. Wang, W. W. Prosise, J. Chen et al., "Construction and characterization of a fully active PXR/SRC-1 tethered protein with increased stability," *Protein Engineering, Design and Selection*, vol. 21, no. 7, pp. 425–433, 2008.

[31] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities," *Nucleic Acids Research*, vol. 35, supplement 1, pp. D198–D201, 2007.

[32] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proceedings of the IEEE Data Mining Conference*, pp. 19–26, Brighton, UK, November 2004.

[33] V. Pihur and S. Datta, "RankAggreg, an R package for weighted rank aggregation," *BMC Bioinformatics*, vol. 10, article 62, 2009.

[34] R. Fagin, E. Berbescu, S. Landis, K. Strumpf, and U. Patil, "Juvenile granulosa cell tumor of the testis," *Urology*, vol. 62, no. 2, p. 351, 2003.

[35] V. Pihur, S. Datta, and S. Datta, "Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach," *Bioinformatics*, vol. 23, no. 13, pp. 1607–1615, 2007.

[36] S. Raychaudhuri, V. Jain, and M. Dongre, "Identification of a constitutively active variant of LuxO that affects production of HA/protease and biofilm development in a non-O1, non-O139 *Vibrio cholerae* O110," *Gene*, vol. 369, no. 1-2, pp. 126–133, 2006.

[37] K. T. Butler, F. J. Luque, and X. Barril, "Toward accurate relative energy predictions of the bioactive conformation of drugs," *Journal of Computational Chemistry*, vol. 30, no. 4, pp. 601–610, 2009.

[38] R. Zhu, L. Hu, H. Li, J. Su, Z. Cao, and W. Zhang, "Novel natural inhibitors of CYP1A2 identified by *in silico* and *in vitro* screening," *International Journal of Molecular Sciences*, vol. 12, no. 5, pp. 3250–3262, 2011.

[39] P. Ferrara, H. Gohlke, D. J. Price, G. Klebe, and C. L. Brooks III, "Assessing scoring functions for protein-ligand interactions," *Journal of Medicinal Chemistry*, vol. 47, no. 12, pp. 3032–3047, 2004.

[40] G. L. Warren, C. W. Andrews, A.-M. Capelli et al., "A critical assessment of docking programs and scoring functions," *Journal of Medicinal Chemistry*, vol. 49, no. 20, pp. 5912–5931, 2006.

[41] H. Kang, Z. Sheng, R. Zhu, Q. Huang, Q. Liu, and Z. Cao, "Virtual drug screen schema based on multiview similarity integration and ranking aggregation," *Journal of Chemical Information and Modeling*, vol. 52, no. 3, pp. 834–843, 2012.

[42] M. Yamazaki, "The pharmacological studies on matrine and oxymatrine," *Yakugaku Zasshi*, vol. 120, no. 10, pp. 1025–1033, 2000.

[43] J. F. Donohue, "Therapeutic responses in asthma and COPD: bronchodilators," *Chest*, vol. 126, no. 2, supplement, pp. 125S–161S, 2004.

[44] Z. Y. Zhang and L. S. Kaminsky, "Characterization of human cytochromes P450 involved in theophylline 8-hydroxylation," *Biochemical Pharmacology*, vol. 50, no. 2, pp. 205–211, 1995.

[45] K. H. Yang, J. H. Lee, and M. G. Lee, "Effects of CYP inducers and inhibitors on the pharmacokinetics of intravenous theophylline in rats: Involvement of CYP1A1/2 in the formation of 1,3-DMU," *Journal of Pharmacy and Pharmacology*, vol. 60, no. 1, pp. 45–53, 2008.

[46] Y. F. Ueng, C. C. Tsai, W. S. Lo, and C. H. Yun, "Induction of hepatic cytochrome p450s by the herbal medicine sophora flavescens extract in rats: impact on the elimination of theophylline," *Drug Metabolism and Pharmacokinetics*, vol. 25, no. 6, pp. 560–567, 2010.

[47] K. Hayashi, K. Homma, S. Wakino et al., "T-type Ca channel blockade as a determinant of kidney protection," *The Keio Journal of Medicine*, vol. 59, no. 3, pp. 84–95, 2010.

[48] J. P. Chovan, S. C. Ring, E. Yu, and J. P. Baldino, "Cytochrome P450 probe substrate metabolism kinetics in Sprague Dawley rats," *Xenobiotica*, vol. 37, no. 5, pp. 459–473, 2007.

[49] A. L. Wong and T. Y. Chan, "Interaction between warfarin and the herbal product *Quilinggao*," *The Annals of Pharmacotherapy*, vol. 37, no. 6, pp. 836–838, 2003.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 121575, 19 pages
http://dx.doi.org/10.1155/2015/121575

Research Article

Gene Signature of Human Oral Mucosa Fibroblasts: Comparison with Dermal Fibroblasts and Induced Pluripotent Stem Cells

Keiko Miyoshi, Taigo Horiguchi, Ayako Tanimura, Hiroko Hagita, and Takafumi Noma

Department of Molecular Biology, Institute of Health Biosciences, The University of Tokushima Graduate School, 3-18-15 Kuramoto-cho, Tokushima 770-8504, Japan

Correspondence should be addressed to Takafumi Noma; ntaka@tokushima-u.ac.jp

Received 21 January 2015; Revised 3 April 2015; Accepted 10 April 2015

Academic Editor: Feng Luo

Oral mucosa is a useful material for regeneration therapy with the advantages of its accessibility and versatility regardless of age and gender. However, little is known about the molecular characteristics of oral mucosa. Here we report the first comparative profiles of the gene signatures of human oral mucosa fibroblasts (hOFs), human dermal fibroblasts (hDFs), and hOF-derived induced pluripotent stem cells (hOF-iPSCs), linking these with biological roles by functional annotation and pathway analyses. As a common feature of fibroblasts, both hOFs and hDFs expressed glycolipid metabolism-related genes at higher levels compared with hOF-iPSCs. Distinct characteristics of hOFs compared with hDFs included a high expression of glycoprotein genes, involved in signaling, extracellular matrix, membrane, and receptor proteins, besides a low expression of HOX genes, the hDFs-markers. The results of the pathway analyses indicated that tissue-reconstructive, proliferative, and signaling pathways are active, whereas senescence-related genes in p53 pathway are inactive in hOFs. Furthermore, more than half of hOF-specific genes were similarly expressed to those of hOF-iPSC genes and might be controlled by WNT signaling. Our findings demonstrated that hOFs have unique cellular characteristics in specificity and plasticity. These data may provide useful insight into application of oral fibroblasts for direct reprograming.

1. Introduction

Oral mucosa is a convenient cell source for regenerative medicine, having the following advantages: (1) simple operation, (2) no cosmetic and functional problems after operation, (3) fast wound healing without scar formation [1], (4) nonkeratinizing epithelia, and (5) no need to consider age and gender differences. Practically, epithelial cell-sheets of human oral mucosa have been used as the grafting material for corneal and esophageal mucosal reconstructions after surgically removing damaged mucosal tissue in regeneration therapy [2, 3]. However, few studies have focused on human oral mucosa fibroblasts (hOFs) as material for regenerative medicine, and little is known about the molecular basis of their characteristics.

Recently, induced pluripotent stem cell (iPSC) technology has shown remarkable progress and has been applied to personalized medicine for diagnostics, drug screening, and regenerative therapy [4]. We also generated human iPSCs from oral mucosa fibroblasts (hOFs-iPSCs), and the excised area of the buccal mucosa was completely healed within a week without any scar formation, as expected [5]. So far, scarless healing is well recognized in fetal, but not adult skin [6]. Therefore, molecular events of the healing process have been studied by comparing postnatal (adult) and fetal skin tissues [7–12]. The differences between fetal and adult healing are strongly related to the production of inflammatory-triggered extracellular matrix (ECM), activation of growth factor signaling, and induction of epithelial-mesenchymal transition (EMT) [1, 10–12]. For example, fibronectin, type III collagen, and hyaluronic acid are more abundant in the fetal skin than in adult skin [1, 8, 11, 13–15]. Furthermore, antifibrotic tumor growth factor-beta3 (TGF-beta3) is highly expressed during fetal wound healing, whereas profibrotic TGF-beta1 and TGF-beta2 are low or absent [1, 7, 11]. These results suggest that skin fibroblasts are deeply involved in ECM deposition and remodeling. In the case of hOFs, higher activity of matrix metalloproteinase-2 (MMP-2) combined

with decreased production and activation of tissue inhibitors of metalloproteinases have been demonstrated by comparing hOFs with skin fibroblasts during ECM remodeling [14].

So far, two comprehensive transcriptome studies have been reported using oral mucosa. One included the comparison of the expression profiles between skin and oral mucosal tissue derived from wound healing mouse models [16]. In this report, oral mucosa epithelial cells produced far less amounts of proinflammatory cytokines compared with skin epithelial cells. The other study compared cultured age-matched human skin fibroblasts with hOFs, showing that wounding stimuli induced cell proliferation and reorganization of collagenous environments in hOFs to a greater extent than in skin fibroblasts [17]. Based on these previous studies, we hypothesized that the sensitivity and plasticity of hOFs may explain their uniqueness and hiPSCs can be used as the alternative for fetal skin fibroblasts to compare the gene profiles.

Additionally, we previously found that endogenous *Krüppel-like factor 4 (KLF4)* and *v-myc avian myelocytomatosis viral oncogene homolog (c-MYC)*, which are the reprogramming factors for generating iPSCs, and *maternally expressed gene 3* (*MEG3*), which is an imprinted gene and long noncoding RNA, were highly expressed in hOFs [5]. *Meg3/Gtl2* is located within the *delta-like 1 homolog 1 (Dlk1)-deiodinase, iodothyronine type III (Dio3)* region and the activation of this region is associated with the level of pluripotency in iPSCs or ESCs [18]. These findings may exhibit a part of plasticity in hOFs.

In this study, we performed comparative analyses of gene profiles of hOFs, hDFs, and hOF-iPSCs to understand the molecular characteristics of hOFs. We chose hOFs derived from the buccal region, not other regions of oral mucosa (gingiva, palate, and tongue) because of its superior accessibility as a cell source appropriate for future regenerative medicine. hOF-iPSCs were used as not only the alternative for fetal skin fibroblasts, but also pluripotent stem cells to find out the specificity in the gene signature of hOFs.

2. Materials and Methods

2.1. Human Fibroblasts. hOFs were isolated from individually collected buccal mucosal tissues obtained from four healthy volunteers (26–35 years old) after receiving written agreement including an informed consent at the Tokushima University Medical and Dental Hospital. Approval from the Institutional Research Ethics Committee of the University of Tokushima was obtained (Project number 708). Details on hOFs isolation have been described previously [5]. After isolation, hOFs were individually designated as hOF1 to hOF4. Among them, we failed to establish primary cell culture from hOF1, so we used three successful cell lines, hOF2, hOF3, and hOF4, for further experiments.

hOFs (hOF2, hOF3, and hOF4) were cultured in Dulbecco's Modified Eagle Medium (DMEM; Nissui, Tokyo, Japan) supplemented with 10% FBS (Nichirei Biosciences, Tokyo, Japan). Three types of hDFs derived from individuals aged 33–36 years old were purchased from the Health Science Research Resources Bank (TIG110, TIG111, and TIG114; Osaka, Japan). hDFs were cultured in Eagle's MEM (EMEM; Nissui) supplemented with 10% FBS (Nichirei Biosciences).

2.2. Generation of hOF-iPSCs. hOF-iPSCs were generated as shown previously [5]. Briefly, mouse *solute carrier family 7, member 1 (mslc7a1)*, was introduced into hOFs using lentiviral infection. Then, four reprogramming factors, *POU class 5 homeobox 1/octamer-binding transcription factor 4 (POU5F1/OCT4)*, *KLF4*, *SRY (sex determining region Y)-box 2 (SOX2)*, and *v-myc avian myelocytomatosis viral oncogene homolog c-MYC*, were transduced by retroviral infection. Generated hOF-iPSCs were maintained in human ES medium (ReproCELL, Tokyo, Japan) supplemented with 5 ng/mL of basic fibroblast growth factor (bFGF) on SNL feeder cells. The pluripotency of hOF-iPSCs was confirmed by the expression of the pluripotent cell markers and by *in vitro* differentiation through embryoid body formation.

2.3. RNA Isolation. RNA samples were prepared from three individual samples in each group (hOFs, hDFs, and hOF-iPSCs; a total of nine samples). Total RNA was isolated using TRI Reagent (Molecular Research Center, Cincinnati, OH, USA), according to the manufacturer's protocol.

2.4. Microarray Analyses. Microarray analysis was performed as previously described [19]. In brief, GeneChip Human Gene 1.0 ST Arrays (Affymetrix, Santa Clara, CA, USA) containing 28,869 oligonucleotide probes for known and unknown genes were used to define gene signatures. First-strand cDNA was synthesized with 400 ng of total RNA from hOFs and hDFs or with 220 ng from hOFs-iPSCs using a WT Expression Kit (Affymetrix), according to the manufacturer's instructions, modified with additional ethanol precipitation. With cRNA obtained from the first-strand cDNA, the second-cycle cDNA reaction was performed. Resulting cDNA was end-labeled with a GeneChip WT Terminal Labeling Kit (Affymetrix). Approximately 5.5 μg of labeled DNA target was hybridized to the array for 17 h at 45°C on the GeneChip Hybridization Oven 640 (Affymetrix). After washing, arrays were stained on a GeneChip Fluidics Station 450 and scanned with a GeneChip Scanner 3000 7G (Affymetrix). A CEL file was generated for each array. All microarray data from the three groups (nine samples in total) have been deposited in Gene expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) under GEO Accession number GSE56805.

2.5. In Silico Data Analyses. The data were analyzed with GeneSpring GX12.0 (Agilent Technologies, Santa Clara, CA, USA). The normalization and summarization of CEL files were performed by "Exon RMA 16" algorithm. After that, the signal values of probe sets were transformed to the value of $\log_2$. For the technological variability, we checked several quality controls including Hybridization Controls (provided by Affymetrix), Histogram, Profile Plot, Matrix Plot, 3D PCA, Pearson's correlation coefficient, and hierarchical clustering

analyses following the standard protocols provided by the manufacturers. Among them, the results of Hybridization controls and Pearson's correlation coefficient were shown in Supplementary Figures S1B and S1C in Supplementary Materials available online at http://dx.doi.org/10.1155/2015/121575, respectively. Each value of Pearson's correlation coefficient is indicated as follows: 1 indicates perfect positive correlation between two samples, 0.80 to 1.0 indicates very strong correlation, and 0.60 to 0.79 indicates strong correlation. Expressed genes that showed a fluorescence intensity greater than 100 were further analyzed. Average gene expression level was calculated for three samples in each group and used for the comparison. To make the stringent criteria, several statistical analyses were performed. First, the data obtained from the differently expressed genes between the 2 groups were analyzed by one-way ANOVA and cut off with the corrected p-value ($p < 0.05$) according to Benjamini-Hochberg (BH) method. Furthermore, Tukey's honestly significant difference (HSD) test was used as the post hoc test, and the differently expressed genes between the 2 groups were extracted. Among 28,869 gene probes, 12,713 gene probes were left after one-way ANOVA and BH analyses (all data was $p < 0.05$, Supplementary Table S1). From these 12,713 gene probes, more than 2-fold differentially expressed gene probes were selected between the two paired groups.

Functional analyses were performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (http://david.abcc.ncifcrf.gov/) [20, 21]. Major biological significance and importance were evaluated by functional annotation clustering (FAC) tool in DAVID. To obtain enrichment clusters of functionally significant and important genes, FAC analysis was performed with the enrichment scores below medium stringency. Pathway analyses were conducted using Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/) pathway tools.

3. Results

3.1. Gene Profiles of Microarray Data. To analyze the molecular profile of hOFs, we prepared three types of cells, hOFs, hDFs, and hOF-iPSCs. Three independent cell lines from different donors were chosen to obtain accurate results from each cell type. Heat map and hierarchical clustering analysis revealed that the gene expression pattern in each group was conserved, except for hDF3 (TIG114) and hOF4, for which intermediate patterns between fibroblasts and hOF-iPSCs were identified (Figure 1(a)). Notably, 56% of probes were expressed at the similar levels (16,156 out of 28,869 probes; less than 2-fold difference) among hOFs, hDFs, and OF-iPSCs. While our samples were not exact age- and gender-matched samples, we observed the strong correlation among the samples by Pearson's correlation coefficient analysis (Supplementary Figure S1A). Each correlation coefficient value among the samples in each group, and also that between hOFs and hDFs, was within the range between 0.9 and 1.0. Furthermore, each correlation coefficient value between either hOFs or hDFs and hOF-iPSCs was within the range between 0.7 and 0.8. These results indicated that our data may, at least in part, exclude the issues about age and gender difference with the strong correlation among the samples. The reliability of microarray hybridization techniques were confirmed by the company-supplied hybrydization control (Figure S1B).

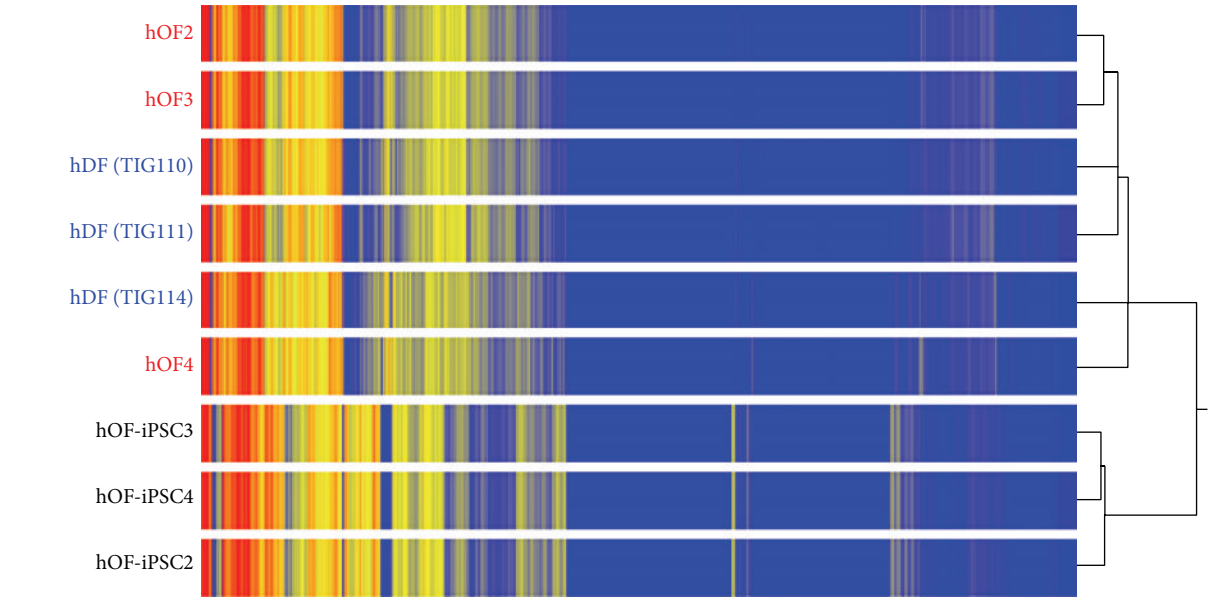
Next, average gene expression signal values in each group were calculated and used for further comparative analyses. Figure S1C shows the scattered plot of gene profile comparison between hOFs and hDFs. Each gene expression of samples was indicated as a spot, and most of them were exhibited within 2-fold line (green line). Therefore, threshold can be set at 2-fold to find the difference of gene expression profile between hOFs and hDFs.

Out of 12,713 probes, we found that 5,738 probes and 5,672 probes (45%) were more than 2-fold differently expressed in hOFs and hDFs, respectively, compared with those in hOF-iPSCs (Figure 1(b), upper panel, left). Approximately 2,300 probes were highly expressed, whereas the expression of 3,400 probes was lower in both hOFs and hDFs than in hiPSCs (Figure 1(b), lower panel). In contrast, only 3.4% (434/12,713 probes) of differentially expressed probes were observed between hOFs and hDFs (Figure 1(b), upper, right). Among these, 272 probes had a high expression and 162 probes had a low expression in hOFs compared with the expression in hDFs (Figure 1(b), upper panel, right).

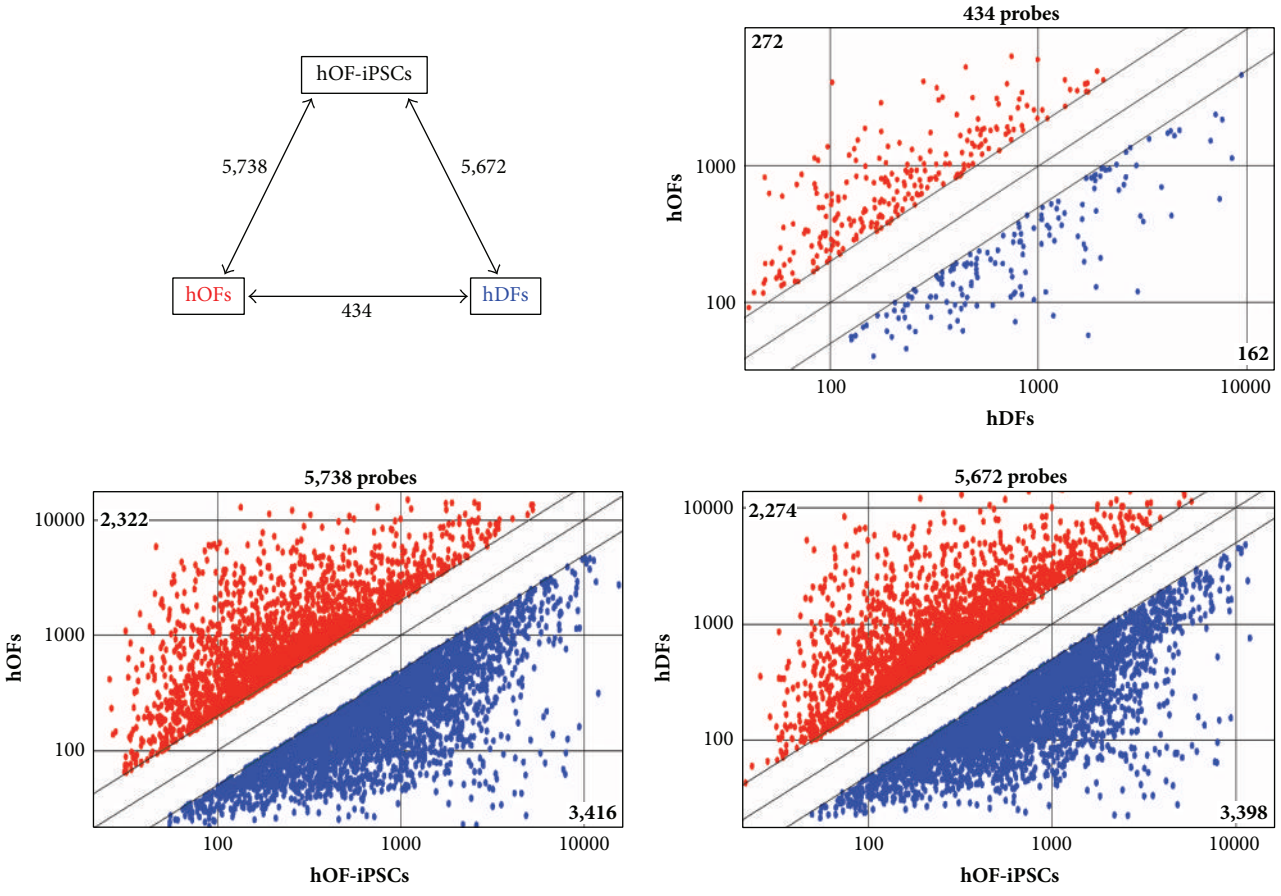
3.2. Enriched Pathways in Fibroblasts. At the beginning, we confirmed the expression levels of several embryonic stem cells (ESCs) markers and reprogramming factors that had been generally observed in iPSCs (Figure 2(a)). As expected, hOF-iPSCs highly expressed all pluripotent markers tested for, that is, *micro RNA302a (MIR302A)*, *MIR302B*, *lin-28 homolog A (LIN28A)*, *Nanog homeobox (NANOG)*, *developmental pluripotency associated 4 (DPPA4)*, *glypican 4 (GPC4)*, *prominin 1 (PROM1)*, *growth differentiation factor 3 (GDF3)*, *POU5F1/OCT4*, and *SOX2*. We also found that the reprogramming factors *KLF4* and *c-MYC* were highly expressed in hOFs and hDFs than in hOF-iPSCs. These results were consistent with previous observations [5].

To elucidate the characteristics of hOFs, we first compared the gene profiles of fibroblasts (hOFs or hDFs) with hOF-iPSCs in steady-state condition. For prediction of the biological function of respective gene profiles, we matched functionally related gene groups to the known pathways by pathway analysis using DAVID linked with KEGG. Genes in thirty pathways were expressed at lower levels in hOFs and hDFs than in hOF-iPSCs, suggesting that these pathways are functionally active in hOF-iPSCs (Figure 2(b)). High expression groups in hOF-iPSCs represented pathways of energy metabolism (glycolysis and tricarboxylic acid (TCA) cycle), nucleotide metabolism (DNA replication, DNA repair, and spliceosome), cell cycle metabolism, and membrane lipid metabolism (Figure 2(b) and Supplementary Figure S2).

Conversely, 46 pathways were enriched among the highly expressed genes in hOFs and hDFs compared with those in hOF-iPSCs (Figure 2(c)). We found that the pathways of glycosaminoglycan (GAG) degradation, glycosphingolipid (GSL) biosynthesis, keratan and heparan sulfate biosynthesis, and lysosome metabolism were highly enriched in hOFs



(a)



(b)

Figure 1: Gene expression signatures in hOFs, hDFs, and hOF-iPSCs. (a) Heat map and hierarchical clustering of whole microarray probes for each of the nine samples. Three individual samples were prepared from each of three types of cells, hOFs, hDFs, and hOF-iPSCs. (b) Comparisons of average signal values among the three types of cells, hOFs, hDFs, and hOF-iPSCs. The number indicates differentially expressed genes ($p < 0.05$, ≥2-fold change; upper panel, left). Scatter plots comparing the average signal values of three samples are shown and the number of differentially expressed probes at more than 2-fold levels is indicated as follows: hOFs *versus* hDFs (upper panel, right), hOFs *versus* hOF-iPSCs (lower panel, left), and hDFs *versus* hOF-iPSCs (lower panel, right).

	Gene	hOFs	hDFs	hiPSCs
ES markers	*MIR302B*	22.6	22.8	2,572.0
	MIR302A	23.0	23.1	2,173.7
	LIN28A	96.0	97.1	7,916.2
	NANOG	79.9	77.9	4,640.8
	DPPA4	57.1	53.5	3,322.3
	GPC4	170.5	230.7	2,070.6
	PROM1	46.2	41.6	1,060.1
	GDF3	56.9	58.8	353.1
Reprogramming factors	*POU5F1* (*OCT4*)	139.7	136.9	6,721.7
	SOX2	86.8	88.0	4,183.4
	KLF4	402.1	374.6	155.5
	c-MYC	899.3	1211.8	966.5

(a)

hOFs or hDFs < hiPSCs

Pathways	Enrichment score	
	hOFs	hDFs
DNA replication	4.7	4.2
Mismatch repair	4.6	4.0
One carbon pool by folate	4.1	4.1
Homologous recombination	3.5	3.3
Spliceosome	3.3	3.0
Cell cycle	3.2	3.1
Nonhomologous end-joining	3	3.6
Base excision repair	2.8	2.5
RNA degradation	2.8	2.5
Aldosterone-regulated sodium reabsorption	2.4	1.8
Proteasome	2.4	2.2
Basal transcription factors	2.3	2.3
Glycosphingolipid biosynthesis	2.1	
Cysteine and methionine metabolism	2.1	2.1
Propanoate metabolism	2.1	2.1
Valine, leucine, and isoleucine degradation	2.1	2.1
Progesterone-mediated oocyte maturation	2.1	2.2
Citrate cycle (TCA cycle)		1.9
Nucleotide excision repair	2.1	1.8
Oocyte meiosis	2	1.8
Pyrimidine metabolism	1.9	1.7
Glycine, serine, and threonine metabolism	1.9	
Aminoacyl-tRNA biosynthesis	1.8	1.8
p53 signaling pathway	1.8	1.9
Pyruvate metabolism	1.8	
Systemic lupus erythematosus	1.7	1.5
Purine metabolism	1.6	1.6
Tight junction	1.5	1.5
Ubiquitin-mediated proteolysis	1.4	1.5
Colorectal cancer		1.5

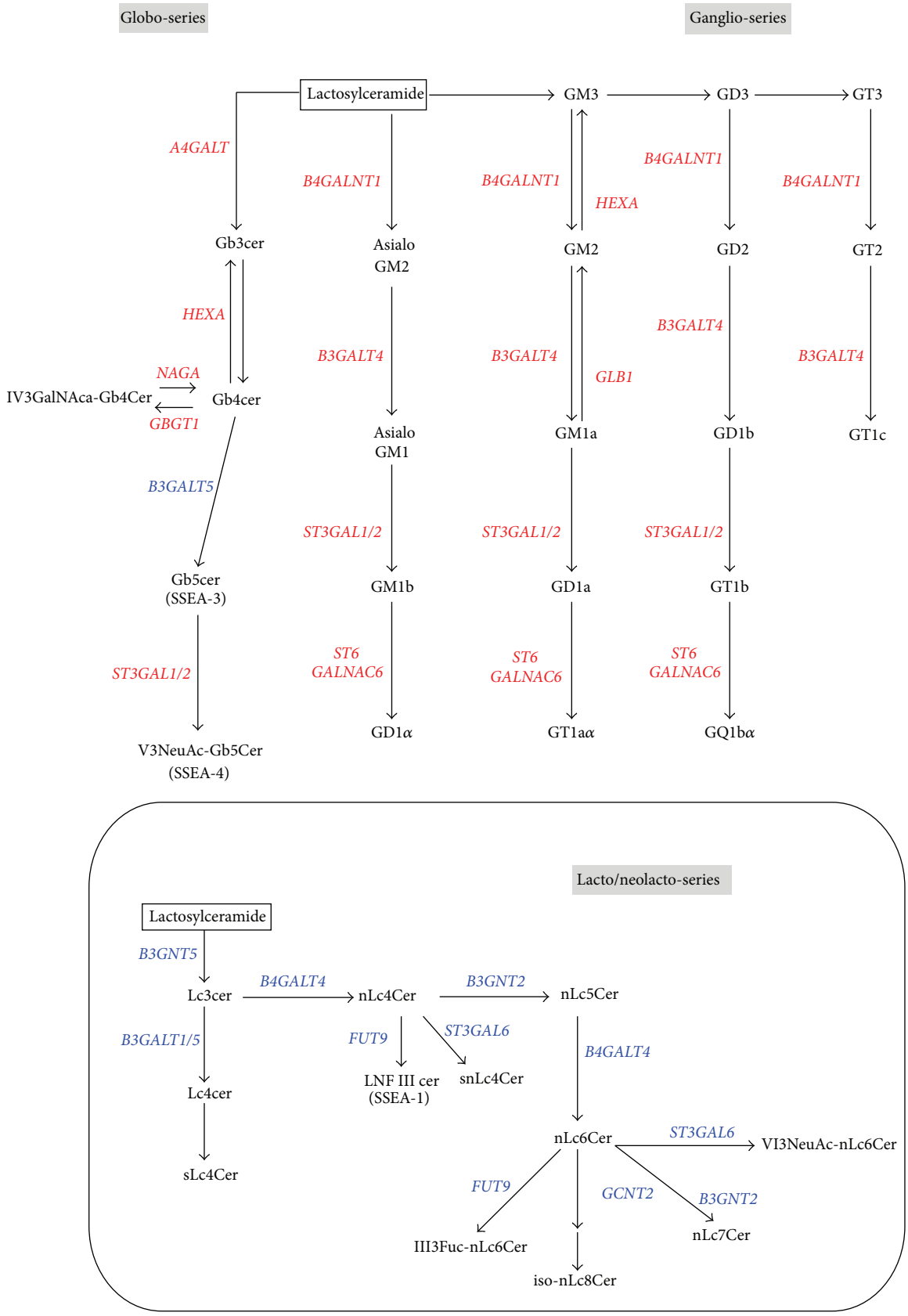
(b)

hOFs or hDFs > hiPSCs

Pathways	Enrichment score	
	hOFs	hDFs
Glycosaminoglycan degradation	4.1	4.8
Glycosphingolipid biosynthesis	3.7	3.6
Bladder cancer	3.5	3.2
Keratan sulfate biosynthesis	3.1	3.2
Chondroitin sulfate biosynthesis		2.9
Heparan sulfate biosynthesis		2.8
Other glycan degradation		2.8
ECM-receptor interaction	2.9	3.4
Pancreatic cancer	2.7	2.6
Lysosome	2.6	2.6
Complement and coagulation cascades	2.5	2.4
mTOR signaling pathway	2.5	2.3
Focal adhesion	2.4	2.7
Melanoma	2.3	2.2
Chronic myeloid leukemia	2.3	2.2
Non-small-cell lung cancer	2.2	2.2
Axon guidance	2.1	1.8
Renal cell carcinoma	2.1	1.8
Acute myeloid leukemia	2.1	2.0
Dilated cardiomyopathy	2.0	2.5
TGF-beta signaling pathway		2.1
Colorectal cancer	1.9	2.1
Intestinal immune network for IgA production		1.9
Arrhythmogenic right ventricular cardiomyopathy (ARVC)		1.8
VEGF signaling pathway	1.8	1.8
Glioma	1.8	1.9
Pathways in cancer	1.7	1.6
Hypertrophic cardiomyopathy (HCM)	1.7	2.1
Adipocytokine signaling pathway	1.7	
PPAR signaling pathway	1.7	
Melanogenesis	1.7	1.6
Vascular smooth muscle contraction	1.7	1.8
GnRH signaling pathway	1.7	1.7
Prostate cancer	1.6	1.8
Gap junction	1.6	1.8
MAPK signaling pathway	1.6	1.7
B cell receptor signaling pathway	1.6	
Progesterone-mediated oocyte maturation		1.7
Fc gamma R-mediated phagocytosis		1.6
Apoptosis	1.6	1.6
Calcium signaling pathway	1.5	1.4
Toll-like receptor signaling pathway	1.5	
Regulation of actin cytoskeleton	1.5	1.6
Cytokine-cytokine receptor interaction	1.3	1.5
Purine metabolism		1.4
Endocytosis		1.4

(c)

Figure 2: Continued.



(d)

Figure 2: Continued.

Glycosphingolipid biosynthesis (ganglio-series)

Genes	Description	hOFs	hDFs	hOF-iPSCs
B3GALT4	*UDP-Gal: betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 4*	310.9	230.4	155.1
B4GALNT1	*Beta-1,4-N-acetyl-galactosaminyl transferase 1*	312.1	287.7	120.7
GLB1	*Galactosidase, beta 1*	2,660.3	2,522.1	1,254.7
HEXA	*Hexosaminidase A (alpha polypeptide)*	2,023.9	1,693.8	495.3
ST3GAL1	*ST3 beta-galactoside alpha-2,3-sialyltransferase 1*	852.1	824.4	109.5
ST3GAL2	*ST3 beta-galactoside alpha-2,3-sialyltransferase 2*	875.9	907.3	435.1
ST6GALNAC6	*ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminidealpha-2,6-sialyltransferase 6*	2,503.2	2,211.2	763.1

Glycosphingolipid biosynthesis (globo-series)

Genes	Description	hOFs	hDFs	hOF-iPSCs
A4GALT	*Alpha 1,4-galactosyltransferase*	683.2	510.6	178.5
GBGT1	*Globoside alpha-1,3-N-acetylgalactosaminyltransferase 1*	572.8	362.6	198.1
HEXA	*Hexosaminidase A (alpha polypeptide)*	2,023.9	1,693.8	495.3
NAGA	*N-acetylgalactosaminidase, alpha-*	1,278.8	852.6	388.3
ST3GAL1	*ST3 beta-galactoside alpha-2,3-sialyltransferase 1*	852.1	824.4	109.5
ST3GAL2	*ST3 beta-galactoside alpha-2,3-sialyltransferase 2*	875.9	907.3	435.1

Glycosphingolipid biosynthesis (lacto/neolacto-series)

Genes	Description	hOFs	hDFs	hOF-iPSCs
B3GNT2	*UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 2*	97.9	105.8	686.1
B3GALT1	*UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 1*	44.3	37.6	1,548.3
B3GALT5	*UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 5*	55.3	57.9	171.7
ST3GAL6	*ST3 beta-galactoside alpha-2,3-sialyltransferase 6*	47.5	48.9	201.2
B3GNT5	*UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 5*	53.5	45.0	217.8
B4GALT4	*UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 4*	142.0	168.0	362.3
GCNT2	*Glucosaminyl (N-acetyl) transferase 2, I-branching enzyme (I blood group)*	43.1	44.2	475.2
FUT9	*Fucosyltransferase 9 (alpha (1,3) fucosyltransferase)*	30.9	32.5	64.4

(e)

Figure 2: Pathway analysis of human fibroblasts and hOF-iPSCs. (a) Gene expression of ESCs markers and reprogramming factors. The numbers indicate average signal values in each cell type. Red: highly expressed genes in hOFs and hDFs compared with hiPSCs. (b) and (c) Pathways with low (b) and high (c) expression in human fibroblasts compared with those in hOF-iPSCs. Numbers indicate enrichment scores provided by DAVID. The top three clusters are colored. Blanks indicate "not listed" in the samples. The top three clusters are highlighted in blue (b) and in red (c), respectively. (d) A diagram of various GSL-biosynthetic pathways. Red and blue colors indicate genes with high and low expressions in human fibroblasts, respectively. cer: ceramide; Gb with subscript: globoside with the number of carbohydrates; G with subscript: ganglioside with subclass; Lc with subscript: lacto- with the number of carbohydrates; nLc-: neolacto-; Fuc: fucose; GalNAc: N-acetylgalactosamine; NeuAc: N-acetylneuraminic acid. (e) Individual gene-expression levels of each GSL-biosynthetic pathway in hOFs, hDFs, and hOF-iPSCs. Red and Blue indicate the same as in (d).

and hDFs. Among them, glycosyltransferases (GTases) in the globo- and ganglio-series of GSL biosynthesis pathways, but not GTases in the lacto- or neolacto-series of the GSL synthetic pathway, were highly expressed in hOFs and hDFs (Figures 2(d) and 2(e)). In addition to these, other signaling components, such as ECM-receptor interaction, complement and coagulation, mammalian target of rapamycin (mTOR) signaling pathway, focal adhesion, and signaling pathways of TGF-beta, mitogen-activated protein kinase (MAPK), vascular endothelial growth factor (VEGF), and calcium were enriched to a greater extent in hOFs and hDFs than in hOF-iPSCs (Figure 2(c)).

3.3. Characterization of hOFs in Comparison with hDFs. Since some of the expressed genes in both hOFs and hDFs must be shared in the biological pathways to display "fibroblastic" characteristics compared with those expressed in hOF-iPSCs, we next tried to elucidate the specificity between hOFs and hDFs. For this purpose, we analyzed a number of genes that were differentially expressed between hOFs and hDFs using microarray analysis, for which overlapping probes were designed and arranged within the same gene to obtain accurate results. Compared with hDFs, 232 genes were overexpressed in hOFs "hOFs > hDFs," whereas 152 genes were underexpressed "hOFs < hDFs." Cranial neural crest

markers especially, such as *distal-less homeobox 5 (DLX5)*, *LIM homeobox 8 (LHX8)*, *paired box 3 (PAX3)*, *PAX9*, and *transcription factor AP-2 alpha (TFAP2A)*, were expressed at a remarkably high level in hOFs (Figure 3(a), left). On the other hand, hDFs expressed homeobox (HOX) cluster genes (Figure 3(a), right) to preserve their positional information as expected [22].

To understand the biological roles of highly expressed genes in hOFs, we performed FAC analysis using DAVID. One hundred and five clusters in hOFs > hDFs and 64 clusters in hOFs < hDFs were observed. The top 12 clusters are shown in Figure 3(b). The top three clusters in hOFs > hDFs were glycoprotein (103 genes), ECM (21 genes), and tube development/embryonic morphogenesis (32 genes). In the glycoprotein cluster, genes related to signaling molecules, extracellular component and matrix, membrane components, and receptors were enriched (Figure 3(c), left), being involved in receiving the extracellular signals. Conversely, transcriptional regulation (20 genes), glycoprotein (63 genes), and transcription activator activity (7 genes) were enriched in hOFs < hDFs. Most of the genes highly enriched in the cluster of transcriptional regulation were HOX genes (Figure 3(c), right), which were shown in Figure 3(a).

Next, we performed pathway analysis to understand the intracellular events in hOFs > hDFs and hOFs < hDFs. In the group of hOFs > hDFs, eleven pathways were enriched (Figure 3(d), Supplementary Table S2). These were categorized into three groups, including (1) tissue-reconstructive pathways (such as complement and coagulation cascades, calcium signaling pathway, endocytosis, chemokine signaling, focal adhesion, and regulation of actin cytoskeleton); (2) differentiation pathways of cranial neural crest lineages (melanogenesis, axon guidance); and (3) growth- and differentiation-inducing factors. The third group comprised three cancer-related pathways (basal cell carcinoma, pancreatic cancer, and pathway in cancer) comprising mainly cytokines, growth factors, and signaling molecules, not oncogenes. In addition, melanogenesis and axon guidance pathways were only detected in hOFs, consistent with hOFs being derived from cranial neural crest cells. In addition, TGF-beta signaling was not enriched independently. Because *TGF-beta3* is expressed higher than *TGF-beta1* and *TGF-beta2* in embryonic skin fibroblasts and opposed to adult skin fibroblasts during wound healing [11], we analyzed TGF-beta signaling pathway-related genes by KEGG program. We found that *TGF-beta2*, *SMAD2* and *SMAD3* were highly expressed, but not *TGF-beta3* (data not shown).

Conversely, only three pathways (p53 signaling, ECM-receptor, and focal adhesion pathways) were enriched in hOFs < hDFs (Figure 3(e), Supplementary Table S3). p53 is known as a tumor suppressor [23], and the expression of *p53* itself showed no difference between hOFs and hDFs (data not shown). However, the downstream genes, *cyclin D1 (CCND1)*, *growth arrest* and *DNA-damage-inducible beta (GADD45B)*, *serpin peptidase inhibitor*, *clade E*, *member 1/plasminogen activator inhibitor type 1 (SERPINE1/PAI-1)*, and *insulin-like growth factor binding protein 3 (IGFBP3)* were downregulated in hOFs. These molecules regulate cell cycle, DNA repair, antiangiogenesis, and the anti-insulin-like growth factor 1 (IGF-1) pathway [23]. *Tenascin C (TNC)*, *integrin*, *alpha 1 (ITGA1)*, *cartilage oligomeric matrix protein (COMP)*, and *ITGA6* were identified and seen to overlap in ECM-receptor and focal adhesion pathways.

3.4. Plasticity and Specificity of hOFs. To further define the characteristics of hOFs, gene groups in hOFs > hDFs and hOFs < hDFs were filtered by similarity in gene-expression level to hOF-iPSCs (Figure 4(a)).

First, we found that 58 genes in hOFs were shared with the similar expression levels in hOF-iPSCs and with the expression levels higher than that in hDFs (hOFs = hiPSCs > hDFs; group G1), suggesting that the genes reflect the plasticity or undifferentiated property of multipotent hOFs by enhancement. Second, 103 genes were highly expressed in hOFs compared with those in hDFs and hiPSCs (hOFs > hiPSCs = hDFs; group G2). The genes in G2 were highly expressed in hOFs but may be kept at low or absent expression levels in hDFs. Therefore, it was suggested that the genes in G2 can exhibit the specificity or differentiated property of hOFs. Third, 70 genes in hOFs had expression levels similar to hOF-iPSCs but were expressed at lower levels than in hDFs (hOFs = hiPSCs < hDFs; group G3). The genes in G3 are defined as the specificity of hDFs; however, these genes could be also involved in the plasticity of hOFs by being suppressed. Twenty-two genes in hOFs were expressed at lower levels than in hDFs that showed expression levels similar to those of hOF-iPSCs (hOFs < hDFs = hiPSCs; group G4), suggesting specificity or differentiated property of hOFs by suppression and, reciprocally, plasticity or undifferentiated property of hDFs.

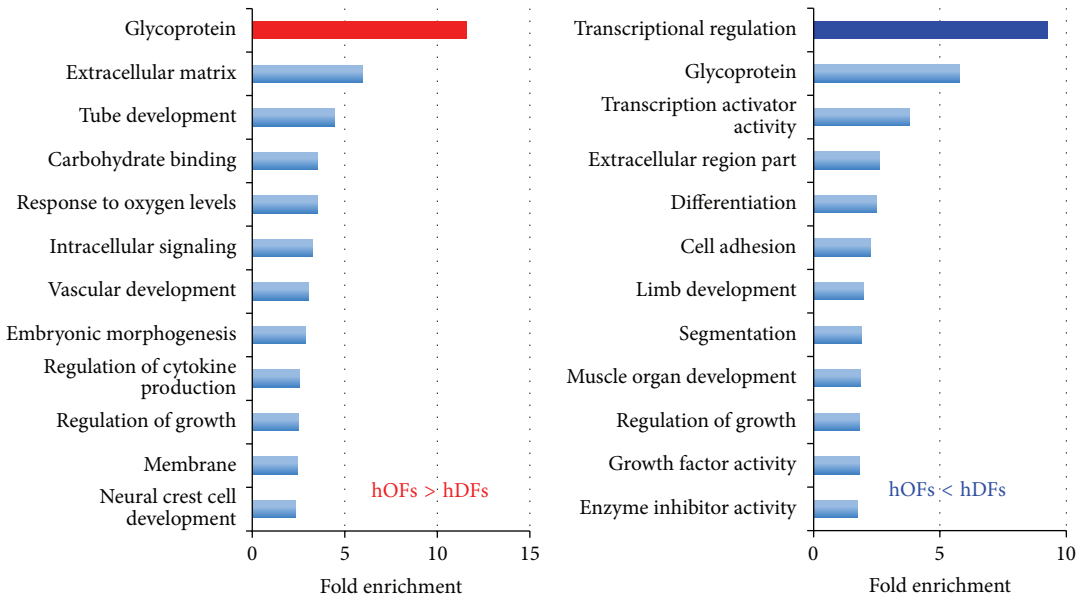
We further analyzed the individual components in the G1–G4 groups, and we categorized them into seven groups, such as ECM/secreted, membrane, receptor, enzyme, signaling, transcriptional regulator, and others (Figure 4(b)). The genes in each group are listed in Supplementary Tables S4–S7. Based on this classification, we found that approximately 30%–40% of the genes in all groups comprised ECM/secreted proteins, membrane proteins, and receptors/transporters, which are highlighted in yellow color in Figure 4(b). These molecular groups are all located at the interface between the cell surface and the extracellular environment, and they may function as a gate of chemical substances and signals (Supplementary Tables S4–S7). Therefore, it is suggested that both fibroblasts are sensitive to environmental factors or cues.

Then we observed that the transcriptional regulator accounted for 10% in both G1 and G2, 34% in G3, and 0% in G4, highlighted by pink color in Figure 4(b) (the gene list, Supplementary Tables S4–S7). This finding is quite important because transcriptional regulators can influence cell fate [24]. Figure 4(c) shows the lists of transcriptional regulators in G1, G2, and G3. The listed genes in G2 and G3 were mostly overlapping with the genes listed in Figures 3(b) and 3(c), which are associated with the fibroblastic specificity of hOFs and hDFs, respectively. The transcriptional regulators in G1 are supposed to represent the gene group related to the plasticity of hOFs because *transcription factor 7-like 1 (T-cell specific, HMG-box)/T-cell factor-3 (TCF7L1/TCF3)* and

	Gene	hOFs	hDFs
Cranial neural crest markers	*DLX5*	418.7	106.4
	LHX8	735.2	68.9
	PAX3	930.1	284.4
	PAX9	4,117.5	102.0
	TFAP2A	2,389.1	404.3

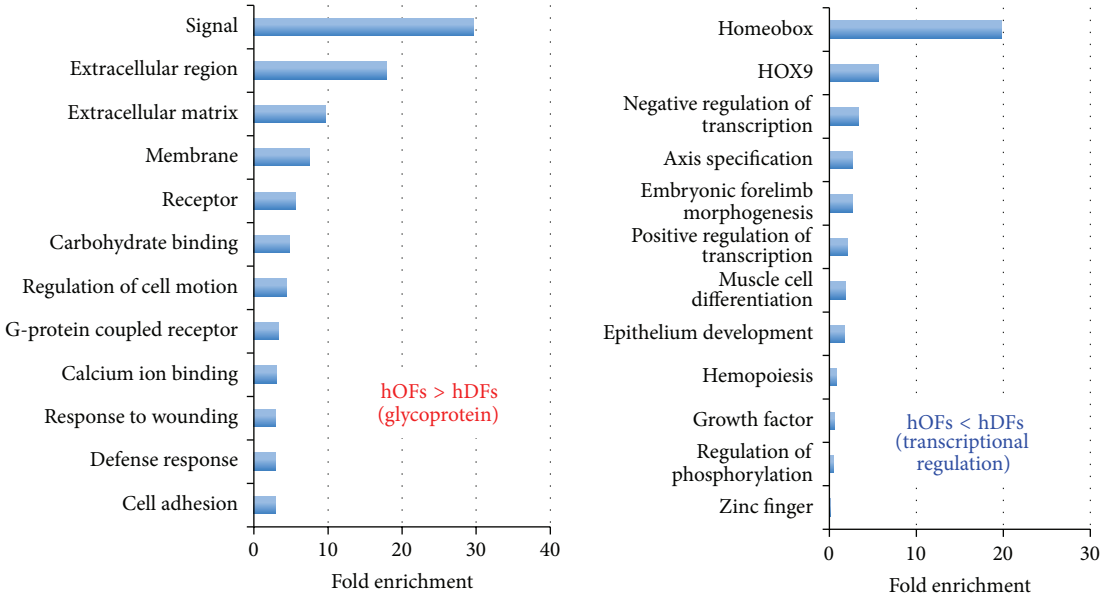
	Gene	hOFs	hDFs
A-P axis markers in the body	*HOXA4*	56.2	197.1
	HOXA6	124.6	251.5
	HOXA7	97.8	443.1
	HOXA9	109.5	453.3
	HOXB2	104.9	949.2
	HOXB3	95.2	504.8
	HOXB5	61.8	186.7
	HOXB6	66.5	375.2
	HOXB7	56.4	125.7
	HOXB8	109.8	301.9
	HOXB9	115.7	759.2
	HOXC10	80.3	1,175.0
	HOXC5	46.1	231.3
	HOXC6	72.6	771.9
	HOXC8	120.0	1,081.5
	HOXC9	62.2	432.3
	HOXD8	138.9	341.9
	HOXD9	172.3	377.6

(a)

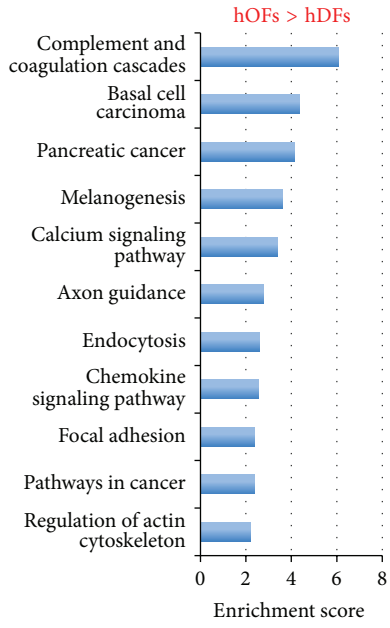


(b)

Figure 3: Continued.



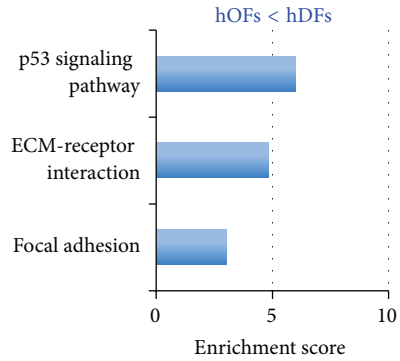
(c)



	Pathways	Genes
1	Complement and coagulation cascades	*F10, F2R, TFPI, C3, CFD, C4A, BDKRB2*
	Calcium signaling pathway	*EDNRA, EDNRB, ADCY9, ADRA1B, F2R, EGFR, PTK2B, BDKRB2, ADCY4, CHRM2*
	Endocytosis	*ARRB1, LDLR, F2R, EGFR, FLT1, DNAJC6, GRK5, EHD3*
	Chemokine signaling pathway	*ARRB1, ADCY9, CXCL14, CXCL16, SHC3, PTK2B, GRK5, ADCY4*
	Focal adhesion	*VTN, ITGA8, EGFR, FLT1, HGF, SHC3, FIGF, PGF*
	Regulation of actin cytoskeleton	*IGF2, FGF18, ITGA8, F2R, EGFR, DIAPH3, BDKRB2, CHRM2*
2	Melanogenesis	*EDNRB, ADCY9, MITF, TCF7L1, ADCY4, WNT16*
	Axon guidance	*SEMA3F, EPHA5, SEMA3D, EPHB6, ROBO2, SEMA4D*
3	Basal cell carcinoma	*BMP4, TCF7L1, PTCH1, WNT16*
	Pancreatic cancer	*TGFB2, EGFR, SMAD3, FIGF, PGF*
	Pathway in cancer	*TCF7L1, EGFR, HGF, BMP4, FIGF, PPARG, PTCH1, WNT16, TGFB2, FGF18, MITF, SMAD3, PGF*

(d)

Figure 3: Continued.



Pathways	Genes
p53 signaling pathway	*CCND1, GADD45B, IGFBP3, SERPINE1*
ECM-receptor interaction	*TNC, ITGA11, COMP, ITGA6*
Focal adhesion	*CCND1, TNC, PDGFC, ITGA11, COMP, ITGA6*

(e)

Figure 3: Comparison of gene profiles in hOFs and hDFs by functional annotation clustering (FAC) and pathway analysis. (a) The positional signatures of hOFs and hDFs as an internal validation. Gene expression of cranial neural crest markers for hOFs (left). Gene expression of anterior-posterior (A-P) axis markers in the body for hDFs (right). Numbers indicate the average signal values in hOFs and hDFs. (b) The top 12 clusters of FAC result in hOFs compared with hDFs. Red bar: the highest enriched cluster in hOFs > hDFs; blue bar: the highest enriched cluster in hOFs < hDFs. (c) The top 12 clusters of FAC result in the individual components of glycoproteins and transcriptional regulation in (b). (d) and (e) Pathway analysis results in genes with high (d) and low (e) expression in hOFs compared with hDFs. Indicated numbers in (b) to (e) represent enrichment scores by DAVID. The number in (d) indicates the three groups categorized in the text. The full names of each gene listed in (d) and (e) are shown in Supplementary Tables S2 and S3, respectively.

transducin-like enhancer of split 1 (E (sp1) homolog, Drosophila Groucho*)* (*TLE1)* are involved in controlling ESCs status by functioning as components of wingless-type MMTV integration site family (WNT) signaling. The transcriptional regulators in G2 are involved in the early developmental regulation, and they are also recognized as markers of the cranial neural crest. The transcriptional regulators in G3 are rich in *HOX* genes, which are involved in determining localization and morphology.

Lastly, we surveyed expression levels of reprogramming regulators because these can support plasticity in fibroblasts. Compared with hOF-iPSCs, the higher expression of reprogramming enhancers, such as *Gli-similar 1 (GLIS1)*, *methyl-CpG-binding domain protein 3 (MBD3)*, *retinoic acid receptor, gamma (RARG)*, and *T-box 3 (TBX3)*, was detected in hOFs and hDFs (Figure 4(d), Supplementary Table S8). Notably, hOFs expressed *RARG* and *TBX3* at the highest level among the three types of cells. Conversely, *LIN28A* was expressed only to a limited extent in hOFs and hDFs. *MEG3*, a human homolog of mouse *Meg3/gene trap locus 2 (Meg3/Gtl2)*, was expressed at the higher level in hOFs than in hDFs and hOF-iPSCs.

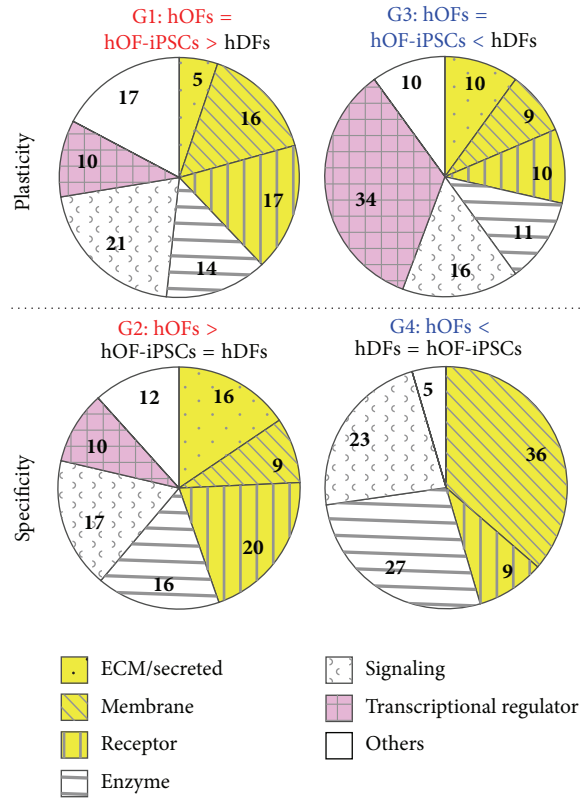
4. Discussion

In this study, we elucidated the unique characteristics of hOFs through comparative analyses of gene expression profiles among hOFs, hDFs, and hOF-iPSCs. In Figure 5(a), we categorized the characteristic gene profile in hOFs that the common fibroblastic features as observed in hOFs and hDFs compared with hOF-iPSCs (upper box) and the specific characteristics of "hOFs" can be demonstrated by comparing with hDFs (lower box). Based on these findings, we developed the possible gene network in hOFs as shown in Figure 5(b).

4.1. Unique Metabolic Pathways in Human Fibroblasts Compared with iPSCs. First, we noted activated GSL metabolism in both hOFs and hDFs compared with hOF-iPSCs (Figures 2 and 5(a)). GSLs are important for membrane organization, signaling interface to ECM, cell-cell adhesion, and cell recognition [25–27]. Furthermore, some GSLs function as sensors in cellular differentiation and tissue patterning [27, 28]. GSLs are basically categorized into three major groups: (1) the ganglio-series and isoganglio-series, (2) the lacto-series and neolacto-series, and (3) the globo-series and isoglobo-series [25, 27]. The ganglio-series and isoganglio-series GSLs are abundant in the brain and are also detected in ESCs of embryoid bodies, neural lineage cells, macrophages, and B cells. The ganglio-series GSLs are functionally involved in cell adhesion and molecular recognition, forming the "glycosynapse" [29, 30]. For example, monosialodihexosylganglioside (GM3) is involved in integrin regulation, epidermal growth factor (EGF) receptor signaling [31], and lipid raft localization [32]. Conversely, lacto-series and neolacto-series GSLs were originally found in erythrocytes as blood group antigen and in tumors as Lewis X (Le^x) GSL antigen. Stage-specific embryonic antigen-1 (SSEA-1), a marker for both mouse ESCs and embryonic carcinoma cells (ECCs), is also included in this group, and it contains Le^x and mediates homotypic adhesion related to compaction or autoaggregation [25]. Furthermore, the absence of lactotriaosylceramide (Lc3cer) synthase, as shown in *UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 5- (B3GNT5-)* deficient mice, has been reported to cause preimplantation lethality [33] or multiple postnatal defects [34]. The globo-series and isoglobo-series GSLs were originally found in human erythrocytes as the major component. Both SSEA-3 and SSEA-4 are common markers for human ESCs and iPSCs [35, 36].

Original category	Gene number	Group	Category	Gene number	Readout for hOFs
hOFs > hDFs	272	G1	hOFs = hOF-iPSCs > hDFs	58	Plasticity
		G2	hOFs > hOF-iPSCs = hDFs	103	Specificity
hOFs < hDFs	162	G3	hOFs = hOF-iPSCs < hDFs	70	Plasticity
		G4	hOFs < hDFs = hOF-iPSCs	22	Specificity

(a)



(b)

G1: hOFs = hOF-iPSCs > hDFs (Plasticity)

Transcriptional regulation	hOFs	hDFs	hOF-iPSCs
ETS2	740.8	348.6	621.0
SIX4	825.4	399.7	1,327.5
TCEAL7	353.0	175.7	221.6
TCF7L1/TCF3	1,981.9	829.8	1,863.2
TLE1	830.8	319.7	1,129.3
TOX	1,289.9	301.6	782.7

G2: hOFs > hOF-iPSCs = hDFs (Specificity)

Transcriptional regulation	hOFs	hDFs	hOF-iPSCs
DLX5	418.7	106.4	81.9
ETS2	740.8	348.6	621.0
FOXF1	768.9	182.8	101.5
LDB2	1,260.1	428.2	800.9
LHX8	735.2	68.9	57.8
MITF	227.5	111.4	103.6
PAX9	4,117.5	102.0	75.4
PITX1	1,039.9	281.7	163.8
PPARG	524.4	100.8	63.0
TCEAL7	353.0	175.7	221.6

G3: hOFs = hOF-iPSCs < hDFs (Plasticity)

Transcriptional regulation	hOFs	hDFs	hOF-iPSCs
BNC1	110.9	302.9	115.2
ETV5	278.6	753.9	316.9
HOXA4	56.2	197.1	47.7
HOXA6	124.6	251.5	94.9
HOXA7	97.8	443.1	94.5
HOXA9	109.5	453.3	76.2
HOXB2	104.9	949.2	92.0
HOXB3	95.2	504.8	72.8
HOXB5	61.8	186.7	49.1
HOXB6	66.5	375.2	50.9
HOXB7	56.4	125.7	42.7
HOXB8	109.8	301.9	90.8
HOXB9	115.7	759.2	87.4
HOXC10	80.3	1,175.0	59.2
HOXC5	46.1	231.3	41.3
HOXC6	72.6	771.9	59.0
HOXC8	120.0	1,081.5	77.6
HOXC9	62.2	432.3	51.3
HOXD8	138.9	341.9	75.4
HOXD9	172.3	377.6	123.8
LHX9	95.7	601.4	67.0
NKX2-6	137.0	388.3	103.7
SERTAD2	320.0	822.4	186.4
TBX5	118.9	770.4	76.3

(c)

Figure 4: Continued.

Reprogramming regulators	hOFs	hDFs	hOF-iPSCs
ESRRB	74.3	70.6	61.2
FOXH1	179.7	176.8	525.9
GLIS1	363.9	392.0	123.7
LIN28A	96.0	97.1	7,916.2
MBD3	479.5	522.7	317.6
NANOG	79.9	77.9	4,640.8
NR5A2	172.1	129.3	287.9
PRDM14	91.1	102.1	1,727.7
RARG	1,131.6	724.6	349.6
SALL4	122.5	146.4	3,185.7
TBX3	1,845.4	839.4	103.0
MEG3	738.9	418.8	476.7

(d)

Figure 4: Characterization of hOFs in comparison with hDFs and hOF-iPSCs. (a) Strategy to define the characteristics of hOFs. The differentially expressed gene groups between hOFs and hDFs were rearranged by the expression similarity with hOF-iPSCs. The readout for hOFs indicates the characteristics of hOFs. Each gene in G1–G4 is listed in Supplementary Tables S4–S7, respectively. (b) The characterization of each gene group categorized in (a). Numbers indicate the percentage of gene numbers in the individual categories compared with total numbers. The groups colored in yellow show the molecules receiving environmental stimuli. The groups colored in pink represent molecules involved in controlling cell fate. (c) The list of individual transcriptional regulators found in (b). Each indicated number is the average signal value in each cell type. (d) Expression levels of reprograming enhancers among the three cell types. Each indicated number is the average signal value in each cell type. Red: hOFs > hDFs > hOF-iPSCs; blue: hOFs = hDFs > hOF-iPSCs. The full names of each gene listed in (d) are shown in Supplementary Table S8.

In our profiles, GSL-related GTs in the globo-series and ganglio-series GSL biosynthetic pathways were highly expressed in both hOFs and hDFs compared with their expression in hOF-iPSCs, whereas GTs in the lacto-series/neolacto-series GSL biosynthetic pathways were less expressed (Figure 2(d)). GSL expression has been demonstrated to be strictly controlled during both mouse embryonic development *in vivo* [25] and differentiation of human ESCs *in vitro* [37, 38]. Globo-series and lacto-series of GSLs are highly expressed in stem cells, whereas gangliosides are contained in further differentiated cells such as embryoid bodies and neuronal cells [37, 38]. Based on these findings, it is suggested that both hOFs and hDFs have the characteristics of differentiated cells except for the high expression of GTs in the globo-series. However, we found that *UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 5 (B3GALT5)*, which catalyzes the conversion from globotetraosylceramide (Gb4cer) to globopentaosylceramide (Gb5cer) (Figures 2(d) and 2(e)), was lower expressed in both hOFs and hDFs than in hOF-iPSCs. Lower expression of *B3GALT5* may cause the accumulation of Gb4cer or globotriaosylceramide (Gb3cer). Because Gb3cer and other glycosphingolipids are also involved in caveolar-1 oligomerization [39], their accumulation may affect the sorting and trafficking of caveolae in the membrane, resulting in the function of signaling in fibroblasts. Taken together, the expression profiles of GSLs-GTs suggested their possible roles in "the environmental sensor" in fibroblasts through membrane metabolism.

Another unique "fibroblastic" feature is the underexpression of aerobic and anaerobic glycolysis-related genes in hOFs and hDFs (Supplementary Figure S2). This finding suggested that hOFs and hDFs are bioenergetically less active than hOF-iPSCs. A recent study reported that the metabolic switching of energy metabolism is linked with cell fate decision [40], consistent with the change from oxidative phosphorylation in mouse embryonic fibroblasts (MEFs) to glycolysis in iPSCs during reprogramming [41]. In addition, it was also demonstrated that active hypoxia inducible factor 1, alpha subunit (HIF1α), and cytochrome c oxidase (COX) could regulate the metabolic transition from aerobic glycolysis in mouse ESCs to anaerobic glycolysis in mouse epiblast stem cells (EpiSCs) and human ESCs [42]. However, we observed that expression levels of *HIF1α* and *COX* were similar among hOFs, hDFs, and hOF-iPSCs in our profiling data (data not shown). Collectively, hOFs and hDFs appear to exhibit the bioenergetically intermediate phenotype between stem cells and terminally differentiated cells, showing the potential to select cell fate, together with membrane sensing of GSLs.

4.2. Gene Signatures Unique to hOFs Compared with hDFs. Next, we elucidated the differences between hOFs and hDFs by comparative *in silico* analyses. The glycoprotein group was highly enriched in hOFs compared with hDFs (Figure 3(b), left); ECM and membrane components, cell motion, adhesion, and defense responses, which are linked

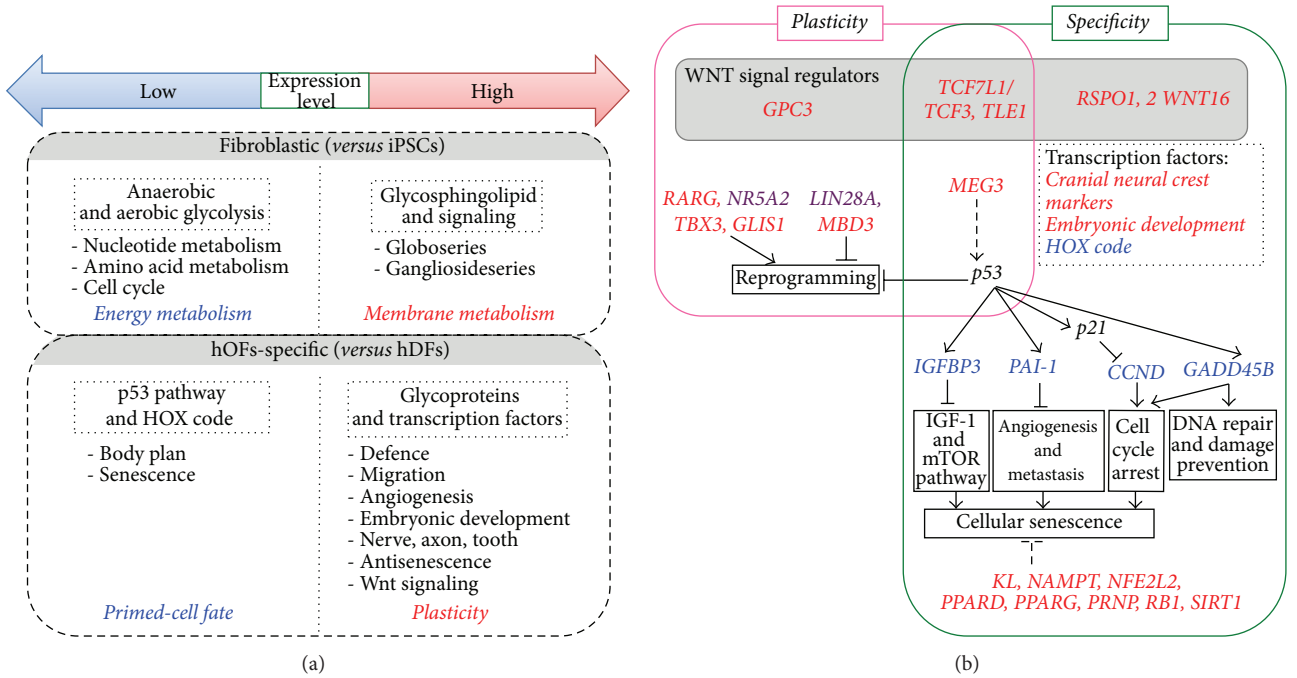


Figure 5: Summary of gene signatures in hOFs. (a) Overview of gene profiles in oral mucosal fibroblasts. Dotted-lined box indicates the category of genes. Red- and blue-colored words in italics show the biological function of gene categories with high and low expressions, respectively. (b) A proposed possible gene network in oral mucosal fibroblasts. Gene names in the different color are indicated as follows. Red: high expression in hOFs compared with that in hDFs (hOFs > hDFs); blue: low expression in hOFs compared with that in hDFs (hOFs < hDFs); purple: similar expression in hOFs and hDFs, but higher than hiPSCs (*hOFs = hDF > hiPSCs*). The box colors indicate biological characteristics or functions as follows. Pink box: possible "plastic" characteristics; green box: possible "specific" characteristics of hOFs; gray box: WNT signal regulators; lined box: biological function; dotted-lined box: the known key transcription factors. Dotted bar indicates indirect effect. Detailed explanations of (a) and (b) are described in the text.

with responses to stimuli from outside the cell, were also sequentially enriched in hOFs (Figure 3(c)). Corresponding pathway analysis revealed that the pathways of tissue reconstruction and differentiation and induction of growth and differentiation factors were active in hOFs (Figure 3(d)). The combination of these highly enriched groups in hOFs may enhance the potential of responding to invasive events or inflammation [43], the advantages of differentiating into melanocytes and neurons (axons) [44, 45], and accessibility of signaling molecules that maintain cell growth or differentiation. These characteristics indicated that hOFs may have the flexibility or plasticity as shown in Figure 5(a).

On the other hand, "transcriptional regulation" was highly enriched in the underexpression group in hOFs (Figure 3(b), right). The components of this group especially were *HOX* genes, conversely representing the specificity of hDFs. Fibroblasts derived from the various anatomical positions in the body have been demonstrated to keep HOX code and position-specified gene signatures to achieve their molecular specification of site-specific variations in fibroblasts [22]. HOX genes are known to regulate anterior-posterior axis, patterning, and timing through development [46]. Although both hOFs and hDFs express their positional information, hOFs might have some plasticity due to low expression of clustered *HOXA* to *HOXD* groups of homeobox genes that tightly control body axis formation (Figure 5(a)).

In addition, we found a low gene expression related to the p53 signaling pathway in hOFs compared with hDFs (Figure 3(e)). p53 is a tumor suppressor gene and its activation regulates multiple events including cell cycle arrest, apoptosis, angiogenesis and metastasis inhibition, DNA repair, IGF-1/mTOR pathway inhibition, reprogramming suppression, and cellular senescence [23, 47, 48]. Although the expression level of *p53* itself was similar in hOFs and hDFs, the downstream genes *CCND1*, *IGFBP3*, and *SERPINE1/PAI-1*, which are involved in p53-induced or stress-induced senescence [49–51], were underexpressed in hOFs. Supportively, we confirmed that the expression of some antisenescence regulators [52] was expressed higher in hOFs compared to those in hDFs (Supplementary Figure S3): *klotho* (*KL*; a membrane protein and suppressor of aging [53]), *nicotinamide phosphoribosyltransferase* (*NAMPT*; a converting enzyme for NAD^+ biosynthesis to increase intracellular NAD^+ levels [54]), *nuclear factor (erythroid derived 2) related factor 2* (*Nrf2*; a transcription factor and induction of antioxidant enzymes [55]), *peroxisome proliferator-activated receptor gamma* (*PPARG*; a transcription factor, antiaging and reduction of physiological stress [56]), *PPAR delta* (*PPARD*, a transcription factor, inhibition of ROS generation [57]), *prion protein* (*PRNP*; a membrane anchored glycoprotein and antioxidant activity [58]), *retinoblastoma 1* (*RB1*; a tumor suppressor protein [59]), and *sirtuin 1* (*SIRT1*; NAD^+

dependent deacetylase and a mammalian longevity protein [60]). These findings suggested that hOFs exhibit not only higher plasticity but also greater longevity compared to hDFs (Figure 5(b)).

4.3. Specificity and Plasticity in hOFs Predicted by the Profiles of Transcription Factors. We performed comparative analyses between hOFs and hOF-iPSCs, which were generated from parental hOFs (Figure 4) to elucidate hOF plasticity and specificity. Focused on the transcription regulators that control cell fate, we developed a plausible gene network to characterize hOFs (Figure 5(b)).

The plasticity of hOFs could be regulated by the high expression of *TCF7L1/TCF3* and *TLE1*, the negative regulators of canonical WNT signaling. In human ESCs, canonical WNT signaling actively regulates pluripotency. However, to differentiate into specific cell types of mesodermal and endodermal lineages, WNT signals need to be transiently downregulated by *TCF7L1/TCF3* and *TLE1* [61–66]. TCF7L1/TCF3 is also defined as a mouse ESC marker [67], and downregulation of TCF7L1/TCF3 has been observed when mouse ESCs differentiate into EpiSCs [68]. Furthermore, *Tcf7l1/Tcf3* regulates stage-specific WNT signaling during the reprogramming of fibroblasts into iPSCs [69], neural stem cell status [70], or epidermal progenitor status [71]. Recently, a new role for *TCF7L1/TCF3* in skin wound healing was reported by demonstrating that *TCF7L1/TCF3* was upregulated in epithelial cells at the site of injury, accelerating wound healing *in vivo* through lipocalin-2 (Lcn2) induction [72]. Another molecule, TLE1, is a transcriptional repressor essential in hematopoiesis and neuronal and epithelial differentiation [73]. Recently, it was reported that TLE1 binds to TCF3 and TCF4 but not to LEF1 and TCF1 and that TCF-TLE1 complexes bind directly to heterochromatin in a specific manner to control transcriptional activation [74]. Furthermore, we found that some positive regulators of WNT signaling were highly expressed by hOFs, for example, a proteoglycan *glypican-3 (GPC3)* [75–77], a secreted protein *R-spondin 1, 2 (RSPO1, 2)* [78, 79], and *WNT16* [80] (Figure 5(b), Supplementary Tables S4 and S5). GPC3 is expressed in pluripotent cells and cancer cells [75–77]. *RSPO1* has been demonstrated to commit to the specification of germ cells, and *RSPO2* plays a role in craniofacial, limb, and branching development [78, 79]. *WNT16* is involved in the specification of hematopoietic stem cells [80]. Taken together, the characteristics of hOFs can be controlled by WNT signaling, and our data is the first report to reveal this by transcriptome profiles. In addition, the cranial neural crest markers were classified into highly expressed gene group of hOFs (Figures 3(a) and 4(c)), and *HOX* genes were repeatedly categorized into the underexpressed gene group of hOFs (Figures 3(b) and 4(c)). These findings suggested that hOFs are differently primed from dermal fibroblasts, but they preserve flexibility or plasticity.

The specificity of hOFs is mainly characterized by a high expression of cranial neural crest markers [44]. *Forkhead box F1 (FOXF1)* (lung), *LIM homeobox 8 (LHX8)* (nerve), *microphthalmia-associated transcription factor (MITF)* (melanogenesis), *PAX9* (tooth, palate, and limb) [81], and *PPARG* (adipocyte) [82, 83] are all involved in embryonic development (Figure 5(b)). These findings suggested that hOFs have some advantage in differentiating into neural crest-derived lineages.

4.4. Plasticity in hOFs is Predicted by Reprogramming Regulators. When we surveyed the detailed hOF gene signatures, we found several important genes associated with plasticity. Recent development of iPSCs technology demonstrated that the cellular plasticity can be acquired by reprogramming with not only four transcription factors, such as Pou5f1/Oct4, Sox2, KLF4, and c-myc [84], but also with additional reprogramming regulators. Among them, we found that two transcription factors, *RARG* [85] and *TBX3* [86], are quite highly expressed in hOFs compared with hDFs (Figure 4(d)). RARG, a nuclear receptor, can form heterodimers with *nuclear receptor subfamily 5*, *group A*, *member 2/liver receptor homolog 1 (NR5A2/LRH-1)* [87], and directly activate *Oct* transcription [88, 89], and the combination with reprogramming factors increased reprogramming efficiency of MEFs into mouse iPSCs [85]. Recently, *Rarg* and *Nr5a2* combined with achaete-scute complex homolog 1 (*Ascl1*), *POU domain*, *class 3*, *transcription factor 2 (Pou3f2/Brn2)*, and *neurogenin 2 (Ngn2)* enhanced the efficiency of transdifferentiation from MEFs to functional neurons [90]. Conversely, *TBX3* is necessary to maintain pluripotency of mouse ESCs and also to regulate differentiation, proliferation, and signaling [86, 91, 92], although *TBX3* in hESC regulates proliferation and differentiation [93]. Although the roles of *RARG* and *TBX3* in hOFs are not fully understood, it might be possible for these to regulate the plasticity of hOFs (Figure 5(b)).

In the other transcription factors, *GLIS1*, [94] was highly expressed, whereas *NR5A2/LRH-1* was underexpressed in both hOFs and hDFs. *LIN28A*, a miRNA and a reprogramming repressor controlling cell plasticity [95, 96], was also expressed at quite low levels in both hOFs and hDFs. Although *MBD3*, the suppression of which can increase reprogramming efficiency [97], was highly expressed in both hOFs and hDFs, these results suggested that both types of fibroblasts might have a similar advantage of reprogramming both cell fate and plasticity. Indeed, in hDFs, less factors or only exogenous *POU5F1/OCT4* can introduce reprogramming [98, 99]. Furthermore, direct induction of transdifferentiation has been reported from hDFs to the other cell lineage without iPSC formation [100–103]. Transdifferentiation has been induced by a combination of specific media and supplements, for example, addition of FGF2 to the culture changed transcriptional profiles in hDFs and promoted regeneration capability [104]. Recently, it was demonstrated that mouse DFs are not a terminally differentiated cell type but can be further differentiated into several different types of fibroblasts to form the dermal structure during skin development and wound healing steps [105]. Since DFs have the plasticity to adapt to the environmental changes *in vitro* and *in vivo* [100–105], hOFs might have similar properties. Further investigations are required to confirm this hypothesis.

In addition, *MEG3* was expressed at a quite high level compared to those of hDFs and hOF-iPSCs (Figure 4(d)).

Because *MEG3* is located within the imprinted *DLK1-DIO3* gene cluster on chromosome 14q32, we further examined the additional imprinted genes (Supplementary Figure S4). Interestingly, hOFs highly expressed both paternal imprinted genes, *DIRAS3* and *IGF2*, and maternal imprinted genes, *H19* and *MEG3,* compared to those in hOF-iPSCs. Furthermore, the expression of *DLK1* and *DIO3*, which are paternally expressed genes and located within the same region as *MEG3*, was lower than that of *MEG3* in hOFs. We found a similar expression pattern within 14q32 in hOF-iPSCs. *MEG8*, known as *Rian* in mouse, is also located within 14q32 and maternally expressed, long noncoding RNAs were not on the lists of gene profiles. The expression of *DIRAS3*, *PEG10*, and *IGF2* was reciprocally observed between hOFs and hOF-iPSCs. Furthermore, although *H19* and *IGF2* exhibit maternal and paternal expressions that are located in the same region, both genes were highly expressed in hOFs. At this moment, we do not know the biological meaning of their expression patterns. Further analyses will be required. *MEG3* is also known as a tumor suppressor via p53 activation [106, 107]. Because the underexpression of p53-downstream genes was observed along with the high expression level of *p53* in hOFs, *MEG3* could also be involved in the specificity of hOFs by controlling p53 signaling as shown in Figure 5(b).

5. Conclusions

We elucidated the fibroblastic plasticity and specificity by analyzing transcriptome profiles of GSL metabolism in hOFs and hDFs. The uniqueness of hOFs is defined as partly primed cells committed to the neural crest cell lineage with plasticity and longevity controlled by *WNT* and *p53* gene network as shown in Figure 5(b). Further analyses are required to prove this hypothesis, but, importantly, our findings in the present study provide a novel basis for discussing the potential application of hOFs in regenerative medicine.

Conflict of Interests

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

Acknowledgments

The authors would like to extend their special thanks to Mr. Hideaki Horikawa, Support Center for Advanced Medical Sciences, the University of Tokushima Graduate School, Institute of Health Biosciences, for his support with the microarray analyses. This work was partly supported by Grants-in-Aid for Scientific Research (no. 24592801), Takeda Science Foundation, and the Presidential Discretion Research budget of the University of Tokushima.

References

[1] J. E. Glim, M. van Egmond, F. B. Niessen, V. Everts, and R. H. J. Beelen, "Detrimental dermal wound healing: what can we learn from the oral mucosa?" *Wound Repair and Regeneration*, vol. 21, no. 5, pp. 648–660, 2013.

[2] K. Nishida, M. Yamato, Y. Hayashida et al., "Corneal reconstruction with tissue-engineered cell sheets composed of autologous oral mucosal epithelium," *The New England Journal of Medicine*, vol. 351, no. 12, pp. 1187–1196, 2004.

[3] T. Ohki, M. Yamato, D. Murakami et al., "Treatment of oesophageal ulcerations using endoscopic transplantation of tissue-engineered autologous oral mucosal epithelial cell sheets in a canine model," *Gut*, vol. 55, no. 12, pp. 1704–1710, 2006.

[4] H. Inoue, N. Nagata, H. Kurokawa, and S. Yamanaka, "IPS cells: a game changer for future medicine," *The EMBO Journal*, vol. 33, no. 5, pp. 409–417, 2014.

[5] K. Miyoshi, D. Tsuji, K. Kudoh et al., "Generation of human induced pluripotent stem cells from oral mucosa," *Journal of Bioscience and Bioengineering*, vol. 110, no. 3, pp. 345–350, 2010.

[6] S. Lamouille, J. Xu, and R. Derynck, "Molecular mechanisms of epithelial-mesenchymal transition," *Nature Reviews Molecular Cell Biology*, vol. 15, no. 3, pp. 178–196, 2014.

[7] M. W. J. Ferguson and S. O'Kane, "Scar-free healing: from embryonic mechanisms to adult therapeutic intervention," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 359, no. 1445, pp. 839–850, 2004.

[8] N. A. Coolen, K. C. W. M. Schouten, B. K. H. L. Boekema, E. Middelkoop, and M. M. W. Ulrich, "Wound healing in a fetal, adult, and scar tissue model: a comparative study," *Wound Repair and Regeneration*, vol. 18, no. 3, pp. 291–301, 2010.

[9] N. A. Coolen, K. C. W. M. Schouten, E. Middelkoop, and M. M. W. Ulrich, "Comparison between human fetal and adult skin," *Archives of Dermatological Research*, vol. 302, no. 1, pp. 47–55, 2010.

[10] M. R. Namazi, M. K. Fallahzadeh, and R. A. Schwartz, "Strategies for prevention of scars: what can we learn from fetal skin?" *International Journal of Dermatology*, vol. 50, no. 1, pp. 85–93, 2011.

[11] S. Kathju, P. H. Gallo, and L. Satish, "Scarless integumentary wound healing in the mammalian fetus: molecular basis and therapeutic implications," *Birth Defects Research Part C: Embryo Today: Reviews*, vol. 96, no. 3, pp. 223–236, 2012.

[12] J. W. Penn, A. O. Grobbelaar, and K. J. Rolfe, "The role of the TGF-beta family in wound healing, burns and scarring: a review," *The International Journal of Burns and Trauma*, vol. 2, no. 1, pp. 18–28, 2012.

[13] K. R. Knight, R. S. C. Horne, D. A. Lepore et al., "Glycosaminoglycan composition of uninjured skin and of scar tissue in fetal, newborn and adult sheep," *Research in Experimental Medicine*, vol. 194, no. 2, pp. 119–127, 1994.

[14] P. Stephens, K. J. Davies, N. Occleston et al., "Skin and oral fibroblasts exhibit phenotypic differences in extracellular matrix reorganization and matrix metalloproteinase activity," *British Journal of Dermatology*, vol. 144, no. 2, pp. 229–237, 2001.

[15] D. J. Whitby and M. W. J. Ferguson, "The extracellular matrix of lip wounds in fetal, neonatal and adult mice," *Development*, vol. 112, no. 2, pp. 651–668, 1991.

[16] L. Chen, Z. H. Arbieva, S. Guo, P. T. Marucha, T. A. Mustoe, and L. A. DiPietro, "Positional differences in the wound transcriptome of skin and oral mucosa," *BMC Genomics*, vol. 11, no. 1, article 47, 2010.

[17] S. Enoch, M. A. Peake, I. Wall et al., "'Young' oral fibroblasts are geno/phenotypically distinct," *Journal of Dental Research*, vol. 89, no. 12, pp. 1407–1413, 2010.

[18] L. Liu, G.-Z. Luo, W. Yang et al., "Activation of the imprinted Dlk1-Dio3 region correlates with pluripotency levels of mouse

stem cells," *Journal of Biological Chemistry*, vol. 285, no. 25, pp. 19483–19490, 2010.

[19] T. W. Utami, K. Miyoshi, H. Hagita, R. D. Yanuaryska, T. Horiguchi, and T. Noma, "Possible linkage of SP6 transcriptional activity with amelogenesis by protein stabilization," *Journal of Biomedicine and Biotechnology*, vol. 2011, Article ID 320987, 10 pages, 2011.

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

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

[22] J. L. Rinn, C. Bondre, H. B. Gladstone, P. O. Brown, and H. Y. Chang, "Anatomic demarcation by positional variation in fibroblast gene expression programs," *PLoS Genetics*, vol. 2, no. 7, article e119, 2006.

[23] B. Vogelstein, D. Lane, and A. J. Levine, "Surfing the p53 network," *Nature*, vol. 408, no. 6810, pp. 307–310, 2000.

[24] C.-F. Pereira, I. R. Lemischka, and K. Moore, "Reprogramming cell fates: insights from combinatorial approaches," *Annals of the New York Academy of Sciences*, vol. 1266, no. 1, pp. 7–17, 2012.

[25] S.-I. Hakomori, "Structure and function of glycosphingolipids and sphingolipids: recollections and future trends," *Biochimica et Biophysica Acta*, vol. 1780, no. 3, pp. 325–346, 2008.

[26] A. R. Todeschini, J. N. Dos Santos, K. Handa, and S.-I. Hakomori, "Ganglioside GM2/GM3 complex affixed on silica nanospheres strongly inhibits cell motility through CD82/cMet-mediated pathway," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 6, pp. 1925–1930, 2008.

[27] G. D'Angelo, S. Capasso, L. Sticco, and D. Russo, "Glycosphingolipids: synthesis and functions," *FEBS Journal*, vol. 280, no. 24, pp. 6338–6353, 2013.

[28] R. Jennemann and H.-J. Gröne, "Cell-specific in vivo functions of glycosphingolipids: lessons from genetic deletions of enzymes involved in glycosphingolipid synthesis," *Progress in Lipid Research*, vol. 52, no. 2, pp. 231–248, 2013.

[29] S.-I. Hakomori, "Glycosynaptic microdomains controlling tumor cell phenotype through alteration of cell growth, adhesion, and motility," *FEBS Letters*, vol. 584, no. 9, pp. 1901–1906, 2010.

[30] A. Prinetti, N. Loberto, V. Chigorno, and S. Sonnino, "Glycosphingolipid behaviour in complex membranes," *Biochimica et Biophysica Acta—Biomembranes*, vol. 1788, no. 1, pp. 184–193, 2009.

[31] S. M. Pontier and F. Schweisguth, "Glycosphingolipids in signaling and development: from liposomes to model organisms," *Developmental Dynamics*, vol. 241, no. 1, pp. 92–106, 2012.

[32] K. Furukawa, Y. Ohkawa, Y. Yamauchi, K. Hamamura, Y. Ohmi, and K. Furukawa, "Fine tuning of cell signals by glycosylation," *Journal of Biochemistry*, vol. 151, no. 6, pp. 573–578, 2012.

[33] F. Biellmann, A. J. Hülsmeier, D. Zhou, P. Cinelli, and T. Hennet, "The Lc3-synthase gene *B3gnt5* is essential to pre-implantation development of the murine embryo," *BMC Developmental Biology*, vol. 8, article 109, 2008.

[34] C.-T. Kuan, J. Chang, J.-E. Mansson et al., "Multiple phenotypic changes in mice after knockout of the B3gnt5 gene, encoding Lc3 synthase—a key enzyme in lacto-neolacto ganglioside synthesis," *BMC Developmental Biology*, vol. 10, article 114, 2010.

[35] H. Suila, V. Pitkänen, T. Hirvonen et al., "Are globoseries glycosphingolipids SSEA-3 and -4 markers for stem cells derived from human umbilical cord blood?" *Journal of Molecular Cell Biology*, vol. 3, no. 2, pp. 99–107, 2011.

[36] A. J. Wright and P. W. Andrews, "Surface marker antigens in the characterization of human embryonic stem cells," *Stem Cell Research*, vol. 3, no. 1, pp. 3–11, 2009.

[37] Y. J. Liang, H. H. Kuo, C. H. Lin et al., "Switching of the core structures of glycosphingolipids from globo- and lacto- to ganglio-series upon human embryonic stem cell differentiation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 52, pp. 22564–22569, 2010.

[38] Y.-J. Liang, B.-C. Yang, J.-M. Chen et al., "Changes in glycosphingolipid composition during differentiation of human embryonic stem cells to ectodermal or endodermal lineages," *Stem Cells*, vol. 29, no. 12, pp. 1995–2004, 2011.

[39] L. Shu and J. A. Shayman, "Glycosphingolipid mediated caveolin-1 oligomerization," *Journal of Glycomics & Lipidomics*, supplement 2, pp. 1–6, 2012.

[40] C. D. L. Folmes, T. J. Nelson, P. P. Dzeja, and A. Terzic, "Energy metabolism plasticity enables stemness programs," *Annals of the New York Academy of Sciences*, vol. 1254, no. 1, pp. 82–89, 2012.

[41] C. D. L. Folmes, T. J. Nelson, A. Martinez-Fernandez et al., "Somatic oxidative bioenergetics transitions into pluripotency-dependent glycolysis to facilitate nuclear reprogramming," *Cell Metabolism*, vol. 14, no. 2, pp. 264–271, 2011.

[42] W. Zhou, M. Choi, D. Margineantu et al., "HIF1α induced switch from bivalent to exclusively glycolytic metabolism during ESC-to-EpiSC/hESC transition," *The EMBO Journal*, vol. 31, no. 9, pp. 2103–2116, 2012.

[43] T. J. Shaw and P. Martin, "Wound repair at a glance," *Journal of Cell Science*, vol. 122, no. 18, pp. 3209–3213, 2009.

[44] N. M. Le Douarin, S. Creuzet, G. Couly, and E. Dupin, "Neural crest cell plasticity and its limits," *Development*, vol. 131, no. 19, pp. 4637–4650, 2004.

[45] A. J. Thomas and C. A. Erickson, "The making of a melanocyte: the specification of melanoblasts from the neural crest," *Pigment Cell & Melanoma Research*, vol. 21, no. 6, pp. 598–610, 2008.

[46] A. J. Durston, S. Wacker, N. Bardine, and H. J. Jansen, "Time space translation: a hox mechanism for vertebrate A-P patterning," *Current Genomics*, vol. 13, no. 4, pp. 300–307, 2012.

[47] P. Hasty and B. A. Christy, "p53 as an intervention target for cancer and aging," *Pathobiology of Aging & Age-Related Diseases*, vol. 3, Article ID 22702, 2013.

[48] H. Hong, K. Takahashi, T. Ichisaka et al., "Suppression of induced pluripotent stem cell generation by the p53-p21 pathway," *Nature*, vol. 460, no. 7259, pp. 1132–1135, 2009.

[49] J. P. Dean and P. S. Nelson, "Profiling influences of senescent and aged fibroblasts on prostate carcinogenesis," *British Journal of Cancer*, vol. 98, no. 2, pp. 245–249, 2008.

[50] R. M. Kortlever and R. Bernards, "Senescence, wound healing and cancer: The PAI-1 connection," *Cell Cycle*, vol. 5, no. 23, pp. 2697–2703, 2006.

[51] F. Lanigan, J. G. Geraghty, and A. P. Bracken, "Transcriptional regulation of cellular senescence," *Oncogene*, vol. 30, no. 26, pp. 2901–2911, 2011.

[52] E. S. Hwang, "Senescence suppressors: their practical importance in replicative lifespan extension in stem cells," *Cellular and Molecular Life Sciences*, vol. 71, no. 21, pp. 4207–4219, 2014.

[53] M. Kuro-o, Y. Matsumura, H. Aizawa et al., "Mutation of the mouse klotho gene leads to a syndrome resembling ageing," *Nature*, vol. 390, no. 6655, pp. 45–51, 1997.

[54] E. van der Veer, C. Ho, C. O'Neil et al., "Extension of human cell lifespan by nicotinamide phosphoribosyltransferase," *The Journal of Biological Chemistry*, vol. 282, no. 15, pp. 10841–10845, 2007.

[55] T. W. Kensler, N. Wakabayashi, and S. Biswal, "Cell survival responses to environmental stresses via the Keap1-Nrf2-ARE pathway," *Annual Review of Pharmacology and Toxicology*, vol. 47, pp. 89–116, 2007.

[56] Y. M. Ulrich-Lai and K. K. Ryan, "PPARγ and stress: implications for aging," *Experimental Gerontology*, vol. 48, no. 7, pp. 671–676, 2013.

[57] H. J. Kim, S. A. Ham, M. Y. Kim et al., "PPARδ coordinates angiotensin II-induced senescence in vascular smooth muscle cells through PTEN-mediated inhibition of superoxide generation," *Journal of Biological Chemistry*, vol. 286, no. 52, pp. 44585–44593, 2011.

[58] W. Rachidi, D. Vilette, P. Guiraud et al., "Expression of prion protein increases cellular copper binding and antioxidant enzyme activities but not copper delivery," *The Journal of Biological Chemistry*, vol. 278, no. 11, pp. 9064–9072, 2003.

[59] C. J. Sherr and F. McCormick, "The RB and p53 pathways in cancer," *Cancer Cell*, vol. 2, no. 2, pp. 103–112, 2002.

[60] C. Ho, E. van der Veer, O. Akawi, and J. G. Pickering, "SIRT1 markedly extends replicative lifespan if the NAD^+ salvage pathway is enhanced," *FEBS Letters*, vol. 583, no. 18, pp. 3081–3085, 2009.

[61] T. A. Blauwkamp, S. Nigam, R. Ardehali, I. L. Weissman, and R. Nusse, "Endogenous Wnt signalling in human embryonic stem cells generates an equilibrium of distinct lineage-specified progenitors," *Nature Communications*, vol. 3, article 1070, 2012.

[62] S. Dalton, "Signaling networks in human pluripotent stem cells," *Current Opinion in Cell Biology*, vol. 25, no. 2, pp. 241–246, 2013.

[63] K. C. Davidson, A. M. Adams, J. M. Goodson et al., "Wnt/β-catenin signaling promotes differentiation, not self-renewal, of human embryonic stem cells and is repressed by Oct4," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 12, pp. 4485–4490, 2012.

[64] K. Gertow, C. E. Hirst, Q. C. Yu et al., "WNT3A promotes hematopoietic or mesenchymal differentiation from hESCs depending on the time of exposure," *Stem Cell Reports*, vol. 1, no. 1, pp. 53–65, 2013.

[65] W. Jiang, D. Zhang, N. Bursac, and Y. Zhang, "*WNT3* is a biomarker capable of predicting the definitive endoderm differentiation potential of hESCs," *Stem Cell Reports*, vol. 1, no. 1, pp. 46–52, 2013.

[66] M. Katoh and M. Katoh, "WNT signaling pathway and stem cell signaling network," *Clinical Cancer Research*, vol. 13, no. 14, pp. 4042–4045, 2007.

[67] W. Zhao, X. Ji, F. Zhang, L. Li, and L. Ma, "Embryonic stem cell markers," *Molecules*, vol. 17, no. 6, pp. 6196–6236, 2012.

[68] J. Wray and C. Hartmann, "WNTing embryonic stem cells," *Trends in Cell Biology*, vol. 22, no. 3, pp. 159–168, 2012.

[69] R. Ho, B. Papp, J. A. Hoffman, B. J. Merrill, and K. Plath, "Stage-specific regulation of reprogramming to induced pluripotent stem cells by Wnt signaling and T cell factor proteins," *Cell Reports*, vol. 3, no. 6, pp. 2113–2126, 2013.

[70] A. Kuwahara, H. Sakai, Y. Xu, Y. Itoh, Y. Hirabayashi, and Y. Gotoh, "Tcf3 represses Wnt-β-catenin signaling and maintains neural stem cell population during neocortical development," *PLoS ONE*, vol. 9, no. 5, Article ID e94408, 2014.

[71] H. Nguyen, M. Rendl, and E. Fuchs, "Tcf3 governs stem cell features and represses cell fate determination in skin," *Cell*, vol. 127, no. 1, pp. 171–183, 2006.

[72] Q. Miao, A. T. Ku, Y. Nishino et al., "Tcf3 promotes cell migration and wound repair through regulation of lipocalin 2," *Nature Communications*, vol. 5, article 4088, 2014.

[73] M. Buscarlet and S. Stifani, "The 'Marx' of Groucho on development and disease," *Trends in Cell Biology*, vol. 17, no. 7, pp. 353–361, 2007.

[74] J. V. Chodaparambil, K. T. Pate, M. R. D. Hepler et al., "Molecular functions of the TLE tetramerization domain in Wnt target gene repression," *The EMBO Journal*, vol. 33, no. 7, pp. 719–731, 2014.

[75] M. Capurro, T. Martin, W. Shi, and J. Filmus, "Glypican-3 binds to Frizzled and plays a direct role in the stimulation of canonical Wnt signaling," *Journal of Cell Science*, vol. 127, no. 7, pp. 1565–1575, 2014.

[76] W. Dormeyer, D. van Hoof, S. R. Braam, A. J. R. Heck, C. L. Mummery, and J. Krijgsveld, "Plasma membrane proteomics of human embryonic stem cells and human embryonal carcinoma cells," *Journal of Proteome Research*, vol. 7, no. 7, pp. 2936–2951, 2008.

[77] P. J. Rugg-Gunn, B. J. Cox, F. Lanner et al., "Cell-surface proteomics identifies lineage-specific markers of embryo-derived stem cells," *Developmental Cell*, vol. 22, no. 4, pp. 887–901, 2012.

[78] Y.-R. Jin, T. J. Turcotte, A. L. Crocker, X. H. Han, and J. K. Yoon, "The canonical Wnt signaling activator, R-spondin2, regulates craniofacial patterning and morphogenesis within the branchial arch through ectodermal-mesenchymal interaction," *Developmental Biology*, vol. 352, no. 1, pp. 1–13, 2011.

[79] Y.-R. Jin and J. K. Yoon, "The R-spondin family of proteins: emerging regulators of WNT signaling," *International Journal of Biochemistry and Cell Biology*, vol. 44, no. 12, pp. 2278–2287, 2012.

[80] W. K. Clements, A. D. Kim, K. G. Ong, J. C. Moore, N. D. Lawson, and D. Traver, "A somitic Wnt16/Notch pathway specifies haematopoietic stem cells," *Nature*, vol. 474, no. 7350, pp. 220–225, 2011.

[81] H. Peters, A. Neubüser, K. Kratochwil, and R. Balling, "Pax9-deficient mice lack pharyngeal pouch derivatives and teeth and exhibit craniofacial and limb abnormalities," *Genes and Development*, vol. 12, no. 17, pp. 2735–2747, 1998.

[82] E. D. Rosen, P. Sarraf, A. E. Troy et al., "PPARγ is required for the differentiation of adipose tissue in vivo and in vitro," *Molecular Cell*, vol. 4, no. 4, pp. 611–617, 1999.

[83] P. Tontonoz, E. Hu, and B. M. Spiegelman, "Stimulation of adipogenesis in fibroblasts by PPARγ2, a lipid-activated transcription factor," *Cell*, vol. 79, no. 7, pp. 1147–1156, 1994.

[84] K. Takahashi and S. Yamanaka, "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors," *Cell*, vol. 126, no. 4, pp. 663–676, 2006.

[85] W. Wang, J. Yang, H. Liu et al., "Rapid and efficient reprogramming of somatic cells to induced pluripotent stem cells by retinoic acid receptor gamma and liver receptor homolog 1," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 45, pp. 18283–18288, 2011.

[86] J. Han, P. Yuan, H. Yang et al., "Tbx3 improves the germ-line competency of induced pluripotent stem cells," *Nature*, vol. 463, no. 7284, pp. 1096–1100, 2010.

[87] J.-C. D. Heng, B. Feng, J. Han et al., "The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells

to pluripotent cells," *Cell Stem Cell*, vol. 6, no. 2, pp. 167–174, 2010.

[88] E. Barnea and Y. Bergman, "Synergy of SF1 and RAR in activation of Oct-3/4 promoter," *The Journal of Biological Chemistry*, vol. 275, no. 9, pp. 6608–6619, 2000.

[89] R. T. Wagner and A. J. Cooney, "Minireview: the diverse roles of nuclear receptors in the regulation of embryonic stem cell pluripotency," *Molecular Endocrinology*, vol. 27, no. 6, pp. 864–878, 2013.

[90] Z. Shi, T. Shen, Y. Liu, Y. Huang, and J. Jiao, "Retinoic acid receptor γ (Rarg) and nuclear receptor subfamily 5, group a, member 2 (Nr5a2) promote conversion of fibroblasts to functional neurons," *The Journal of Biological Chemistry*, vol. 289, no. 10, pp. 6415–6428, 2014.

[91] H. Niwa, K. Ogawa, D. Shimosato, and K. Adachi, "A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells," *Nature*, vol. 460, no. 7251, pp. 118–122, 2009.

[92] Y. Takashima and A. Suzuki, "Regulation of organogenesis and stem cell properties by T-box transcription factors," *Cellular and Molecular Life Sciences*, vol. 70, no. 20, pp. 3929–3945, 2013.

[93] T. Esmailpour and T. Huang, "TBX3 promotes human embryonic stem cell proliferation and neuroepithelial differentiation in a differentiation stage-dependent manner," *Stem Cells*, vol. 30, no. 10, pp. 2152–2163, 2012.

[94] M. Maekawa, K. Yamaguchi, T. Nakamura et al., "Direct reprogramming of somatic cells is promoted by maternal transcription factor Glis1," *Nature*, vol. 474, no. 7350, pp. 225–229, 2011.

[95] N. Shyh-Chang and G. Q. Daley, "Lin28: primal regulator of growth and metabolism in stem cells," *Cell Stem Cell*, vol. 12, no. 4, pp. 395–406, 2013.

[96] K. Tanabe, M. Nakamura, M. Narita, K. Takahashi, and S. Yamanaka, "Maturation, not initiation, is the major roadblock during reprogramming toward pluripotency from human fibroblasts," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 30, pp. 12172–12179, 2013.

[97] Y. Rais, A. Zviran, S. Geula et al., "Deterministic direct reprogramming of somatic cells to pluripotency," *Nature*, vol. 502, no. 7469, pp. 65–70, 2013.

[98] A. Radzisheuskaya, G. le Bin Chia, R. L. dos Santos et al., "A defined Oct4 level governs cell state transitions of pluripotency entry and differentiation into all embryonic lineages," *Nature Cell Biology*, vol. 15, no. 6, pp. 579–590, 2013.

[99] J. Sterneckert, S. Höing, and H. R. Schöler, "Concise review: Oct4 and more: the reprogramming expressway," *Stem Cells*, vol. 30, no. 1, pp. 15–21, 2012.

[100] A. I. Abdullah, A. Pollock, and T. Sun, "The path from skin to brain: generation of functional neurons from fibroblasts," *Molecular Neurobiology*, vol. 45, no. 3, pp. 586–595, 2012.

[101] R. Mitchell, E. Szabo, Z. Shapovalova, L. Aslostovar, K. Makondo, and M. Bhatia, "Molecular evidence for OCT4-induced plasticity in adult human fibroblasts required for direct cell fate conversion to lineage specific progenitors," *Stem Cells*, vol. 32, no. 8, pp. 2178–2187, 2014.

[102] Y.-J. Nam, K. Song, X. Luo et al., "Reprogramming of human fibroblasts toward a cardiac fate," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 14, pp. 5588–5593, 2013.

[103] M. Osonoi, O. Iwanuma, A. Kikuchi, and S. Abe, "Fibroblasts have plasticity and potential utility for cell therapy," *Human Cell*, vol. 24, no. 1, pp. 30–34, 2011.

[104] O. Kashpur, D. LaPointe, S. Ambady, E. F. Ryder, and T. Dominko, "FGF2-induced effects on transcriptome associated with regeneration competence in adult human fibroblasts," *BMC Genomics*, vol. 14, no. 1, article 656, 2013.

[105] R. R. Driskell, B. M. Lichtenberger, E. Hoste et al., "Distinct fibroblast lineages determine dermal architecture in skin development and repair," *Nature*, vol. 504, no. 7479, pp. 277–281, 2013.

[106] V. Balik, J. Srovnal, I. Sulla et al., "MEG3: a novel long noncoding potentially tumour-suppressing RNA in meningiomas," *Journal of Neuro-Oncology*, vol. 112, no. 1, pp. 1–8, 2013.

[107] Y. Zhou, X. Zhang, and A. Klibanski, "MEG3 noncoding RNA: a tumor suppressor," *Journal of Molecular Endocrinology*, vol. 48, no. 3, pp. R45–R53, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 185179, 10 pages
http://dx.doi.org/10.1155/2015/185179

Research Article

Improving the Mapping of Smith-Waterman Sequence Database Searches onto CUDA-Enabled GPUs

Liang-Tsung Huang,[1] Chao-Chin Wu,[2] Lien-Fu Lai,[2] and Yun-Ju Li[2]

[1]*Department of Medical Informatics, Tzu Chi University, Hualien 970, Taiwan*
[2]*Department of Computer Science and Information Engineering, National Changhua University of Education, Changhua 500, Taiwan*

Correspondence should be addressed to Chao-Chin Wu; ccwu@cc.ncue.edu.tw

Received 23 January 2015; Revised 25 May 2015; Accepted 8 June 2015

Academic Editor: Liam McGuffin

Sequence alignment lies at heart of the bioinformatics. The Smith-Waterman algorithm is one of the key sequence search algorithms and has gained popularity due to improved implementations and rapidly increasing compute power. Recently, the Smith-Waterman algorithm has been successfully mapped onto the emerging general-purpose graphics processing units (GPUs). In this paper, we focused on how to improve the mapping, especially for short query sequences, by better usage of shared memory. We performed and evaluated the proposed method on two different platforms (Tesla C1060 and Tesla K20) and compared it with two classic methods in CUDASW++. Further, the performance on different numbers of threads and blocks has been analyzed. The results showed that the proposed method significantly improves Smith-Waterman algorithm on CUDA-enabled GPUs in proper allocation of block and thread numbers.

1. Introduction

Sequence alignment is one of the most important methodologies in the field of computational biology [1]. It describes the way of arrangement of DNA/RNA or protein sequences, in order to identify the regions of similarity among them and to infer structural, functional, and evolutionary relationship between the sequences. Sequence alignment enables researchers to compare the sequences of genes or proteins with unknown functions to sequences of well-studied genes or proteins. When a new sequence is found, the structure and function can be easily predicted by performing sequence alignment because a sequence sharing common ancestor would exhibit similar structure or function.

The most widely used sequence alignment algorithm may be the Smith-Waterman algorithm that was first proposed by Smith and Waterman in 1981 [2] and optimized by Gotoh in 1982 [3]. It performs local sequence alignment, which is designed especially for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. To determine similar regions between two nucleotide or protein sequences, the Smith-Waterman algorithm, instead of looking at the total sequence, compares segments of all possible lengths and optimizes the similarity measure. The Smith-Waterman finds the alignment in a more quantitative way by giving scores for matches and mismatches for every possible pair of residues. The scores are predefined in scoring matrices, such as PAM (point accepted mutation) [4] and BLOSUM (blocks substitution matrix) [5]. In general, a positive score is assigned for a match, a negative score for a mismatch, and a negative score for a gap penalty.

Although the Smith-Waterman algorithm is one of the most advanced and sensitive pairwise sequence comparison algorithms currently available, it is theoretically about 50 times slower than other popular heuristic-based algorithms, such as FASTA [6, 7] and BLAST (Basic Local Alignment Search Tool) [8, 9]. The Smith-Waterman algorithm is slow because it imposes no constraints on the alignment; that is, no sequences will be filtered out if the final alignment score is not above a predefined threshold. However, the Smith-Waterman algorithm is still widely used because of its high sensitivity of sequence alignment even though it has higher time complexity of algorithm. To enable the

Smith-Waterman algorithm to produce exact results in a reasonably shorter time, much research has been focusing on using various high-performance architectures to accelerate the processing speed of the algorithm [10–27]. In particular, it becomes a recent trend to use the emerging accelerators and many-core architectures, such as field-programmable gate arrays (FPGAs) [10–12], cell/BEs [13–17], and general-purpose graphics processing units (GPUs), to run the Smith-Waterman algorithm [18–26].

FPGAs allow customers to configure large resources of logic gates and RAM blocks to implement complex digital computations after manufacturing. Since an FPGA can be configured to execute the Smith-Waterman algorithm, it can be regarded as special-purpose hardware for the Smith-Waterman, resulting in high execution speed. Note the FPGA-based implementation of the Smith-Waterman is more hardware centric. Cell/BEs are multicore microarchitecture that combines a general-purpose power architecture core with streamlined coprocessing elements. The primary feature of cell/BEs is to greatly accelerate multimedia and vector processing applications by introducing the streaming SIMD extensions 2 (SSE2) technology. SIMD instructions can greatly increase performance when exactly the same operations are to be performed on multiple data objects. The SIMD instructions on cell/BEs are used by several research projects to parallelize the Smith-Waterman algorithm.

Modern general-purpose GPUs are not only powerful graphics engines but also highly parallel programmable processors. Today's GPUs use hundreds of parallel processor cores executing tens of thousands of parallel threads to rapidly solve large problems, now available in many PCs, laptops, workstations, and supercomputers. Because of the availability and the popularity, GPUs have been used to implement the Smith-Waterman algorithm, where CUDASW++ is the leading research that provides the fast, publicly available solution to the exact Smith-Waterman algorithm on commodity hardware [18–20]. CUDASW++ 3.0 is the latest version, which couples CPU and GPU SIMD instructions and carries out concurrent CPU and GPU computations [20].

This study aimed at how to improve CUDASW++, especially for short query sequences. Since we observed that the shared memory in each streaming multiprocessor is not fully utilized in CUDASW++, the execution flow of the Smith-Waterman algorithm was rearranged to fully utilize the shared memory for reducing the amount of slow global memory access. This paper is organized as follows. Section 2 introduces CUDASW++ and CUDA-Enabled GPUs. Section 3 presents our method to map the Smith-Waterman database search algorithm onto a CUDA-Enabled GPU for short query sequences. Section 4 demonstrates the experimental results and analyse the performance. Finally, conclusions are given in Section 5.

2. Related Work

CUDASW++ is one of the key projects for implementing the Smith-Waterman sequence database search algorithm on general-purpose GPUs, where the source code of CUDASW++ is publicly available [18–20]. Liu et al. have proposed three versions of CUDASW++ so far, to map the Smith-Waterman database search algorithm onto nVIDIA GPUs. CUDASW++ 1.0 completes all the Smith-Waterman computations on GPUs by fully exploiting the aggregate power of multiple G200 (and higher) GPUs [18].

CUDASW++ 2.0 aims at optimizing the performance of CUDASW++ 1.0 based on the SIMT abstraction of CUDA-enabled GPUs [19]. Two optimization approaches have been implemented in CUDASW++ 2.0. In the first approach, the authors defined a length threshold to partition the database into two parts. For those sequences of length shorter than the threshold, CUDASW++ 2.0 adopts the intertask parallelization method for their alignments with the query sequence. For the other sequences, the system uses the intratask parallelization method. The intertask parallelization method uses one thread to align one subject sequence with the query sequence. It means that multiple subject sequences are aligned with the query sequence concurrently, without interthread communication. On the other hand, the intratask parallelization method uses all the threads in a block to align one subject sequence with the query sequence. Since the intratask parallelization method imposed communication and synchronization among threads, it has better performance than the intertask method only for the sequences of lengths larger than the predefined threshold. According to the statistics, over 99% of the subject sequences will be aligned by the interthread parallelization method. Our work focuses on improving the inter-task parallelization method due to this observation.

The second approach proposed in CUDASW++ 2.0 is the column-major parallelization method. Similar to the intratask method, all the threads in a block work together to align one subject sequence with the query sequence. However, the column-major method aims to exploit more thread parallelism by speculative computation. That is, threads start the computation of H scores before the dependent data, F scores, are available. The speculative computation assumes the speculative H scores will be larger than or equal to F scores. Since it is speculative computation, the lazy-F loop is used to verify whether the assumption is correct for each H score or not. If any answer is false, all the H scores on the same column have to be recalculated. According to the evaluation results reported, the column-major method outperforms the first approach only for few cases. In practice, CUDASW++ adopts the first approach to perform the alignment.

CUDASW++ 3.0 is written in CUDA C++ and PTX assembly languages, targeting GPUs based on the Kepler architecture. It conducts concurrent CPU and GPU computations to accelerate the Smith-Waterman algorithm [20]. According to the compute powers of the CPU and the GPU used in the system, CUDASW++ 3.0 dynamically distributes all sequence alignment workloads over CPUs and GPUs to balance the runtimes of CPU and GPU computations. On the CPU side, the streaming SIMD extensions- (SSE-) based vector execution units and multithreading are employed to speed up the Smith-Waterman algorithm. On the GPU side, PTX SIMD video instructions are used to parallelize the Smith-Waterman algorithm.

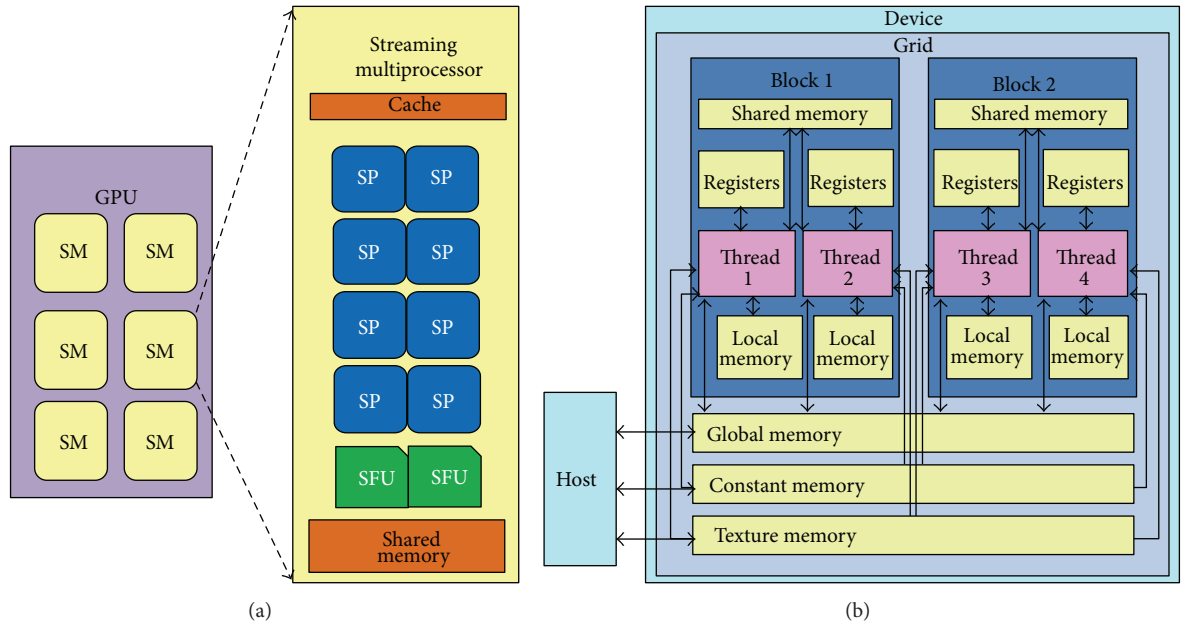


Figure 1: The block diagram of (a) CUDA-enabled GPUs and (b) the memory hierarchy.

Manavski and Valle introduced the idea of query profile and the sorted database [21]. Ligowski and Rudnicki reported their research result almost at the same time as CUDASW++ 1.0 and they investigated how to use shared memory to improve the performance of the Smith-Waterman algorithm [22]. The above two projects both did not exploit the intratask parallelism for long subject sequences. Khajeh-Saeed et al. proposed an interesting parallel scan Smith-Waterman algorithm [23]. They argued that the classic diagonal parallelization approach suffers from nonuniform parallelism distribution across phases of dynamic programming and the memory access pattern is hard to the advantage of memory coalescing. Instead, they aimed to fully parallelize the computation of the cells in one row of the dynamic programming matrix at the same time. To enforce the data dependence between the cells in the same row, they needed to perform the parallel scan to update the values of the cells, resulting in high synchronization overhead between threads and blocks. The parallel scan algorithm can be used in the intratask kernel of CUDASW++. Blazewicz et al. mainly focused on how to improve the backtracking procedure of the Smith-Waterman algorithm [24]. They proposed to use four Boolean matrices to indicate the proper direction of backward moves for every position during the process of backtracking. Their method can be adopted by other packages for further performance improvement, including CUDASW++. Hains et al. proposed using a tiling approach to improve the performance of the intratask kernel of CUDASW++ [25]. They also pointed out several important design issues for tuning performance, including how to ensure that registers are used, instead of global memory, even when the capacity is not exceeded.

CUDA is a new language and development environment, allowing execution of general-purpose applications on NVIDIA's GPUs [28]. The hardware model is comprised of several highly threaded streaming multiprocessors (SMs), where each SM consists of a set of streaming processors (SPs), as shown in Figure 1(a). The computing system consists of a host that is a traditional CPU, also called *host*, and one or more GPUs, also called *device*, as shown in Figure 1(b).

3. Materials and Methodology

3.1. Multiple Subject Sequences of Parallel Method on Smith-Waterman Algorithm. The Smith-Waterman algorithm has been mathematically proven to find the best local alignment of two sequences. The algorithm compares two sequences by computing a score that represents the minimal cost of transforming one sequence to another using two elementary operations: match/mutation and insertion/deletion. If two characters from two sequences match, the cumulative score is increased. However, if one character in the first sequence can be mutated from another character in the second sequence, the cumulative score is either increased or decreased depending on the relationship between these two characters defined in the adopted substitution matrix. There are different substitution matrices for scoring alignment of two sequences. For instance, a BLOSUM (blocks substitution matrix) is a substitution matrix used for sequence alignment of proteins and it records a score for each of the 210 possible substitution pairs of the 20 standard amino acids. Several sets of BLOSUMs exist using different alignment database, where each is named with a number. For two sequences, S_1 and S_2 with lengths L_1 and L_2, the first elementary operation, match/mutation, computes the similarity $H(i, j)$ of these sequences ending at positions i and j in order to identify common subsequences. We call the sequence S_1 the query sequence and the sequence S_2 the subject sequence.

The computation of $H(i, j)$, for $1 \le i \le L_1$ and $1 \le j \le L_2$, is formulated by the following recurrences:

$$\begin{aligned}
&\text{Deletion } E(i, j) = \max\{E(i, j-1), H(i, j-1) - \rho\} \\
&\quad - \sigma, \\
&\text{Insertion } F(i, j) = \max\{F(i-1, j), H(i-1, j) - \rho\} \\
&\quad - \sigma, \\
&\text{Similarity } H(i, j) \\
&\quad = \max\{0, E(i, j), F(i, j), H(i-1, j-1) \\
&\quad + \text{sbt}(S_1[i], S_2[j])\},
\end{aligned} \tag{1}$$

where $\text{sbt}(S_1[i], S_2[j])$ represents the score for the ith character in the sequence S_1 and the jth character in the sequence S_2 defined in the specified substitution matrix. If $S_1[i]$ and $S_2[j]$ are the same character, they are matched; otherwise, it is assumed that the two are derived from an ancestral character, that is, mutation.

Furthermore, in the recurrences, $E(i, j)$ and $F(i, j)$ represent the two cases when gaps are inserted because of different sequence length, where a gap is a consecutive run of spaces in an alignment, represented as a dash on a protein/DNA sequence alignment. To perform a sequence alignment, we write one sequence on top of another, where the characters in one position are deemed to have a common evolutionary origin. If the two sequences are of different lengths, gaps are inserted to make them of equal length. Gaps are used to create alignments that are better conformed to underlying biological models and more closely fit patterns that one expects to find in meaningful alignments. A gap can be inserted into either a query sequence or a subject sequence. Since an insertion in one sequence can always be seen as a deletion in the other one, when a gap is used in the query sequence, it is a gap insertion; otherwise, it is a gap deletion. Gap penalty values are designed to reduce the score when a sequence alignment has been disturbed by gaps. An initial penalty is assigned for a gap opening, ρ, and an additional penalty is assigned for gap extensions that increase the gap length, σ.

We investigate the problem of aligning each subsequence in a database with the query sequence using the Smith-Waterman algorithm, where the database consists of N subject sequences. The problem can be divided into N independent subproblems, where each subproblem is to use the Smith-Waterman algorithm to align the query sequence and one subject sequence. Basically, we solve the problem in a way that is the same as CUDASW++ 2.0. That is, a thread will be assigned to align one subject sequence with the query sequence if the length of the subject sequence is not larger than the user predefined threshold. However, how each thread uses the memory resources on a CUDA-enabled GPU in our method is different from that in CUDASW++ 2.0. On the other hand, for those subject sequences of lengths larger than the threshold, all the threads in a block will perform the alignment in parallel for only one subject sequence. For simplicity, the details of how to use multiple threads to perform an instance of the Smith-Waterman algorithm are omitted.

3.2. Thread Assignment and Sequence Alignment in CUDA-Enabled GPU. We assign one thread for solving one subproblem in a CUDA-enabled GPU. The advantage of such a thread assignment is no interthread communication incurred. Because of the different lengths of the subject sequences in the database, the execution times of the threads are not the same. Therefore, the execution times of threads in a warp will be different and the warp cannot be complete until the slowest thread finishes its work, resulting in a longer warp execution time. Since a block will execute warps one by one, longer warp execution times lead to longer block execution time. To address the problem, the subject sequences allocated to the same warp should be of similar length. To meet this requirement, we preprocess the database by sorting the subject sequences by the length in the ascending order. At the run time, every 32 continuous subject sequences will be assigned to one warp.

Even though the sorted subject sequences can shorten the execution times of warps, the sorted subject sequences cannot be accessed efficiently in the global memory. The reason is explained as follows. Because CUDA-enabled GPUs are SIMD architecture, the 32 threads in a warp will access the ith characters from their assigned subject sequences in parallel, respectively. However, subject sequences are stored one by one in the global memory, resulting in that every 32 ith elements in 32 continuous subject sequences are not stored contiguously. Therefore, the 32 threads in the same warp cannot access the 32 ith elements in one bus transaction. To conquer this problem and to take the advantage of memory coalescing for accessing global memory, the sorted subject sequence database is transformed based on the following method before sending subject sequences to the global memory on a GPU. The elements from each 32 continuous subject sequences are stored interleavingly. In other words, the ith elements from the kth sequences in every 32 continuous subject sequences are stored at $(32 \times i + k)$th memory location.

To align the query sequence with each subject sequence in the database, the query sequence will be used repeatedly. On the other hand, the number of amino acids is only 20. That is, each character of the query sequence will be pairing with the 20 amino acids repeatedly. It is time consuming if each time of pairing has to access the substitution matrix one tome for scoring. Therefore, it is usual to construct a query profile to address the problem. A query profile is a two-dimensional array. The row fields consist of the characters of the query sequence in order and the column fields consist of the 20 amino acids. The value of each cell of a query profile is the score of the relationship between the corresponding amino acid and the corresponding character in the query sequence. The query profile is saved on the texture memory. Each time we can fetch 4 consecutive scores from the texture memory in one access and use four registers to save the 4 scores in a vector fashion. In this way, the cost of accessing the scores can be reduced.

3.3. Memory Allocation at Run Time. To perform the Smith-Waterman algorithm, at run time we need to allocate memory space for the three matrices: H, E, and F. Since the matrices might be very large but the data are intermediate, it is better

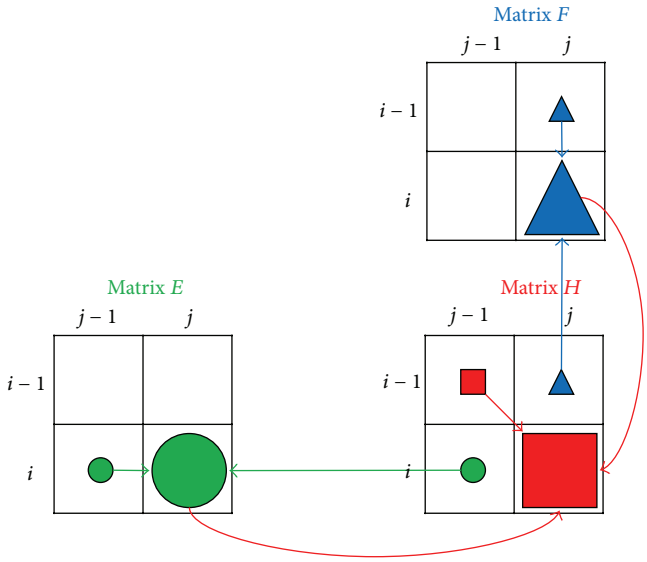


FIGURE 2: The dependence relationship between matrices H, E, and F.

not to save all the matrices data in the global memory. To reduce the required memory space as much as possible, we analyze the dependency relationship among the three matrices, as shown in Figure 2. To compute the element, $H(i, j)$, we require to access the values of $E(i, j)$ and $F(i, j)$, meaning that we have to calculate $E(i, j)$ and $F(i, j)$ before $H(i, j)$ in the same loop iteration. On the other hand, $E(i, j)$ depends on $E(i, j - 1)$ and $H(i, j - 1)$ while $F(i, j)$ depends on $F(i - 1, j)$ and $H(i - 1, j)$. It means that, when executing iteration (i, j), we require the intermediate data produced in iterations $(i - 1, j)$ and $(i, j - 1)$ only. Therefore, we have no need to save all the intermediate data of matrices H, E, and F on global memory. Instead, we can use registers to save the intermediate data produced in the previous iterations and the computation result in the current iteration. Assume the index of the inner loop is j; two registers are sufficient for storing E: one for $E(i, j - 1)$ and one for $E(i, j)$. However, a row of registers is required for F, which is infeasible because of very limited number of registers for a thread. Similarly, we need a row of registers plus two for H. Due to the limited number of registers available in a thread, we use shared memory to buffer the spilled values of registers.

However, all the threads in a block, on a streaming multiprocessor, share the shared memory and its space is not as large as the global memory's. For instance, the amount of shared memory on C1060 is 16 K bytes. If there are 256 threads in a block, each thread can have 64 bytes, that is, 16 words. Therefore, we can swap at most 16 register values out to shared memory for each of the threads. Otherwise, the shared memory is overflow and the slow global memory has to be used to buffer the data due to overflow, which will degrade the overall performance. Therefore, we use our method only when shared memory can store all the spilled register values. In other cases, the original CUDASW++ will be invoked to process the query. Our method will be built in the CUDASW++ package as an execution option. In general, if there are T threads in each block and the amount of available shared memory is S bytes per streaming multiprocessor, the longest length of one query sequence is equal to $S/(4 \times T)$ since matrices H and E require shared memory for buffering and each cell requires two bytes.

3.4. Proposed Mapping Algorithm on CUDA-Enabled GPUs. Registers are the fastest memory and shared memory is faster than global memory one hundred times. To use registers to solve dependence as much as possible and to efficiently use shared memory to buffer spilled registers, we propose the following algorithm to perform the Smith-Waterman algorithm, as shown in Figure 3. Every K consecutive residue on the assigned subject sequence for a thread is grouped in order, padding dummy residues at the end of the subject sequence when necessary. Similarly, every P consecutive residue on the query sequence is in an ordered partition, padding dummy residues at the end of the query sequence when necessary. A tile is defined as the alignment of one group of ordered residues on the subject sequence with a partition of ordered residues on the query sequence. To enforce the dependence, tiles will be aligned one by one in the column-major order, where the tiles on the same column are processed from top to bottom. Furthermore, for a tile, the first residue on the query sequence is aligned with the K residues on the subject sequence one by one, from left to right; then the second, the third, and to the Pth residues are aligned with the K residues on the subject sequence, respectively.

For each tile, the values on the Kth column have to be read by the tile next to it on the right hand side. Similarly, the values on the Pth row have to be read by the tile next to it to the bottom. For each of the F and H matrices, we use K registers to save K consecutive values on the same row, respectively. To calculate the next row, the K registers can be reused because the source operand and the destination operand of an instruction can use the same register. Therefore, the values on the Pth row of a tile can be forwarded through registers to the first row on the next tile, right below the tile. However, the values on the Kth column cannot be forwarded through registers to the next tile on the right hand side because the Kth register is reused to save the value for the next row on the Kth column, even in the same tile. Note that the next tile on the right hand side of the current tile is not the next tile to be processed; it is the next tile on the bottom of the current tile. Consequently, shared memory is used to buffer P values on the Kth column for each tile. The buffered values can be read to calculate the values for the first column of the tile on a tile's immediate right hand side.

Since tiles are processed in column-major order, the shared memory for buffering the Kth column of tiles on the same column can be reused when the next column of tiles is processed. If the query sequence length is not too long, shared memory can buffer all the values on the Kth column of tiles on the same column, no global memory access is required. Let the number of bytes per shared memory be X, the number of threads per block be T, and the number of residues in the query sequence be Q. If $X \geq T \times Q$, no global memory access is required. At the run time, the system can decide if shared memory can buffer all required values on a column based on the hardware configuration and the length of the

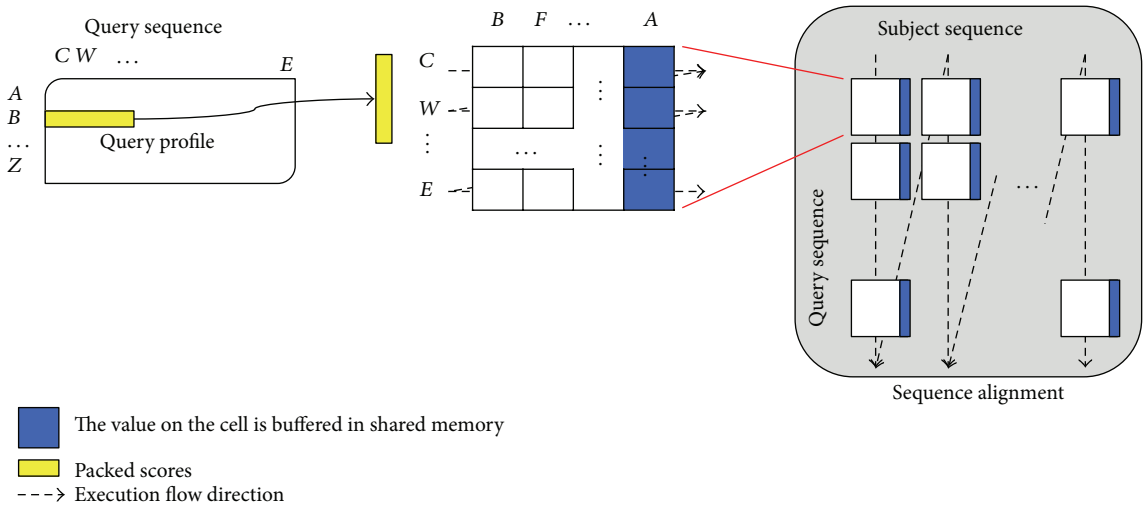


Figure 3: The overview of the proposed mapping algorithm.

query sequence specified by users. Since the shared memory is rather limited, the query sequence length cannot be too long for applying our method.

Since the query profile packs the scores of every four continuous residues and saves the packed scores on texture memory, we can get four scores whenever texture memory is accessed. The fetched four scores have to be saved in four registers because these scores are for the calculation of the four consecutive cells on the same column inside a tile. Consequently, P should be a multiple of four. The pseudo code of our method is shown in Pseudocode 1, where P and K are set to four to reduce the pressure on the register requirement.

4. Results and Discussion

4.1. Analysis of Experimental Platforms between Tesla C1060 and Tesla K20. We used CUDA version 5.0 to extend CUDASW++ 2.0 with our proposed method. Two platforms are used to evaluate our proposed method. The Dell Precision T5500 computer workstation is our first experimental platform, consisting of one Intel Xeon CPU and one nVIDIA Tesla C1060 GPU. Another experimental platform is comprised of Intel Core i7 CPU and nVIDIA Tesla K20, where Tesla K20 is based on the new Kepler architecture, providing a 15-time increase in double precision performance compared Tesla C1060. Tesla K20 consists of 13 streaming multiprocessors with 2496 cores totally while Tesla C1060 has 30 streaming multiprocessor with 240 cores totally.

The Kepler architecture employs a new streaming multiprocessor architecture, called SMX, which deliver more processing performance and efficiency. An SMX allows a greater percentage of space to be applied to processing cores versus control logic. In addition, the Kepler architecture simplifies GPU programming by allowing programmers to easily accelerate all parallel nested loops, resulting in a GPU dynamically spawning new threads on its own without going back to the CPU. Finally, the Kepler architecture also uses the Hyper-Q technique to slash CPU idle time by allowing multiple CPU cores to simultaneously utilize a single Kepler GPU.

Table 1: The hardware configuration of the first experimental platform, where Tesla C1060 is included.

Intel Xeon processor E5504		NVIDIA Tesla C1060	
Number of CPUs	1	Number of GPUs	1
Number of processor cores	4	Number of processor cores	240
Clock speed	2 GHz	Clock speed	1.3 GHz
Memory size	6 GB	Memory size	4 GB
Memory types	DDR3 800	Memory types	GDDR3
Cache	4 MB	Memory clock	800 MHz

Table 2: The hardware configuration of the second experimental platform, where Tesla K20 is included.

Intel Core i7-4790		NVIDIA Tesla K20	
Number of CPUs	4	Number of GPUs	1
Number of processor cores	8	Number of processor cores	2496
Clock speed	3.6 GHz	Clock speed	0.71 GHz
Memory size	8 GB	Memory size	4.8 GB
Memory types	DDR3 1600	Memory types	GDDR5
Cache	8 MB	Memory clock	2600 MHz

The hardware configurations of Tesla C1060 and K20 are shown in Tables 1 and 2, respectively. The operating system installed is Linux and its version is Ubuntu 12.04 64-bit. The BLOSUM64 protein sequences database is used for the performance evaluation. Moreover, the short query sequences are also from the BLOSUM 64.

4.2. Performance Evaluation on Tesla C1060. We present the speedup of our method in the following, where the speedup

```
/* * * * * * * * * * * * * * * *ASSUMPTION* * * * * * * * * * * * * * * *
Query sequence length → qlen, aligned to 4 bytes and padded with dummy residues
Subject sequence length → dblen, aligned to 4 bytes and padded with dummy residues
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * */
for (j = 1; j ≤ dblen; j += 4){
  Initialize all the relevant variables;
  Load the packed 4 residues between j and j + 3 from texture memory to register;
  for (i = 1; i ≤ qlen; i += 4){
    Get the j to j + 3 residues of the subject sequence from register;
    Load substitution scores for cells (i, j) to (i + 3, j + 3) from query profile;
    for (k = 0; k < 4; k++){
      Load H and F values of the cell (i + k, j − 1) from shared memory;
      Compute the H, E and F values of the cells from (i + k, j) to (i + k, j + 3),
      and calculate the maximum score;
      Save H and F values of the cell (i + k, j + 3) to shared memory;
    }
  }
}
```

Pseudocode 1: The pseudocode of the scoring function in our proposed method.

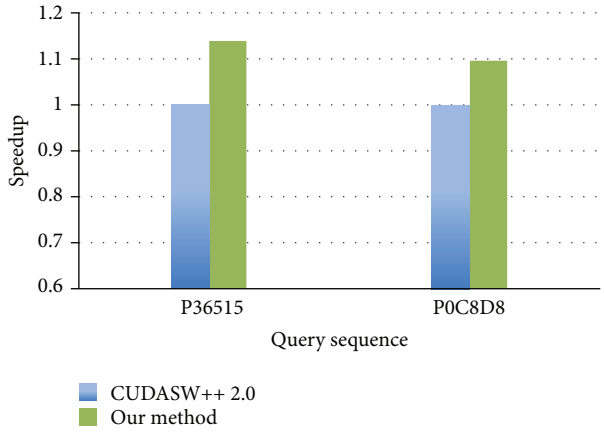


Figure 4: The speedup of our method over CUDASW++ 2.0 on Tesla C1060 with 256 blocks.

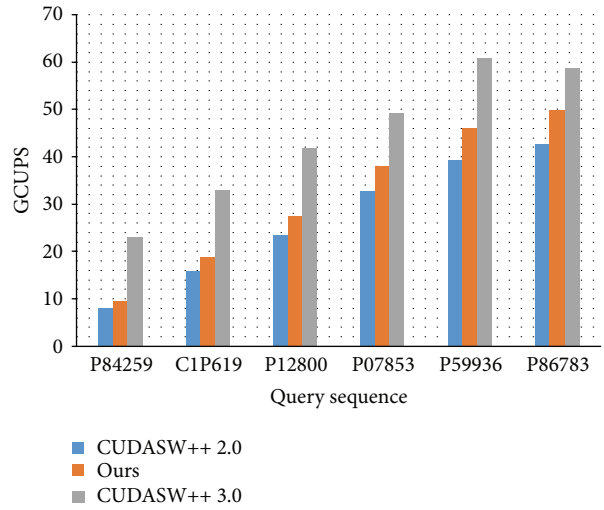


Figure 5: The GCUPS comparison of our method with CUDASW++ 2.0 and CUDASW++ 3.0 on Tesla K20.

is to divide the execution time of a method by the execution time of CUDASW++ 2.0. The performance improvement of our method over CUDASW++ 2.0 is shown in Figure 4. The C1060 has 16 K-byte shared memory. Because there are 256 threads in each block, each thread is assigned with 64-byte, that is, 16-word, shared memory for storing the matrices H, E, and F. The best speedup is about 1.14 when the query sequence is P36515 consisting of four amino acids.

4.3. Performance Evaluation with CUDASW++ 2.0 and CUDASW++ 3.0 on K20. We further compared with different methods on Tesla K20 with 64 blocks and 64 threads. On the Kepler GPU, K20, the space of shared memory per streaming multiprocessor is much larger than that on Tesla C1060. This characteristic can store more spilled register values for CUDASW++ series and thus reduce the frequency of swapping some 10 shared memory values out to/in from the slow global memory. The feature enables our method to process longer query sequences with more parallel threads per block. The GCUPS comparison between our method and CUDASW++ 2.0 as well as CUDASW++ 3.0 is shown in Figure 5, where GCUPS stands for giga cell updates per second. Our method outperforms CUDASW++ 2.0 for all of the query sequences because ours can fully utilize the shared memory without the need of swapping data between shared memory and global memory. When the query sequence length becomes larger, our method can provide more performance improvement. The reason is because CUDASW++ 2.0 required more data swapping between shared memory and global memory when processing a longer query sequence.

However, our method could not outperform CUDASW++ 3.0 because the latter performed the sequence alignment by coupling both the compute powers of CPU and

GPU while our method was built upon CUDASW++ 2.0, which utilized the GPU compute power only. That is, the subject sequences were divided and allocated to CPU and GPU according to each individual compute power. On the GPU side, CUDASW++ 3.0 used PTX assembly instructions to implement the key recurrence equation and procedure of finding the optimal local alignment score, where every assembly instruction operated on quads of 8-bit signed values, corresponding four independent alignments. The idea of our proposed method can be applied to CUDASW++ 3.0 in the future to accelerate the processing of alignments with short query sequences for the intertask kernel.

CUDASW++ 3.0 used local memory to buffer one row of matrices H and E while we used shared memory instead for the buffering because the access latency of local memory is much longer than that of shared memory. However, the shared memory space is so much smaller than the local memory space that it is impossible to only use shared memory for all the buffering without the help of global memory or local memory if sequences are too long. That is why our method is applicable for those alignments involving short query sequences only. Our method can be integrated into the CUDASW++ 3.0 package as an execution option and the length of the input query sequence is used to determine which method will be invoked for the alignments. If the length is not too long to fit all the required buffering into shared memory, our method is invoked. Otherwise, the original CUDASW++ 3.0 is invoked instead. Our method was originally designed based on CUDASW++ 2.0. To integrate our method into CUDASW++ 3.0, the PTX assembly instructions can be used in our proposed algorithm for further performance improvement.

4.4. Performance Evaluation of Different Threads and Blocks. This subsection explores the influence of the numbers of threads and blocks. We take the query sequence, P86783, for the following study. First, we set the number of the blocks as 64 and change the number of threads, as shown in Figure 6. When the number of threads is increased, our approach and CUDASW++ 2.0 obtained almost the same GCUPS while CUDASW++ 3.0 has higher performance. When the number of threads per block becomes larger, the length deviation of the subject sequences per block becomes higher, resulting in poorer load balance between threads in the same block. Moreover, the amount of shared memory allocated to each thread is reduced when more threads in a block contend for the shared memory. On the other hand, more subject sequences per block can be aligned concurrently. For CUDASW++ 3.0, it adopts advanced scheduling designed especially for Kepler architecture, which prefers more threads per block.

Next, we take the query sequence, P86783, to investigate the influence of the number of blocks, as shown in Figure 7. The number of threads per block is set to 64. When there are more blocks, it means that the total number of threads in a grid is increased. Consequently, the number of subject sequences allocated to each thread is decreased. On the other hand, we cannot run more than one block at the same time on any streaming multiprocessor since each block required almost all the shared memory space in its resident streaming multiprocessor. As a result, increasing the number of blocks incurs higher overhead for context switching between blocks.

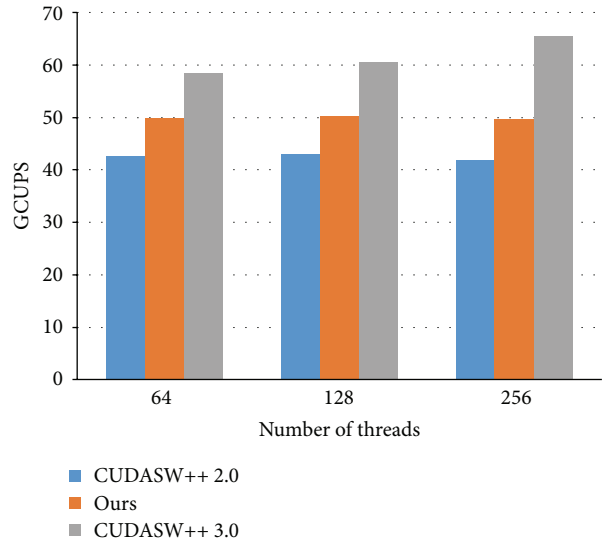


Figure 6: The performance analysis of 64 blocks based on different number of threads.

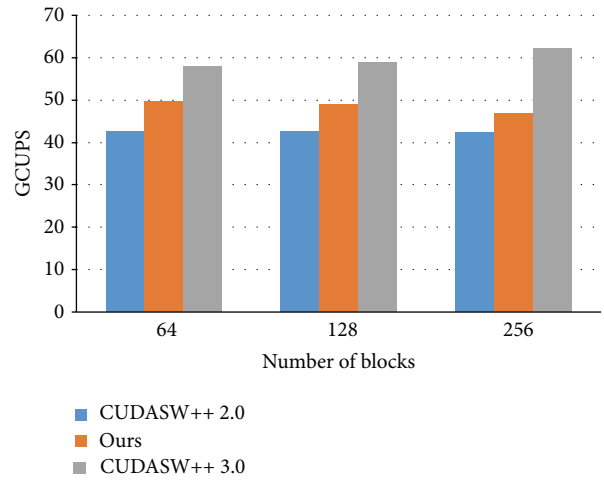


Figure 7: The performance analysis of 64 threads based on different number of blocks.

5. Conclusions

The inexpensive general-purpose GPUs give engineers a great choice to accelerate time-consuming applications. In this work, we discussed how to use NVIDIA GPUs to implement the Smith-Waterman database search algorithm. We have added our method to the advanced package of CUDASW++ 2.0 as an option of execution. When users input a query sequence, the extended package will determine how to run the query based on the length of the query sequence as well as the space of shared memory per streaming multiprocessor. If the query sequence length is short with the calculation of

the available shared memory per streaming multiprocessor, the extended package will use our method to run the query. Otherwise, the original CUDASW++ 2.0 will be used. Our idea can be applied to CUDASW++ 3.0 to improve the intertask kernel in the future.

We have evaluated our method on Tesla C1060 and K20 using the benchmark BLUSOM64. Further, we analyze the performance on different number of threads and blocks. The results suggested that the proposed method may improve Smith-Waterman algorithm on CUDA-enabled GPUs in proper allocation of block and thread numbers.

Conflict of Interests

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

Authors' Contribution

Chao-Chin Wu, Liang-Tsung Huang, and Lien-Fu Lai contributed to the algorithm. Yun-Ju Li implemented the algorithm and carried out experiments. Chao-Chin Wu and Liang-Tsung Huang analyzed the results and wrote the report. Chao-Chin Wu supervised the work. All authors read, edited, and approved the final paper.

Acknowledgments

The authors would like to thank the Ministry of Science and Technology (the successor to National Science Council), Taiwan, for financially supporting this research under Contract no. NSC102-2221-E-018-014 and MOST104-3011-E-018-001. In addition, thanks are due to Jr-Wei Li for helping in conducting experiments.

References

[1] D. M. Mount, *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA, 2nd edition, 2004.

[2] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[3] O. Gotoh, "An improved algorithm for matching biological sequences," *Journal of Molecular Biology*, vol. 162, no. 3, pp. 705–708, 1982.

[4] M. O. Dayhoff, R. Schwartz, and B. C. Orcutt, "A model of evolutionary change in proteins," *Atlas of Protein Sequence and Structure*, vol. 5, supplement 3, pp. 345–358, 1978.

[5] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 22, pp. 10915–10919, 1992.

[6] D. J. Lipman and W. R. Pearson, "Rapid and sensitive protein similarity searches," *Science*, vol. 227, no. 4693, pp. 1435–1441, 1985.

[7] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, no. 8, pp. 2444–2448, 1988.

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

[9] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[10] T. Oliver, B. Schmidt, D. Nathan, R. Clemens, and D. Maskell, "Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW," *Bioinformatics*, vol. 21, no. 16, pp. 3431–3432, 2005.

[11] T. F. Oliver, B. Schmidt, and D. L. Maskell, "Reconfigurable architectures for bio-sequence database scanning on FPGAs," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 52, no. 12, pp. 851–855, 2005.

[12] I. T. S. Li, W. Shum, and K. Truong, "160-fold acceleration of the Smith-Waterman algorithm using a field programmable gate array (FPGA)," *BMC Bioinformatics*, vol. 8, article 185, 2007.

[13] M. Farrar, "Striped Smith-Waterman speeds database searches six times over other SIMD implementations," *Bioinformatics*, vol. 23, no. 2, pp. 156–161, 2007.

[14] T. Rognes, "Faster smith-waterman database searches with inter-sequence SIMD parallelisation," *BMC Bioinformatics*, vol. 12, article 221, 2011.

[15] A. Wirawan, C. K. Kwoh, N. T. Hieu, and B. Schmidt, "CBESW: sequence alignment on the playstation 3," *BMC Bioinformatics*, vol. 9, article 377, 2008.

[16] A. Szalkowski, C. Ledergerber, P. Krähenbühl, and C. Dessimoz, "SWPS3—fast multi-threaded vectorized Smith-Waterman for IBM Cell/B.E. and x86/SSE2," *BMC Research Notes*, vol. 1, article 107, 2008.

[17] M. S. Farrar, *Optimizing Smith-Waterman for the Cell Broadband Engine*, 2008, http://cudasw.sourceforge.net/sw-cellbe.pdf.

[18] Y. Liu, D. L. Maskell, and B. Schmidt, "CUDASW++: optimizing Smith-Waterman sequence database searches for CUDA-enabled graphics processing units," *BMC Research Notes*, vol. 2, article 73, 2009.

[19] Y. Liu, B. Schmidt, and D. L. Maskell, "CUDASW++2.0: enhanced Smith-Waterman protein database search on CUDA-enabled GPUs based on SIMT and virtualized SIMD abstractions," *BMC Research Notes*, vol. 3, article 93, 2010.

[20] Y. Liu, A. Wirawan, and B. Schmidt, "CUDASW++ 3.0: accelerating Smith-Waterman protein database search by coupling CPU and GPU SIMD instructions," *BMC Bioinformatics*, vol. 14, article 117, 2013.

[21] S. A. Manavski and G. Valle, "CUDA compatible GPU cards as efficient hardware accelerators for smith-waterman sequence alignment," *BMC Bioinformatics*, vol. 9, supplement 2, article S10, 2008.

[22] L. Ligowski and W. Rudnicki, "An efficient implementation of Smith Waterman algorithm on GPU using CUDA, for massively parallel scanning of sequence databases," in *Proceedings of the 23rd IEEE International Parallel & Distributed Processing Symposium (IPDPS '09)*, pp. 1–8, May 2009.

[23] A. Khajeh-Saeed, S. Poole, and J. B. Perot, "Acceleration of the Smith-Waterman algorithm using single and multiple graphics processors," *Journal of Computational Physics*, vol. 229, no. 11, pp. 4247–4258, 2010.

[24] J. Blazewicz, W. Frohmberg, M. Kierzynka, E. Pesch, and P. Wojciechowski, "Protein alignment algorithms with an efficient backtracking routine on multiple GPUs," *BMC Bioinformatics*, vol. 12, article 181, 2011.

[25] D. Hains, Z. Cashero, M. Ottenberg, W. Bohm, and S. Rajopadhye, "Improving CUDASW, a parallelization of smith-waterman for CUDA enabled devices," in *Proceedings of the 25th IEEE International Parallel and Distributed Processing Symposium, Workshops and Phd Forum (IPDPSW' 11)*, pp. 490–501, May 2011.

[26] E. F. D. O. Sandes and A. C. M. A. de Melo, "Retrieving smith-waterman alignments with optimizations for megabase biological sequences using GPU," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 5, pp. 1009–1021, 2013.

[27] M. Noorian, H. Pooshfam, Z. Noorian, and R. Abdullah, "Performance enhancement of Smith-Waterman algorithm using hybrid model: comparing the MPI and hybrid programming paradigm on SMP clusters," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pp. 492–497, October 2009.

[28] CUDA GPUs, https://developer.nvidia.com/cuda-gpus.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 731938, 17 pages
http://dx.doi.org/10.1155/2015/731938

Research Article

Similarities in Gene Expression Profiles during *In Vitro* Aging of Primary Human Embryonic Lung and Foreskin Fibroblasts

Shiva Marthandan,[1] Steffen Priebe,[2] Mario Baumgart,[1] Marco Groth,[1] Alessandro Cellerino,[1,3] Reinhard Guthke,[2] Peter Hemmerich,[1] and Stephan Diekmann[1]

[1]*Leibniz-Institute for Age Research-Fritz Lipmann Institute e.V. (FLI), 07745 Jena, Germany*
[2]*Leibniz Institute for Natural Product Research and Infection Biology-Hans-Knöll-Institute e.V. (HKI), 07745 Jena, Germany*
[3]*Laboratory of Neurobiology, Scuola Normale Superiore, University of Pisa, 56126 Pisa, Italy*

Correspondence should be addressed to Shiva Marthandan; smarthandan@fli-leibniz.de

Received 23 January 2015; Revised 14 June 2015; Accepted 22 June 2015

Academic Editor: Hesham H. Ali

Replicative senescence is of fundamental importance for the process of cellular aging, since it is a property of most of our somatic cells. Here, we elucidated this process by comparing gene expression changes, measured by RNA-seq, in fibroblasts originating from two different tissues, embryonic lung (MRC-5) and foreskin (HFF), at five different time points during their transition into senescence. Although the expression patterns of both fibroblast cell lines can be clearly distinguished, the similar differential expression of an ensemble of genes was found to correlate well with their transition into senescence, with only a minority of genes being cell line specific. Clustering-based approaches further revealed common signatures between the cell lines. Investigation of the mRNA expression levels at various time points during the lifespan of either of the fibroblasts resulted in a number of monotonically up- and downregulated genes which clearly showed a novel strong link to aging and senescence related processes which might be functional. In terms of expression profiles of differentially expressed genes with age, common genes identified here have the potential to rule the transition into senescence of embryonic lung and foreskin fibroblasts irrespective of their different cellular origin.

1. Introduction

Cellular senescence is a terminal phase observed towards the end of a primary human fibroblast cell population after numerous cell divisions; it is considered to be the cellular aging process. Cellular senescence occurs either naturally or stress induced; that is, cells stop dividing after a finite number of cell divisions (termed "replicative senescence"), reaching the final cell cycle arrested state called the "Hayflick limit" [1]. The process of senescence is associated with a number of phenotypes; in general, the integrity and function of tissues decline, resulting in the body being susceptible to diseases associated with age [2, 3]. Key factors driving cellular senescence are induced increase in Cyclin dependent kinase inhibitors (CDKIs) [4], oxidative stress [5], and DNA damage [6, 7]. In senescence, despite their viability and active metabolism, cells are resistant to mitogenic or apoptotic stimuli [8, 9]. On the one hand, cellular senescence results in irreversible growth arrest, limiting the proliferation of damaged cells susceptible to neoplastic transformation resulting in a decreased incidence of cancer. However, on the other hand, senescence results in *in vivo* aging, weakening the function and renewal of stem cells [10]. Markers are able to identify cellular senescence *in vitro* and *in vivo*: enlarged cell morphology, increase in amount of cellular debris, changes in chromatin structure, increase in Cyclin dependent kinase inhibitors (CDKIs) expression, presence of senescence associated secretory phenotype (SASP), and senescence associated ß-galactosidase (SA ß-Gal) [11–13]. DNA damage response and the p53-p21 and p16-pRb pathways are crucial for senescence induction [14], together with additional pathways including telomere uncapping, DNA damage (UV, ionizing radiation, and chemicals), cytoskeletal genes, the interferon pathway, nutrient imbalances, oncogenic activities, and oxidative stress [7, 15, 16]. In primates,

the percentage of senescent skin fibroblasts increases with age *in vivo* [17]. Here, we therefore used primary human fibroblasts [9, 13, 18] as our model system.

Recently, we identified individual gene expression patterns during replicative senescence among five fibroblast cell lines of different cell origins [13]. In this study, we determined mRNA expression changes during different stages of their lifespan in two fibroblast cell lines of different cell origin. We analyzed the transcriptome, determined by RNA-seq, at five separate population doublings (PDs) between young and senescent embryonic lung (MRC-5) and foreskin (HFF) fibroblasts. Using both molecular and systems biology approach, we studied the growth pattern of the two fibroblast cell lines in detail. By comparing fibroblasts from two different origins we were able to determine either mRNA changes specific for one of the cell lines or common transcriptomic patterns which underlie the process of replicative senescence.

2. Materials and Methods

2.1. Cell Lines. Primary human fibroblasts (MRC-5, primary cells, 14-week-gestation male, fibroblasts from normal lung, normal diploid karyotype) were obtained from ATCC (LGC Standards GmbH, Wesel, Germany). Human foreskin fibroblasts (HFFs; primary cells, fibroblasts from foreskin, normal diploid karyotype) cells were kind gifts of T. Stamminger (University of Erlangen [19]).

2.2. Cell Culture. Primary human fibroblast cells were cultured in Dulbeccos Modified Eagle's Low Glucose Medium (DMEM) with L-glutamine (PAA Laboratories, Pasching, Austria), supplemented with 10% fetal bovine serum (FBS) (PAA Laboratories) under normal air conditions in a 9.5% CO_2 atmosphere at 37°C. The cells were subcultured by removing the remaining medium followed by washing in 1x PBS (pH 7.4) (PAA Laboratories) and detachment using trypsin/EDTA (PAA Laboratories). Primary fibroblasts were subcultured in a 1 : 4 (= 2 population doublings (PDs)) or 1 : 2 (= 1 PD) ratio. For stock purposes, cryoconservation of the cell lines at various PDs was undertaken in cryoconserving medium (DMEM + 10% FBS + 5% DMSO). Cells were immediately frozen at −80°C and stored for two to three days. Afterwards, cells were transferred to liquid nitrogen for long time storage. Refreezing and rethawing was not performed to avoid premature senescence [20].

A vial of each of the two fibroblast cell lines (MRC-5 and HFF) was obtained and maintained in culture from an early PD. On obtaining enough stock on confluent growth of the fibroblasts in 75 cm^2 flasks, cells were subcultured into three separate 75 cm^2 flasks and were passaged until they were senescent in culture. At five different time points of the fibroblast's span in culture (MRC-5 = PDs 32, 42, 52, 62, and 72 and HFFs = PDs 16, 26, 46, 64, and 74), the total RNA was extracted and used for high-throughput sequencing.

2.3. Detection of Senescence Associated ß-Galactosidase (SA ß-Gal). The SA β-Gal assay was performed as described by [11] at each of the five PDs in both MRC-5 and HFF. Paired two-sample type 2 Student's t-tests assuming equal variances were done to examine the values obtained from SA ß-Gal assay for statistical significance [9].

2.4. Western Blotting. The protocol was carried out as explained in [9, 21]. The optimal concentration of all primary antibodies was estimated in primary human fibroblasts. Primary antibodies are as follows: anti-p21 mouse antibody (OP64; Calbiochem; dilution 1 : 200), anti-p15 rabbit antibody (4822; Cell Signaling Technology; 1 : 250), anti-p16 mouse antibody (550834; BD Pharmingen; 1 : 200), anti-p27 rabbit antibody (sc-528; Santa Cruz; 1 : 200), anti-Cyclin B1 mouse antibody (CCNB1; ab72; Abcam; 1 : 1000), anti-Eg5 rabbit antibody (KIF11; ab61199; Abcam; 1 : 500), anti-Histone H1.2 rabbit antibody (HIST1H1C; ab17677; Abcam; 1 : 1000), anti-ID3 mouse antibody (ab55269; Abcam; 1 : 100), anti-Cathepsin K rabbit antibody (CTSK; ab19027; Abcam; 1 : 50), anti-DKK3 goat antibody (ab2459; Abcam; 1 : 5000), anti-TMEM47 rabbit antibody (SAB1104840; SIGMA-Aldrich; 1 : 250), anti-IGFBP7 rabbit antibody (ab74169; Abcam; 1 : 500), anti-IGFBP2 rabbit antibody (ab91404; Abcam; 1 : 500), anti-MMP3 rabbit antibody (ab53015; Abcam; 1 : 200), anti-Thymosin beta 10 rabbit antibody (TMSB10; ab14338; Abcam; 1 : 10000), anti-Egr1 mouse antibody (ab55160; Abcam; 1 : 100), anti-RPS23 mouse antibody (ab57644; Abcam; 1 : 200), anti-LIF mouse antibody (SAB1406083; SIGMA-Aldrich; 1 : 100), anti-FBL rabbit antibody (SAB1101099; SIGMA-Aldrich; 1 : 500), anti-Id1 rabbit antibody (ab52998; Abcam; 1 : 500), anti-IL11 rabbit antibody (ab76589; Abcam; 1 : 500), anti-CLDN11 rabbit antibody (HPA013166; SIGMA-Life Sciences; 1 : 50), anti-NADH Dehydrogenase subunit 6 rabbit antibody (MT-ND6; ab81212; Abcam; 1 : 1000), anti-MT-ND5 rabbit antibody (ab83985; Abcam; 1 : 500), anti-Granulin rabbit antibody (GRN; ab108608; Abcam; 1 : 1000), anti-Cyclin D1 rabbit antibody (CCND1; 2922; Cell Signaling; 1 : 500), anti-Cyclin D2 mouse antibody (CCND2; ab3085; Abcam; 1 : 500), anti-Cyclin A2 rabbit antibody (CCNA2; NBP1-31330; Novus Biologicals; 1 : 1000), anti-Wnt16 rabbit antibody (ab109437; Abcam; 1 : 500), anti-Cystatin C rabbit antibody (CST3; ab109508; Abcam; 1 : 10000), anti-MOXD1 mouse antibody (SAB1409086; SIGMA-Aldrich; 1 : 200), anti-PERP rabbit antibody (ab5986; Abcam; 1 : 500) and anti-tubulin mouse antibody (T-9026; SIGMA-Aldrich; 1 : 5000). After development of film in the Western Blots procedure, intensity of the signals was quantified using Metamorph software [22]. The signal intensity values were examined for statistical significance using unpaired two-tailed two-sample Student's t-tests assuming unequal variances.

2.5. RNA Extraction. Total RNA was isolated using Qiazol (Qiagen) according to the manufacturer's protocol, with modifications as explained in [9].

2.6. Quantitative Real-Time PCR. Real-time PCR was performed using CFX384 thermocycler Biorad and Quantitect PCR system (Qiagen) as described earlier in [23]. Three reference genes (GAPDH, ACTB, and RAB10) were used for

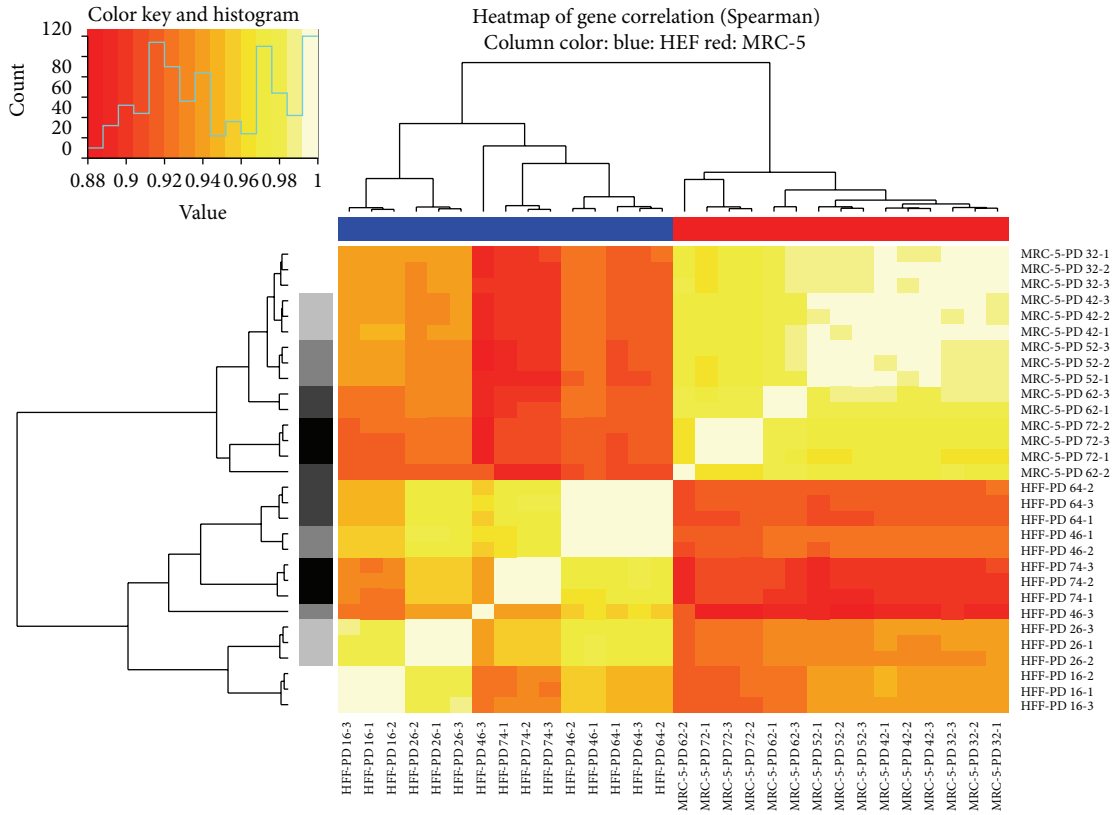


Figure 1: Heatmap showing Spearman correlation values computed for all expressed genes and all 30 samples sequences for this study. The histogram on top left denotes the distribution of the correlation values as well as the colors for each value. The dendrograms on top and left of the histogram are identical but show different colors at their leaves: top: color denotes the cell line (blue: HFF; red: MRC-5); left: color denotes the PD (bright to dark → young to old). Samples are clustered first according to cell line and second according to PD. Highest correlations were found between the three replicates (light areas on the main diagonal).

normalization of the CT values. Since our RNA-seq results revealed a stable expression of RAB10 for both cell lines across the PDs, it was selected as reference gene. An unpaired two-tailed two-sample Student's t-tests assuming unequal variances was used for examination for statistical significance based on the ΔCT values.

2.7. RNA Sequencing. For quality check, total RNA was analyzed using Agilent Bioanalyzer 2100 (Agilent Technologies) and RNA 6000 Nano Kit (Agilent) to ensure appropriate RNA quality in terms of degradation. The RNA integrity number (RIN) varies between 8 and 10 with an average of around 9.65. Total RNA was used for Illumina library preparation and next-generation sequencing [24]. About 2.5 μg total RNA was used for indexed library preparation using Illumina's TruSeq RNA Sample Prep Kit v2 following the manufacturer's instruction. Libraries were pooled and sequenced (4 samples per lane) using a HiSeq2000 (Illumina) in single read mode with 50 cycles using sequencing chemistry v3. Sequencing resulted in approximately 43 million reads with a length of 50 bp (base pairs) per sample. Reads were extracted in FASTQ format using CASAVA v1.8.2 (Illumina).

2.8. RNA-seq Data Analysis. Raw sequencing data were received in FASTQ format. Read mapping was performed using Tophat 2.0.6 [25] and the human genome references assembly GRCh37.66 (http://feb2012.archive.ensembl.org). The resulting SAM alignment files were processed using featureCounts v1.4.3-p1 [26] and the respective GTF gene annotation, obtained from the Ensembl database [27]. Gene counts were further processed using the R programming language [28] and normalized to RPKM values. RPKM values were computed using exon lengths provided by featureCounts and the sum of all mapped reads per sample.

2.9. Sample Clustering and Analysis of Variance. Spearman correlation between all samples was computed in order to examine the variance and the relationship of global gene expression across the samples, using genes with raw counts larger than zero. Correlation values were visualized using a heatmap (Figure 1). Additionally, principal component analysis (PCA) was applied using the log 2 RPKM values for genes with raw counts larger than zero. Results were visualized in a three-dimensional scatterplot (Figure 2).

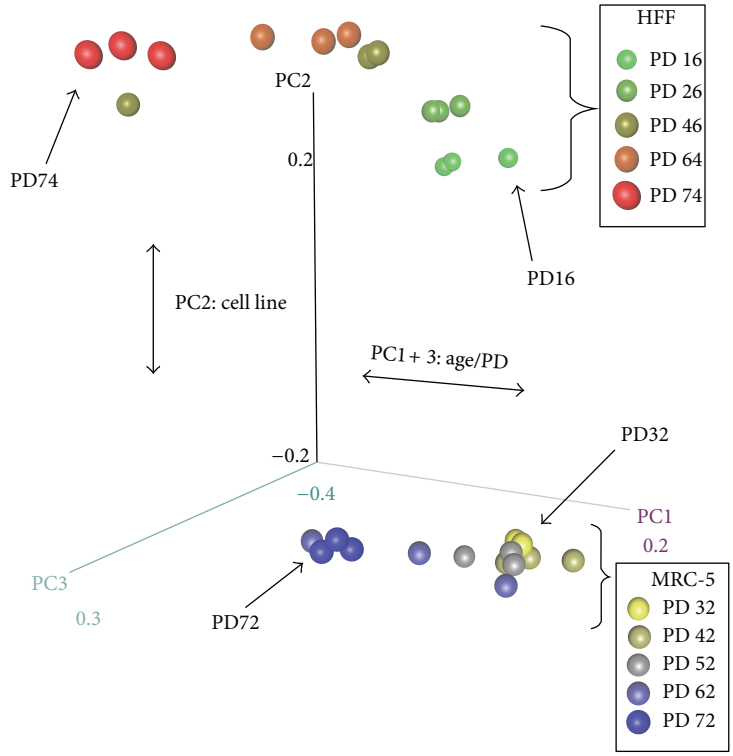


FIGURE 2: 3D PCA plot for 30 samples (2 cell lines, 5 different PDs, triplicates). Both cell lines are clearly separated by PC2. The effect of aging is partly explained by PC1 and PC3. Colors: yellow to blue: young to old MRC-5 cells; green to red: young to old HFF cells.

2.10. Detection of Differential Expression. The Bioconductor packages *DESeq* 1.10.4 [29] and *edgeR* 3.4.2 [30] were used to identify differentially expressed genes. Both packages provide statistics for determining of differential expression in digital gene expression data using a model based on the negative binomial distribution. The nonnormalized gene counts have been used here, since both packages include internal normalization procedures. The resulting p values were adjusted using Benjamini and Hochberg's approach for controlling the false discovery rate (FDR) [31]. Genes with an adjusted p value < 0.05 found by both packages were assigned as differentially expressed. Since large sets of DEG were found more strict selection cutoffs have been used: adjusted p value < 0.01 (by both packages) and absolute log 2 fold-changes > 1. See Supplemental Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/731938 for complete test results.

2.11. Comparison of RNA-seq with qRT-PCR and Protein Expression. Correlation analysis was performed using all 15 samples (3 replicates for each of the 5 PDs) for MRC-5 and HFF, respectively. Spearman correlation coefficients were estimated using the RPKM values (RNA-seq data) and $2^{-\Delta CT}$ values (qRT-PCR data). For comparison of RNA-seq with Western Blot data, only the first and the last PD were used. Log 2 fold-changes were calculated based on RPKM values (RNA-seq data), $2^{-\Delta\Delta CT}$ ratios (rRT-PCR), and protein expression ratios (Western Blots).

2.12. Clustering of Expression Profiles. Genes were clustered according to their temporal profiles using a fuzzy c-means algorithm. We used the function *cmeans* from the package *e1071* 1.6-2 of the R programming language. Parameters were defined as $m = 1.2$, iter.max = 500, d.obj.fun = 10^{-8}. The number of trials for the fuzzy algorithm was set to 30. The optimum number of clusters was determined using a combination of several cluster validation indexes as described by [32]. See Supplemental Table 2 for detailed assignment of the genes to clusters.

2.13. Functional Enrichment Analysis. Singular gene set enrichment analysis was performed using FungiFun2 [33] for selected sets of genes based on the clustering results. Although FungiFun2 is mainly suited for fungal gene enrichment analysis annotation for human genes is included as well and was recently updated. Default parameters were used while significant Gene Ontology (GO) terms and KEGG pathways were selected according to FDR corrected p values < 0.05. Complete lists of GO-terms and KEGG pathways are available from Supplemental Table 3. The list of GO-terms was further summarized using TreeMaps of the REVIGO online tool [34]. Default parameters and GO term with adjusted p values were used as input.

2.14. Monotonically Expressed Genes. In order to identify genes that change their expression levels monotonically with age, we calculated the Spearman correlation coefficient $c(i)$

of each gene's temporal profile with the linearly increasing curve $f(x) = x$. In order to incorporate the replicates at each time point, we repeated the calculations by randomly sampling over the replicates at each time point and by calculating an average correlation coefficient from the resampled curves afterwards. p values $p(i)$ were computed using the R base function *cor.test*. We used the calculated correlation coefficient $c(i)$ of gene i with the linear increasing curve as a criterion to split the genes into the following three groups: if $c(i) > 0$ and $p(i) < 0.05$, we considered a gene to be monotonically increasing with age, if $c(i) < 0$ and $p(i) < 0.05$, the gene was considered to be monotonically decreasing with age, and if $p(i) > 0.05$, the expression of the corresponding gene was considered nonuniformly [35]. See Supplemental Table 4 for detailed test results.

2.15. Functional Association Networks. Gene symbols were used as input for functional association network creation using the STRING database [36], Cognoscente [37], and GeneMANIA [38] online tools. For STRING we used "multiple names" input and selected "Homo sapiens" as organism. In Cognoscente we selected "Human" and "Radius = 0 with intermediates" as input parameters in addition to the list of genes. The GeneMANIA network was created using default settings. The resulting networks are shown in Figure 9 and Supplemental Figure 6.

3. Results and Discussion

We studied the growth of two primary human fibroblast cell lines, MRC-5 and HFF, throughout their span in culture from an early PD until they achieved senescence at late PDs. Analysis of their growth behaviour (Supplemental Figure 1A) and their entry into senescence (Supplemental Figure 1B), measured by the induction of SA β-Gal, revealed a cell line specific transition into senescence of these two fibroblasts (MRC-5 derived from embryonic lung and HFFs derived from human foreskin). Fibroblast cell line specific growth has been observed by us before [13, 39]. Total RNA was extracted at five different time points of the fibroblasts span in culture and was subjected to high-throughput RNA sequencing (RNA-seq).

3.1. Global Expression Profiles Cluster according to Cell Line and Age. Overall, the RNA-seq data of this study comprise 30 samples: 15 samples for each cell line (HFF and MRC-5), consisting of five different PDs, each with three biological replicates. For each sample, mapping and counting resulted in 56,299 raw gene count expression values (using Ensembl gene annotation). The largest group of these genes (21,226) belongs to the group of protein coding genes. For all the 30 samples, 19,237 genes have raw gene counts larger than zero; these genes were considered for further analysis.

First, we studied primary clustering of the global gene expression. We therefore created a heatmap showing the Spearman correlation for all 30 samples using the nonzero genes (Figure 1). In this heatmap, both cell lines were clearly separated. In eight out of ten cases, the three values of the replicates were clustered together, showing the good quality of the data and low noise between the replicates. Next, we applied principal component analysis (PCA) to further investigate the effect of aging in the individual cell lines. Figure 2 shows the first three principal components which explain ~97% of the variances in a three-dimensional plot. MRC-5 and HFF again were clearly separated (by PC2) and the effect of aging was covered by PC1 and PC3, with a larger separation between young and old HFF compared to MRC-5. Already at this global level, similarity between both cell lines is perceptible, since young and senescent samples are grouped concordantly.

Table 1: Number of DEG across different PD in MRC-5 and HFF for two significance criteria. See Supplemental Table 1 for detailed test results.

Comparison	Number of DEG (FDR < 0.05)	Number of DEG (FDR < 0.01; | log 2FC| > 1)
MRC5: PD 32 to PD 42	2,050	131
MRC5: PD 42 to PD 52	1,248	185
MRC5: PD 52 to PD 62	2,617	773
MRC5: PD 62 to PD 72	4,582	1,516
MRC5: PD 32 to PD 72	8,992	2,117
HFF: PD 16 to PD 26	7,873	1,083
HFF: PD 26 to PD 46	2,228	1,366
HFF: PD 46 to PD 64	15	8
HFF: PD 64 to PD 74	7,002	1,553
HFF: PD 16 to PD 74	12,529	4,651

3.2. MRC-5 and HFF Share Common Differentially Expressed Genes Regulated by Aging. Differentially expressed genes (DEG) were identified by comparing all consecutive PDs as well as the first with the last PD in MRC-5 and HFF cells (10 comparisons; Table 1). Figures 3(a) and 3(c) show the absolute number of DEG found as well as the intersection of sets of DEG (indicated by color). Overall, considering all five comparisons in each cell line, we identified more DEG in HFF (14,511) compared to MRC-5 (10,517). Due to the strong effect of aging on gene expression and the large number of detected DEG, more stringent selection cutoffs ($p < 0.01$ and |log 2 fold-change| > 1) were used beyond the standard p value threshold of 0.05 (Figures 3(b) and 3(d)). Figure 3 reveals that DEG were not specific for a certain PD comparison but recurred when later PDs were compared. MRC-5 and HFF shared a large fraction of DEG; only a minor fraction of DEG was identified uniquely in one of the cell lines (bars on the right in Figure 3). This indicates common processes which occur during aging in both cell lines rather than cell line specific changes. Most DEG were found when comparing the first with the last PD, leading to new DEG which had not previously been detected between consecutive PDs (orange and turquoise coloured bars in Figure 3). Both cell lines differed between the absolute number of DEG as well as the increased percentage of DEG for the first two transitions in HFF (PD 16 to PD 26; PD 26 to PD 46). In

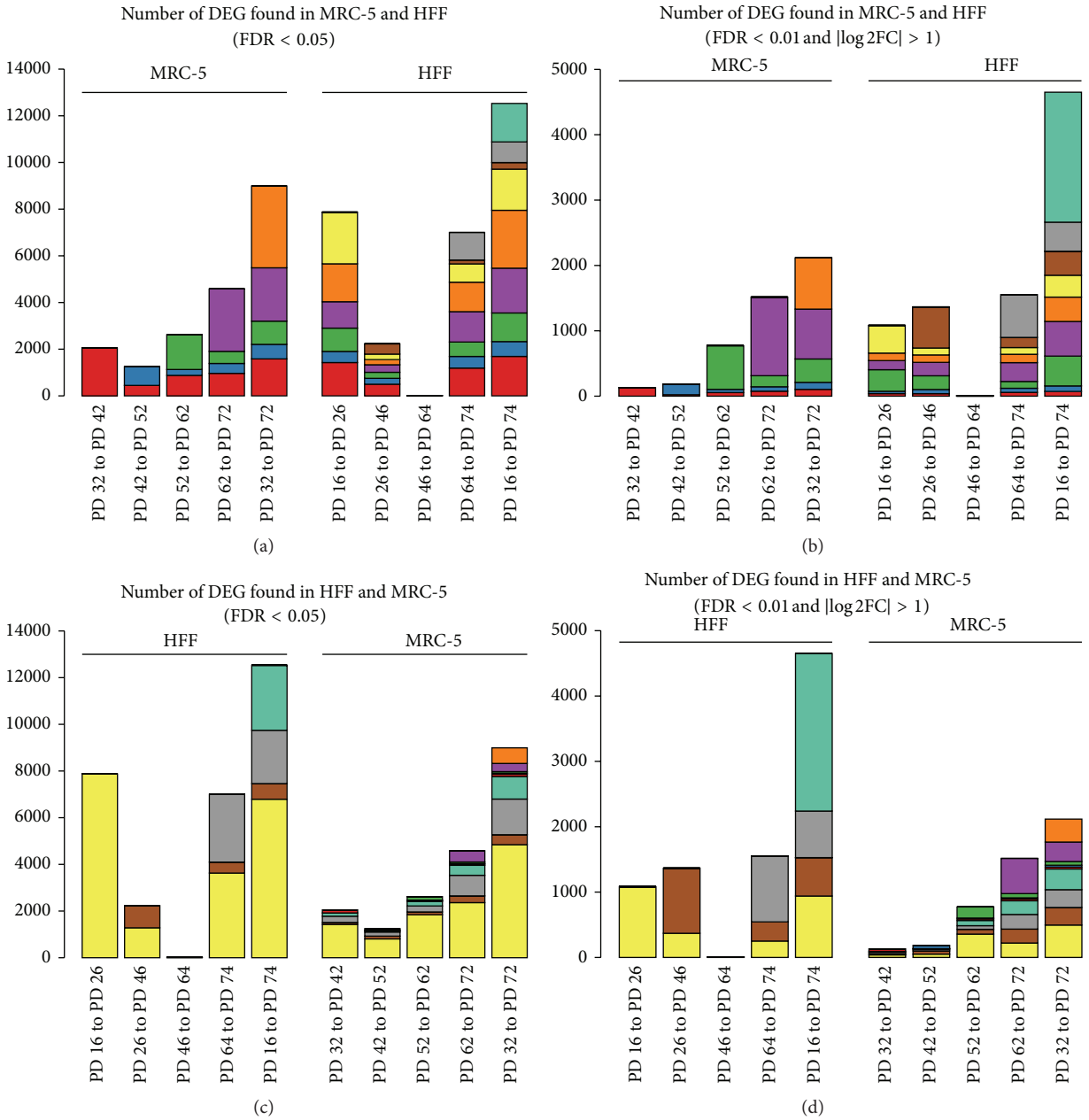


Figure 3: Intersection barplot showing the number of DEG between several consecutive PDs in MRC-5 and HFF for two different cutoffs ((a) + (c): FDR < 0.05; (b) + (d): FDR < 0.01 and |log 2-fold-change| > 1). In each plot, identical colors across different bars indicate the same set of DEG (intersection) while new colors indicate a new set of DEG (from left to right). (a) and (b) indicate genes which were found in MRC-5 and likewise in HFF. (c) and (d) show the same number of DEG but HFF is listed first in order to show the number of genes which were found likewise in MRC-5. For instance in (a), the red colored parts of all bars encode DEG found in MRC-5 between PD32 and PD42 while yellow colored parts of bars denote "new" DEG found in HFF between PD16 and PD26.

MRC-5, most of the changes seemed to occur at late PDs, while HFF cells indicated larger changes already after early PDs. This effect is also perceivable by the distances in the PCA plot (Figure 2).

3.3. High Correlation of RNA-seq with qRT-PCR for Selected DEG. For validation of the RNA-seq data, qRT-PCR was applied. Here, triplicates for all five PDs were measured. Selection of genes was based on the comparison of the first with the last PD, using the strict DEG criteria (p < 0.01 and |log 2 fold-change| > 1), resulting in 2,117 DEG for MRC-5 and 4,651 DEG for HFF (5th and 10th bars in Figures 3(b) and 3(d)). We further filtered the intersection of those two gene sets (1,139) according to common differences in both cell lines. The majority (917) of these DEG were commonly regulated, either up (385) or down (532),

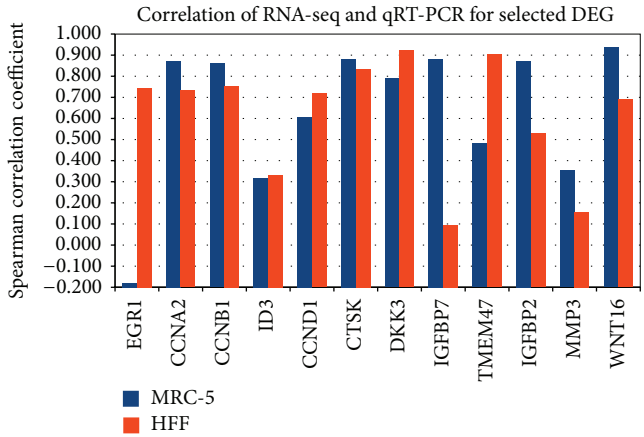


Figure 4: Spearman correlation of RNA-seq and qRT-PCR expression profiles for selected DEG commonly regulated in MRC-5 and HFF. Almost all genes exhibit a positive correlation between both measurement techniques (exception: EGR1 in MRC-5). With age, the first four genes (EGR1-ID3) are downregulated (first PD compared to last PD), while the later are upregulated (CCND1-WNT16).

showing again the similarity of gene expression changes in both cell lines. Overall 12 DEG were selected which either showed strong expression in the RNA-seq data (RPKM > 50; genes: *EGR1*, *CCND1*, *CTSK*, *DKK3*, *IGFBP7*, and *TMEM47*) or were proven to have an established role in cell cycle and senescence pathways (*CCNA2*, *CCNB1*, *ID3*, *IGFBP2*, *MMP3*, and *WNT16*). The minimal RPKM criterion was applied to ensure a strong expression signal in at least one condition for a set of genes. The expression profiles from both measurement techniques were then confronted using Spearman rank correlation in each individual cell line. The results showed high correlation coefficients indicating a good overlap of both measurement techniques and quality of high-throughput gene expression analysis (Figure 4). In 17 out of 24 cases, correlation was larger than 50% (mean correlation of 63%), while only once negative correlation was found (EGR1 in MRC-5).

3.4. Consistent Changes of mRNA and Protein Expression. Although mRNA expression changes are generally considered to consequently lead to corresponding changes in protein levels, correlation between both can be as little as 40%, as observed in large-scale proteome- and transcriptome-profiling experiments [40]. We thus asked if the detected changes of strongly altered DEG correlated with corresponding protein expression levels. Triplicates of the first and last PD were selected for comparison. Gene selection was performed as described above, either by strong expression in RNA-seq (RPKM > 35) or by functional relation to cell cycle and senescence pathways. Overall, 28 DEG (16 down- and 12 upregulated DEG) were selected (the genes mentioned above, validated by qRT-PCR, were included in this set). The results of this comparison showed consistent changes, in terms of their direction of regulation, between mRNA expression, measured by RNA-seq, and protein expression, measured by Western Blots, for all selected genes (Figures 5 and 6). 44 out of the 56 protein fold-changes exhibited significant differences between young and old PDs.

3.5. Common Genes Ruling the Transition into Senescence in MRC-5 and HFF. Then, we asked if common cellular markers are involved in the transition into senescence. We thus studied the genes most differentially expressed with age commonly in MRC-5 and HFF fibroblasts. We noticed that a large number of genes among the most differentially expressed genes belonged to the secretory phenotype (Figure 8(a), as explained in Section 3.7). The list of genes included *CTSK*, normally stimulated by inflammatory cytokines released after tissue injury [41], *GRN*, a previously functionally validated gene responsible for wound healing [42], *CST3*, associated with sarcopenia [43], and *PERP*, a p53 apoptosis effector, the mRNA expression level of which is upregulated in human mesenchymal stem cells [44]. We detected significant upregulation of *IGFBP2* which was found upregulated with senescence in retinal pigment epithelial cells [45, 46] and BJ fibroblasts [47]. *ID1*, *ID3*, *CCNA2,* and *CCNB1* showed significant downregulation with age in our study for both human fibroblasts. Downregulation of *ID1* and *ID3* expression with senescence was detected in BJ foreskin, WS1 fetal skin, and LF1 lung human fibroblasts [48] and of *CCNA2* in IMR-90 and WI-38 [49]. Targeting *CCNB1* expression inhibits proliferation of breast cancer cells [50]. The list of most differentially expressed genes also included *IGFBP7* and *MMP3* which encode protein receptors predominantly located on the cell surface. Both *IGFBP7* and *MMP3* are upregulated with senescence in human melanocytes [51–54]. Recently we found that overexpression of recombinant IGFBP7 proteins induced premature senescence in early PD MRC-5 fibroblasts [13].

Among the genes significantly upregulated with age in both MRC-5 and HFFs we identified *DKK3*, having a role in Wnt signaling [55–57]. DKK3 has tumor suppressor activity in breast cancer patients [58] and in papillary thyroid carcinoma [57]. However, we had failed to demonstrate an induction of premature senescence in early PD HFFs on overexpression of recombinant DKK3 proteins [13]. Though not significantly differentially expressed with age in MRC-5 fibroblasts, one of the genes which were most significantly upregulated with age in HFFs was the *SFRP4* gene, an antagonist for Wnt signalling [59]. SFRP4 acts as a tumor suppressor in gastric carcinoma [60] and epithelial ovarian cancer cell lines [61]. In a separate study, we functionally validated the expression of *SFRP4* in early PD HFF and MRC-5 fibroblasts by treating them separately with human recombinant SFRP4 protein. This treatment resulted in premature senescence induction in HFFs but not in early PD MRC-5 fibroblasts [13]. Here, induction of SFRP4 mRNA expression was not detected by RNA-seq, explaining the lack of premature senescence induction in early PD MRC-5 fibroblasts. SFRP4 expression thus showed cell line specific differences.

3.6. Clustering of the Expression Profiles Shows Similar Pattern in Both Cell Lines. We found many differentially expressed

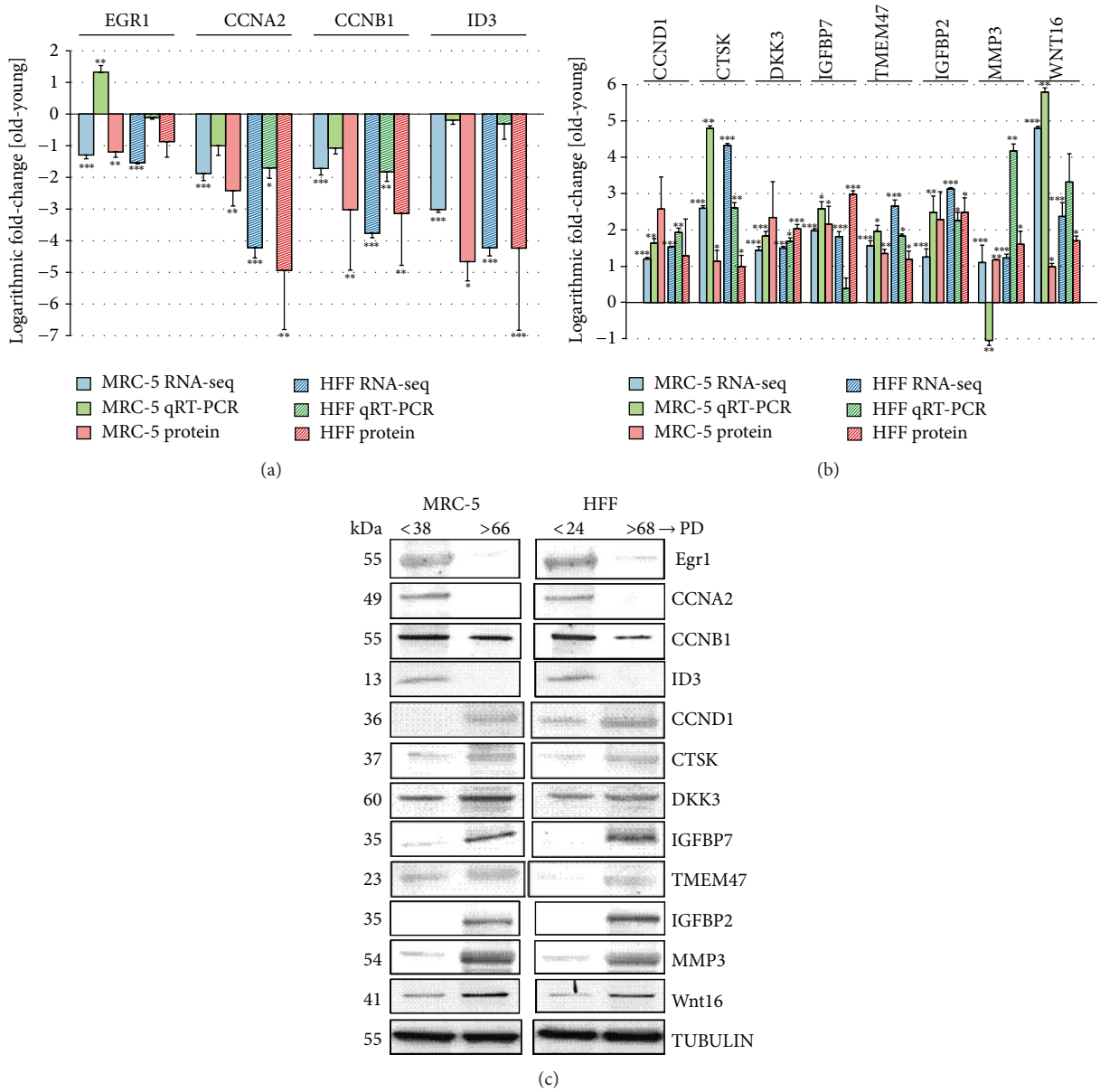


Figure 5: Comparison of expression changes between young and old MRC-5 and HFF fibroblasts measured by RNA-seq, qRT-PCR, and Western Blots. (a) Four genes commonly downregulated and (b) 8 genes commonly upregulated in both cell lines. (a, b) The colors of the bars indicate the measurement technique (blue: RNA-seq; green: qRT-PCR; red: Western Blots/protein expression). Solid colored bars represent MRC-5 while shaded boxes represent HFF cells. The height of the bars corresponds to the logarithmic fold-change (FC) of expression between the first and the last PD investigated here (RNA-seq: log 2 RPKM FC; qRT-PCR: log $2^{-\Delta\Delta CT}$; protein: log 2 expression ratio). Error bars indicate standard deviation from the mean. Changes statistically different comparing young and old PD (RNA-seq: DESeq; rRT-PCR/Protein: Student's t-test; $n = 3$) are indicated with an asterisk: $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. (c) The blots show the protein expression levels in MRC-5 and HFF cells at young compared to old PDs. The up- or downregulation was signified by the presence or absence of bands in Western Blots.

genes commonly regulated in both MRC-5 and HFF. Next, we asked if both cell lines exhibit common temporal expression profiles rather than showing different effects for the same set of genes. Therefore, we applied fuzzy c-means clustering comparing the expression profiles of both cell lines. We used 1,803 genes found to be differentially regulated between the four consecutive PDs and between the first and the last PD in both cell lines, according to the strict cutoffs as shown in Figures 3(b) and 3(d) (FDR < 0.01; |log 2 fold-change| > 1). Using several cluster validation indexes, an optimal number of five clusters were estimated, and each selected DEG was assigned to one out of these five groups (Figure 7). The majority of DEG exhibits similar temporal expression profiles in MRC-5 and HFF. 811 DEG were upregulated (clusters 3 and 5) and 722 are downregulated (clusters 2 and 4). Stronger differences between both cell lines were found for genes

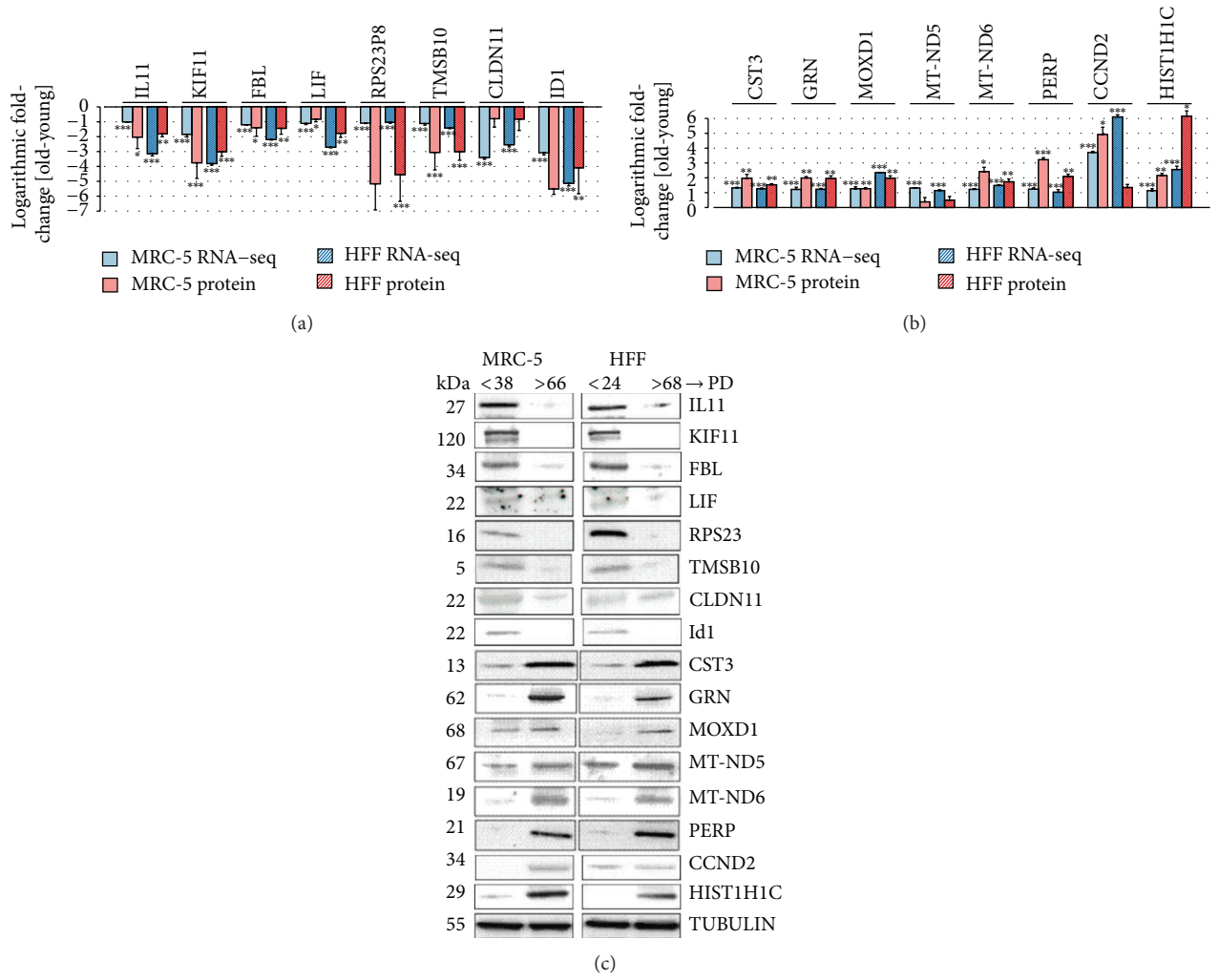


Figure 6: Comparison of expression changes between young and old MRC-5 and HFF fibroblasts measured with RNA-seq and Western Blots. (a) 8 genes commonly downregulated and (b) 8 genes commonly upregulated in both cell lines. (a, b) The colors of the bars indicate the measurement technique (blue: RNA-seq; red: Western Blots/protein expression). Solid colored bars represent MRC-5 while shaded boxes represent HFF cells. The height of the bars corresponds to the logarithmic fold-change (FC) of expression between the first and the last PD investigated here (RNA-seq: log 2 RPKM FC; protein: log 2 expression ratio). Error bars indicate standard deviation from the mean. Changes statistically different comparing young and old PD (RNA-seq: DESeq; rRT-PCR/Protein: Student's t-test; $n = 3$) are indicated with an asterix: $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. (c) The blots show the protein expression levels in MRC-5 and HFF cells at young compared to old PDs. The up- or downregulation was signified by the presence or absence of bands in Western Blots.

grouped in clusters 1 and 4. Interestingly, most genes follow a monotonic profile (either up or down) while only few genes exhibit a parabolic-like shape. Clusters 1 and 3 show major changes in MRC-5 between the second to last and the last PD, while cluster 4 groups genes with large differences between the first and the second PD in HFF. This effect was already observed when comparing DEG between the consecutive PDs (see DEG section above). Figure 7 summarizes the gene expression profiles by showing only the scaled and centred mean and standard deviation of the DEG clustered. In most of the cases, the absolute expression values were different between both cell lines (indicated by the dashed horizontal lines) but the trends of the actual changes across the five PDs were similar. For instance, cluster 3 contains genes which show larger mean expression values for MRC-5 but are upregulated with increasing PDs in both cell lines (*vice versa* in cluster 2).

3.7. Identification of Functional Categories Significantly Enriched for Genes with Common Expression Profiles. Next, we deduced the main biological processes driven by the differentially expressed gene sets obtained from the cluster analysis. Using gene set enrichment analysis, for each of the five clusters significant GO categories and KEGG pathways could be identified. The results indicated a strong connection of upregulated genes (grouped in clusters 3 and 5) to "extracellular space" (GO:0005615) and "membrane" (GO:0016020) components (Figure 8(a)). Corresponding KEGG pathways, found for these genes, were for example, "ECM-receptor interaction" (hsa04512) and

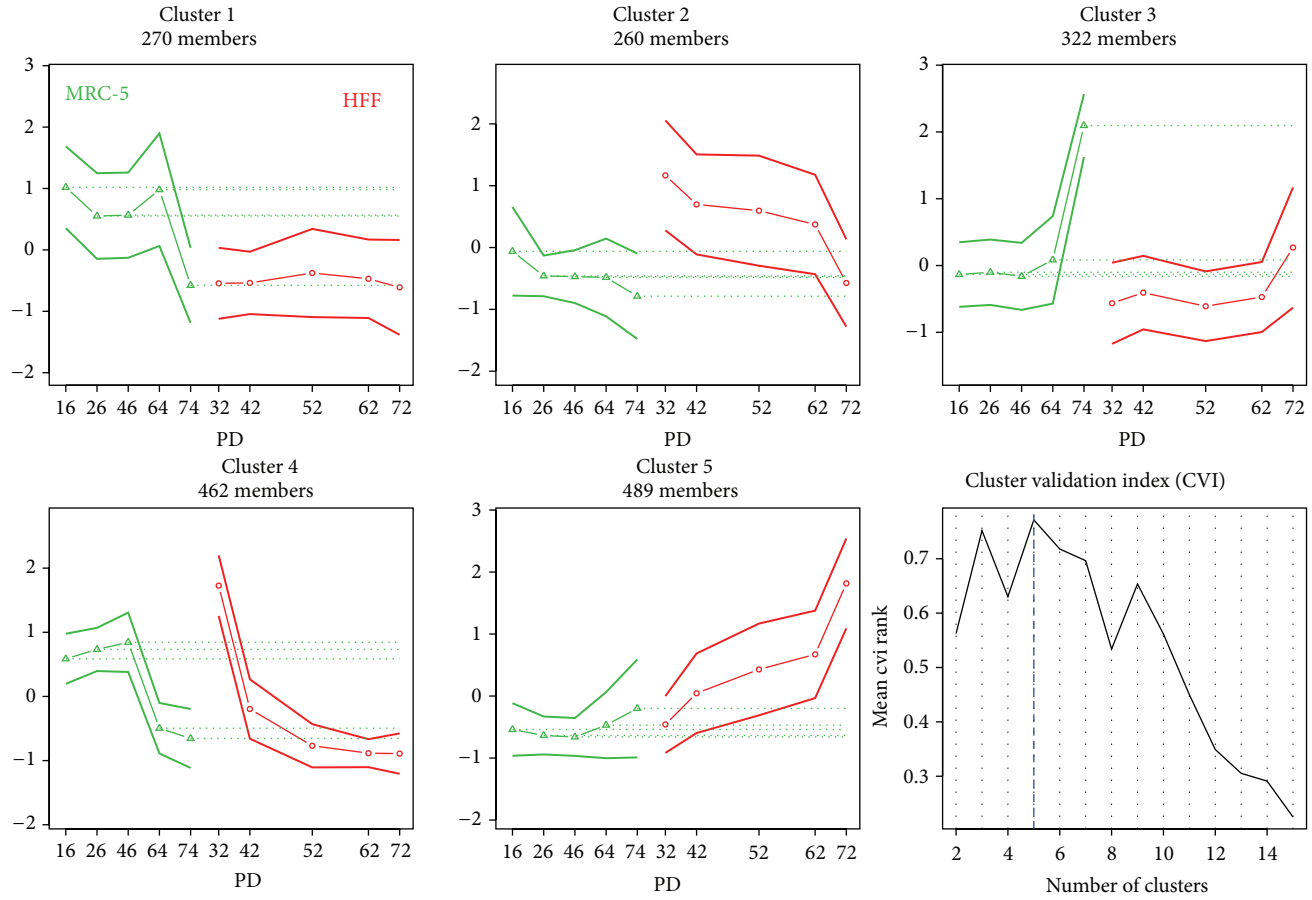
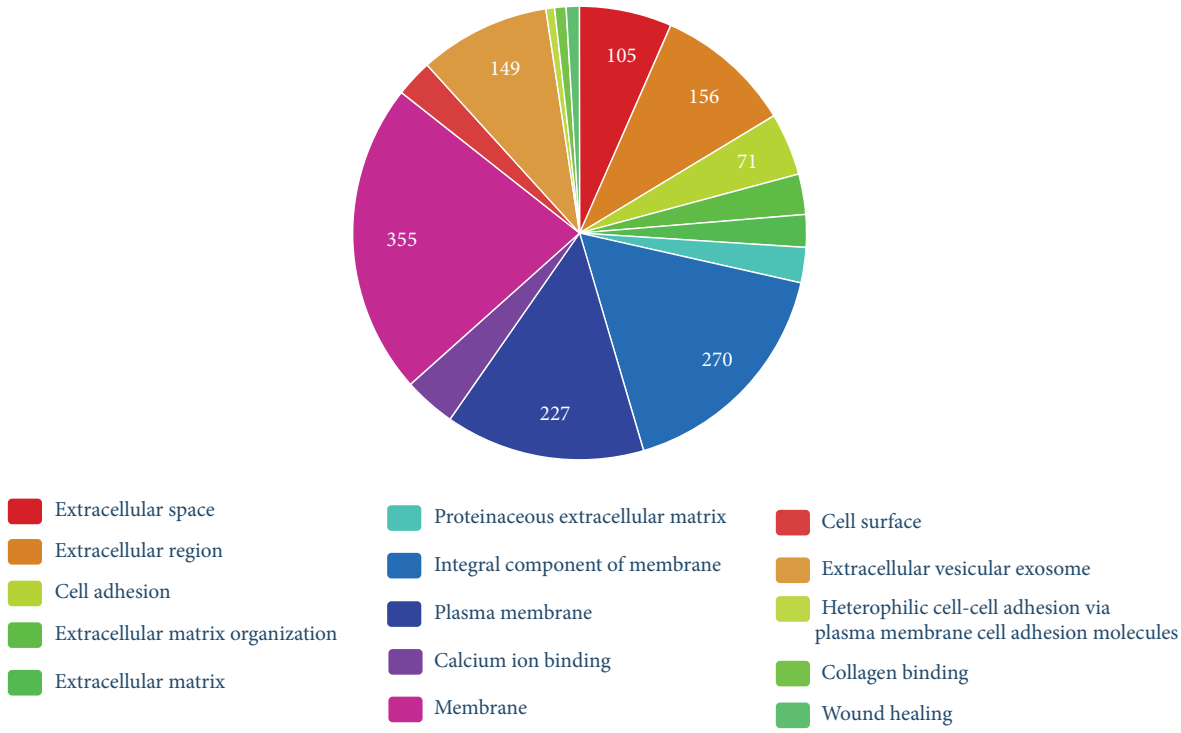


Figure 7: Combined line plot showing the scaled and clustered expression profiles of 1,803 genes detected in both cell lines as differentially expressed. The middle line, including the points, indicates the mean expression profile for all included genes in a certain cluster. The thicker lines above and below indicate the standard deviation. Cell lines are indicated by color (green lines: MRC-5; red lines: HFF cells). The horizontal dotted green lines can be used to compare the mean profiles of both cell lines. The last plot shows a summary of the cluster validation analysis. An optimal number of five clusters were estimated.

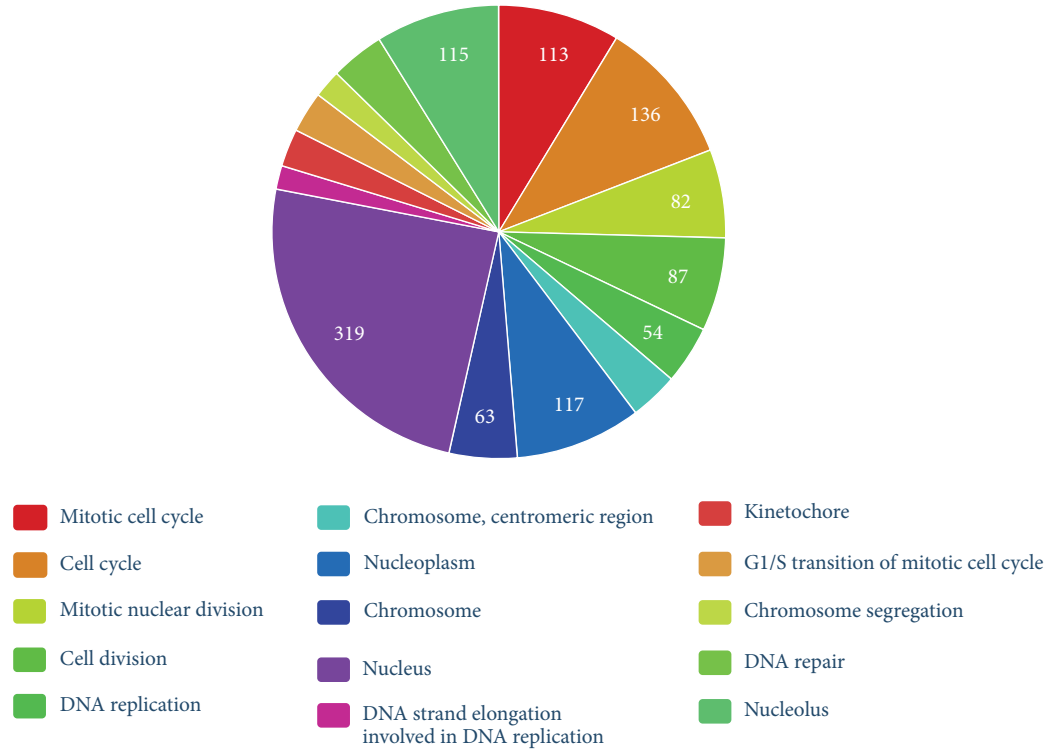
"ABC transporters" (hsa02010; see Supplemental Table 3). ABC proteins transport various molecules across extra- and intracellular membranes and are involved in aging and age-related diseases [62]. Cluster 3 shows stronger upregulation at late PD for MRC-5 cells while HFF cells are upregulated more clearly in cluster 5 (see above). Comparing the GO-terms found for these single clusters, we found links of the stronger upregulation in MRC-5 with "integral component of plasma membrane" (GO:0005887) and the "Golgi apparatus" (GO:0005794), while "sarcolemma" (GO:0042383) and "nucleosome" (GO:0000786) were more specific for upregulation in HFF (Supplemental Figure 2). The structure of the secretion regulating Golgi complex is altered in senescent cells [63]. While our results indicate cell line specific differences during replicative senescence, the GO-term comparison revealed that in both cell lines many genes were similarly upregulated. A large set of GO-terms associated with upregulated genes were related to the senescence associated secretory phenotype [64].

Downregulated genes (grouped in clusters 2 and 4) were associated to strongly enriched GO processes related to, for example, "cell cycle" (GO:0007049), "cell division" (GO:0051301) and "DNA replication" (GO:0006260) (Figure 8(b)). Here, differences between both cell lines are more obvious. After a slight initial gain, expression in MRC-5 cells declined strongly between PD 46 and PD 64. In HFFs, strong decline already started after PD32 without larger changes for late PD (Figure 7; cluster 4). Most of these cell cycle related genes, which account for the above-mentioned profiles, are related to the cellular component "nucleoplasm" (GO:0005654). Associated GO-terms for cluster 2, which depicts moderate downregulation, were more widespread and covered processes like "positive regulation of nitric oxide biosynthetic process" (GO:0045429), "endoderm formation" (GO:0001706), and "response to cAMP" (GO:0051591; Supplemental Figure 3).

Cluster 1 showed the largest differences between both cell lines. While, in MRC-5, genes are downregulated strongly at the last PD, no clear up- or downregulation is observed



(a) 15 most significantly enriched categories. Method: GO



(b) 15 most significantly enriched categories. Method: GO

Figure 8: (a) Top 15 of 171 significantly enriched GO-terms based on 811 upregulated genes (clusters 3 and 5). (b) Top 15 of 253 significantly enriched GO-terms based on 722 downregulated genes (clusters 2 and 4). The numbers in the pie-chart correspond to the number of genes which were included in the clusters and assigned to the GO-terms stated in the chart.

Table 2: Number of genes whose expression values are monotonically up- and downregulated, respectively, for MRC-5 and HFF cells. Note that monotonic behaviour does not necessarily include differential expression. See Supplemental Table 4 for complete test results.

Class	MRC-5	HFF
Monotonically upregulated	47	423
Monotonically downregulated	132	465
Nonuniformly (not significant)	19,058	18,349

for HFF. Significantly enriched GO-terms associated to these genes were, for example, "vasculogenesis" (GO:0001570), "response to lipopolysaccharide" (GO:0032496), and "cell adhesion" (GO:0007155) (Supplemental Figure 4).

3.8. Monotonically Regulated Genes in MRC-5 and HFF Are Connected in Functional Association Networks. Since senescence is a continuous cellular process, it can be hypothesized that genes possessing key relevance for senescence change their expression values monotonically over time, while genes with irregular temporal expression patterns might be associated with response to environmental conditions, with the circadian rhythm or other processes.

Amongst others, continuous increasing and decreasing profiles were found by the clustering analysis. In addition to this nonbiased approach, we intended to identify genes with a strong monotonic behaviour across the PDs investigated here. We calculated the Spearman correlation coefficient of each gene's temporal profile with a linearly increasing sequence. Replicates for each PD were incorporated by a random sampling approach. Subsequently, we classified genes into three classes according to their behaviour with age: (a) monotonically upregulated genes, (b) monotonically downregulated genes, and (c) nonuniformly regulated genes (Table 2).

More monotonically up- and downregulated genes were found for HFFs compared to MRC-5 (888 versus 179). Only a small subset of these genes were commonly regulated in both cell lines (9 up and 14 down) but even less genes showed an opposite monotonic expression profiles (8; see Supplemental Figure 5 and Supplemental Table 4). The 23 commonly monotonically up- or downregulated genes were studied in more detail. Since in both cell lines the regulation of these genes strongly correlated with an increase of senescence, they might play an essential role in cellular aging and may rule common regulatory process. We used several online resources in order to find potential or validated interactions between these genes. The STRING database [36] only provides the interactions between four out of all the 23 genes (Supplemental Figure 6A). Using Cognoscente [37], 17 out of 23 genes were connected within one interaction graph (Supplemental Figure 6B). More interactions could be found using GeneMANIA [38], leading to a network which is widely connected by coexpression and common pathways like, for example, "epithelial cell proliferation" and "extracellular matrix organization" (Figure 9). Both of the latter tools integrate intermediate genes which were not in the input list. Hub genes in these networks included *ATF7*, *MAF*, *UBC,* and *ELAVL* which are interesting candidates for further studies. All the four of these genes were functionally associated with tumorigenesis. Members of the ubiquitin family including UBC have been associated with tumor progression [65]. In terms of ATF7, the activating transcription factor family is associated with cell proliferation and oncogenesis [66]. Both MAF and ELAV1 have been associated with oncogenesis and tumor progression [67, 68]. Thus all the four genes had an association with cell proliferation. Then, we investigated the biological relevance of the monotonically up- and downregulated genes in both fibroblast cell lines. The list of monotonically downregulated genes included *NLE1*, *AMMECR1*, *FIBCD1*, *ENPP2*, *TMTC4*, *ANPEP*, *MYC*, *EFNB3*, *HCLS1*, *FERMT1*, *FABP5*, *SPHK1*, *GOS2,* and *RPL36A*. The genes monotonically upregulated included *LRP10*, *TMCO3*, *CAV2*, *ADAMTS5*, *C5orf15*, *SDC2*, *ANKH*, *PCDHB16,* and *TGFB2*. A number of genes in the above list have been functionally associated with proliferation.

3.8.1. Monotonically Downregulated Genes. *NLE* plays a role in regulating the Notch activity and is involved in embryonic development in mammals by affecting the CDKN1A and Wnt pathways [69]. Forced expression of miR-26 inhibits the growth of stimulated breast cancer cells and tumor in xenograft models by reducing the mRNA expression levels of AMMECR1 and other genes [70]. AMMECR1 is associated with Alport syndrome, mental retardation, midface hypoplasia, and elliptocytosis [71]. FIBCD1 (fibrinogen C domain containing 1) binds to chitin of invading parasites [72]. FIBCD1 is primarily present in the gastrointestinal tract of humans; however, their presence in skin has been highly debated [73, 74]. ENPP2 facilitates cell motility and progression and is related to the invasion of ductal breast carcinomas [75]. *TMTC4* is a gene contributing to embryonic brain development; it interacts with Wntless, an integral Wnt regulator [76]. *EFNB3*, a member of the ephrin gene family, is associated with neural development [77]. ANPEP is a well-known marker for acute myeloid leukemia and tumor invasion; it has a regulatory role in angiogenesis [78, 79]. FERMT1 is overexpressed in colon and lung carcinomas [80]. The *MYC* oncogene is associated with cell growth regulation by driving proliferation via upregulation of Cyclins and downregulation of p21 [81, 82]. *HCLS1* gene which is monotonically downregulated with age is associated with antigen receptor signaling and clonal expansion as well as deletion of lymphoid cells [83]. The *FABP5* gene encodes the fatty acid binding protein in epidermal cells and is upregulated in psoriatic tissues [84]. SPHK1 has been previously associated with melanoma progression and angiogenesis [85, 86]. The *GOS2* gene promotes apoptosis by binding to BCL2, hence preventing the formation of protective BCL2-BAX; its mRNA and protein levels are downregulated in type 2 diabetic patients [87, 88]. Thus, almost all genes, monotonically downregulated with age in both fibroblast cell lines, are associated with proliferation and cell survival.

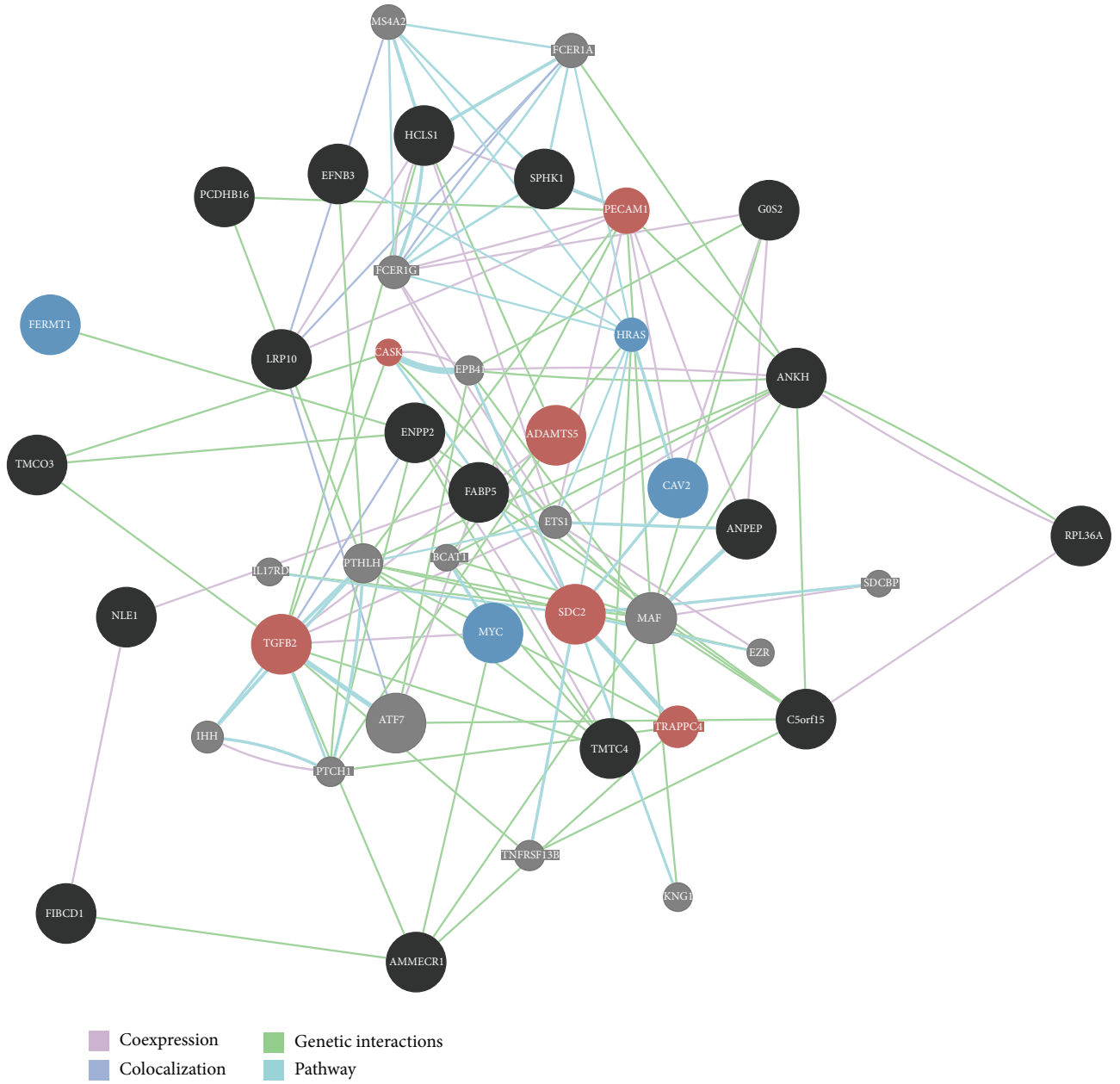


Figure 9: Functional association network by GeneMANIA for 23 monotonically up- or downregulated genes. Input genes are depicted by either black, red, or blue color or large circles. Intermediate genes, added by the tool, are shown using small grey circles. Colored genes denote functions associated by the tools: blue = "epithelial cell proliferation"; red = " extracellular matrix organization." The edge colors indicate the type of interaction, as explained in the legend on the bottom right.

3.8.2. Monotonically Upregulated Genes. *LRP10*, a negative regulator of Wnt signalling, was found monotonically upregulated with age [89]. CAV2 is a scaffolding protein within the caveolar membrane modulating cancer progression [90]. ADAMTS5 enables the destruction of aggrecan in patients with arthritic disease which is prevalent with aging [91]. The *ANKH* gene, associated with regulation of tissue calcification and in turn susceptibility to arthritis, is also monotonically upregulated with age in both fibroblast cell lines [92]. Syndecan-2 protein (SDC2) is upregulated in skin and lung tissues of patients suffering from (age-associated) systemic sclerosis and fibrosis [93, 94]. mRNA expression of PCDHB16 is upregulated in patients with (age-associated) Alzheimer's disease [95]. TGFB2, also monotonically upregulated with age in fibroblasts, has suppressive effects on interleukin-2 dependent T cell proliferation and displays effector functions [96].

In summary, the genes, which we found here monotonically up- and downregulated with age in both fibroblast cell lines, have been studied before. In this study, we explicitly

show for the first time the age-associated regulation of these genes in primary human fibroblast cells of two different origins. In a following study we will determine the protein expression of all age-related genes and functionally validate the expression of these genes.

4. Conclusion

We studied molecular aspects of cellular aging by determining the differential expression of genes during the aging of two primary human fibroblasts, MRC-5 and HFFs. RNA-seq data analysis encompassed different levels, starting from the complete set of annotated and expressed genes, proceeding to different gene subsets and functional categories. Most of the detected changes were found to be common in both cell lines, as indicated by the large number of overlapping DEG and common expression profiles identified by clustering. We validated the expression patterns for selected genes, demonstrating an association of almost all most differentially expressed genes with proliferation or cell cycle arrest, consistent with previous senescence studies. Investigating expression changes across five consecutive PDs and comparing young with senescent cells enabled us to identify both monotonically up- and downregulated genes as well as the most differentially expressed genes. Both sets of genes strongly contributed to the transition into cellular senescence. Thus, we quantitatively describe similarities in gene expression profiles during the aging of two fibroblast cell lines of different origin.

Data Deposition

The RNA-seq data discussed in this publication have been deposited in NCBIs Gene Expression Omnibus and are accessible through GEO Series accession number GSE63577.

Conflict of Interests

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

Authors' Contribution

Shiva Marthandan and Steffen Priebe contributed equally to this work.

Acknowledgments

The work described here is part of the research program of the Jena Centre for Systems Biology of Ageing (JenAge). The authors acknowledge JenAge funding by the German Ministry for Education and Research (Bundesministerium für Bildung und Forschung (BMBF); support code: 0315581). The authors would like to thank Sabine Ohndorf and Sabine Gallert for excellent technical assistance.

References

[1] L. Hayflick and P. S. Moorhead, "The serial cultivation of human diploid cell strains," *Experimental Cell Research*, vol. 25, no. 3, pp. 585–621, 1961.

[2] J. Campisi, "Aging and cancer cell biology, 2008," *Aging Cell*, vol. 7, no. 3, pp. 281–284, 2008.

[3] J. M. Vicencio, L. Galluzzi, N. Tajeddine et al., "Senescence, apoptosis or autophagy? When a damaged cell must decide its path—a mini-review," *Gerontology*, vol. 54, no. 2, pp. 92–99, 2008.

[4] J.-Y. Kato, M. Matsuoka, K. Polyak, J. Massagué, and C. J. Sherr, "Cyclic AMP-induced G1 phase arrest mediated by an inhibitor ($p27^{Kip1}$) of cyclin-dependent kinase 4 activation," *Cell*, vol. 79, no. 3, pp. 487–496, 1994.

[5] T. von Zglinicki, G. Saretzki, W. Docke, and C. Lotze, "Mild hyperoxia shortens telomeres and inhibits proliferation of fibroblasts: a model for senescence?" *Experimental Cell Research*, vol. 220, no. 1, pp. 186–193, 1995.

[6] F. A. Mallette, M.-F. Gaumont-Leclerc, and G. Ferbeyre, "The DNA damage signaling pathway is a critical mediator of oncogene-induced senescence," *Genes & Development*, vol. 21, no. 1, pp. 43–48, 2007.

[7] T. V. Zglinicki, G. Saretzki, J. Ladhoff, F. D. D. Fagagna, and S. P. Jackson, "Human cell senescence as a DNA damage response," *Mechanisms of Ageing and Development*, vol. 126, no. 1, pp. 111–117, 2005.

[8] R. Marcotte, C. Lacelle, and E. Wang, "Senescent fibroblasts resist apoptosis by downregulating caspase-3," *Mechanisms of Ageing and Development*, vol. 125, no. 10-11, pp. 777–783, 2004.

[9] S. Marthandan, S. Priebe, P. Hemmerich, K. Klement, S. Diekmann, and T. G. Hofmann, "Long-term quiescent fibroblast cells transit into senescence," *PLoS ONE*, vol. 9, no. 12, Article ID e115597, 2014.

[10] F. Rodier and J. Campisi, "Four faces of cellular senescence," *Journal of Cell Biology*, vol. 192, no. 4, pp. 547–556, 2011.

[11] G. P. Dimri, X. Lee, G. Basile et al., "A biomarker that identifies senescent human cells in culture and in aging skin in vivo," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 20, pp. 9363–9367, 1995.

[12] C. Lawless, C. Wang, D. Jurk, A. Merz, T. V. Zglinicki, and J. F. Passos, "Quantitative assessment of markers for cell senescence," *Experimental Gerontology*, vol. 45, no. 10, pp. 772–778, 2010.

[13] S. Marthandan, S. Priebe, M. Baumgart et al., "Conserved genes and pathways ruling the transition into senescence in primary human fibroblasts," *Mechanisms of Ageing and Development*. In press.

[14] T. Kuilman, C. Michaloglou, W. J. Mooi, and D. S. Peeper, "The essence of senescence," *Genes and Development*, vol. 24, no. 22, pp. 2463–2479, 2010.

[15] J. Campisi, "Suppressing cancer: the importance of being senescent," *Science*, vol. 309, no. 5736, pp. 886–887, 2005.

[16] J. Campisi and J. Sedivy, "How does proliferative homeostasis change with age? What causes it and how does it contribute to aging?" *Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 64, no. 2, pp. 164–166, 2009.

[17] J. C. Jeyapalan, M. Ferreira, J. M. Sedivy, and U. Herbig, "Accumulation of senescent cells in mitotic tissue of aging primates," *Mechanisms of Ageing and Development*, vol. 128, no. 1, pp. 36–44, 2007.

[18] O. Toussaint, E. E. Medrano, and T. von Zglinicki, "Cellular and molecular mechanisms of stress-induced premature senescence (SIPS) of human diploid fibroblasts and melanocytes," *Experimental Gerontology*, vol. 35, no. 8, pp. 927–945, 2000.

[19] M. Kronschnabl and T. Stamminger, "Synergistic induction of intercellular adhesion molecule-1 by the human cytomegalovirus transactivators IE2p86 and pp71 is mediated via an Sp1-binding site," *Journal of General Virology*, vol. 84, part 1, pp. 61–73, 2003.

[20] S. Honda, L. M. Hjelmeland, and J. T. Handa, "Oxidative stress-induced single-strand breaks in chromosomal telomeres of human retinal pigment epithelial cells in vitro," *Investigative Ophthalmology and Visual Science*, vol. 42, no. 9, pp. 2139–2144, 2001.

[21] T. Ulbricht, M. Alzrigat, A. Horch et al., "PML promotes MHC class II gene expression by stabilizing the class II transactivator," *Journal of Cell Biology*, vol. 199, no. 1, pp. 49–63, 2012.

[22] S. Sivakumar, J. R. Daum, A. R. Tipton, S. Rankin, and G. J. Gorbsky, "The spindle and kinetochore-associated (Ska) complex enhances binding of the anaphase-promoting complex/cyclosome (APC/C) to chromosomes and promotes mitotic exit," *Molecular Biology of the Cell*, vol. 25, no. 5, pp. 594–605, 2014.

[23] M. Baumgart, M. Groth, S. Priebe et al., "RNA-seq of the aging brain in the short-lived fish *N. furzeri*—conserved pathways and novel genes associated with neurogenesis," *Aging Cell*, vol. 13, no. 6, pp. 965–974, 2014.

[24] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow et al., "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature*, vol. 456, pp. 53–59, 2008.

[25] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biology*, vol. 14, no. 4, article R36, 2013.

[26] Y. Liao, G. K. Smyth, and W. Shi, "FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, vol. 30, no. 7, pp. 923–930, 2014.

[27] P. Flicek, M. R. Amode, D. Barrell et al., "Ensemble 2012," *Nucleic Acids Research*, vol. 40, no. 1, pp. D84–D90, 2012.

[28] R Development Core Team, "R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria," 2008, http://www.R-project.org.

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

[30] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[31] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[32] R. Guthke, U. Möller, M. Hoffman, F. Thies, and S. Töpfer, "Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection," *Bioinformatics*, vol. 21, no. 8, pp. 1626–1634, 2005.

[33] S. Priebe, C. Kreisel, F. Horn, R. Guthke, and J. Linde, "FungiFun2: a comprehensive online resource for systematic analysis of gene lists from fungal species," *Bioinformatics*, vol. 31, no. 3, pp. 445–446, 2015.

[34] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, "Revigo summarizes and visualizes long lists of gene ontology terms," *PLoS ONE*, vol. 6, no. 7, Article ID e21800, 2011.

[35] S. Priebe, U. Menzel, K. Zarse et al., "Extension of life span by impaired glucose metabolism in *Caenorhabditis elegans* is accompanied by structural rearrangements of the transcriptomic network," *PloS One*, vol. 8, no. 10, Article ID e77776, 2013.

[36] A. Franceschini, D. Szklarczyk, S. Frankild et al., "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research*, vol. 41, no. 1, pp. D808–D815, 2013.

[37] V. VanBuren and H. Chen, "Managing biological complexity across orthologs with a visual knowledgebase of documented biomolecular interactions," *Scientific Reports*, vol. 2, article 1011, 2012.

[38] D. Warde-Farley, S. L. Donaldson, O. Comes et al., "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Research*, vol. 38, no. 2, Article ID gkq537, pp. W214–W220, 2010.

[39] S. Schäuble, K. Klement, S. Marthandan et al., "Quantitative model of cell cycle arrest and cellular senescence in primary human fibroblasts," *PLoS ONE*, vol. 7, no. 8, Article ID e42150, 2012.

[40] C. Vogel and E. M. Marcotte, "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses," *Nature Reviews Genetics*, vol. 13, no. 4, pp. 227–232, 2012.

[41] X. W. Cheng, R. Kikuchi, H. Ishii et al., "Circulating cathepsin K as a potential novel biomarker of coronary artery disease," *Atherosclerosis*, vol. 228, no. 1, pp. 211–216, 2013.

[42] M. J. Smout, T. Laha, J. Mulvenna et al., "A granulin-like growth factor secreted by the carcinogenic liver fluke, *Opisthorchis viverrini*, promotes proliferation of host cells," *PLoS Pathogens*, vol. 5, no. 10, Article ID e1000611, 2009.

[43] C. F. de Winter, M. A. Echteld, and H. M. Evenhuis, "Chronic kidney disease in older people with intellectual disability: results of the HA-ID study," *Research in Developmental Disabilities*, vol. 35, no. 3, pp. 726–732, 2014.

[44] W. Wagner, P. Horn, M. Castoldi et al., "Replicative senescence of mesenchymal stem cells: a continuous and organized process," *PLoS ONE*, vol. 3, no. 5, Article ID e2213, 2008.

[45] L. M. Hjelmeland, "Senescence of the retinal pigmented epithelium," *Ophthalmology & Visual Science*, vol. 40, pp. 1–2, 1999.

[46] H. Matsunaga, J. T. Handa, C. M. Gelfman, and L. M. Hjelmeland, "The mRNA phenotype of a human RPE cell line at replicative senescence," *Molecular Vision*, vol. 29, no. 5, article 39, 1999.

[47] D. N. Shelton, E. Chang, P. S. Whittier, D. Choi, and W. D. Funk, "Microarray analysis of replicative senescence," *Current Biology*, vol. 9, no. 17, pp. 939–945, 1999.

[48] Y. Kong, H. Cui, and H. Zhang, "Smurf2-mediated ubiquitination and degradation of Id1 regulates p16 expression during senescence," *Aging Cell*, vol. 10, no. 6, pp. 1038–1046, 2011.

[49] T. Chen, L. Xue, J. Niu et al., "The retinoblastoma protein selectively represses E2F1 targets via a TAAC DNA element during cellular senescence," *The Journal of Biological Chemistry*, vol. 287, no. 44, pp. 37540–37551, 2012.

[50] I. Androic, A. Krämer, R. Yan et al., "Targeting cyclin B1 inhibits proliferation and sensitizes breast cancer cells to taxol," *BMC Cancer*, vol. 8, article 391, 2008.

[51] N. Wajapeyee, R. W. Serra, X. Zhu, M. Mahalingam, and M. R. Green, "Oncogenic BRAF induces senescence and apoptosis through pathways mediated by the secreted protein IGFBP7," *Cell*, vol. 132, no. 3, pp. 363–374, 2008.

[52] A. J. T. Millis, M. Hoyle, H. M. McCue, and H. Martini, "Differential expression of metalloproteinase and tissue inhibitor of metalloproteinase genes in aged human fibroblasts," *Experimental Cell Research*, vol. 201, no. 2, pp. 373–379, 1992.

[53] M. K. Kang, A. Kameta, K.-H. Shin, M. A. Baluda, H.-R. Kim, and N.-H. Park, "Senescence-associated genes in normal human oral keratinocytes," *Experimental Cell Research*, vol. 287, no. 2, pp. 272–281, 2003.

[54] J.-P. Coppé, P.-Y. Desprez, A. Krtolica, and J. Campisi, "The senescence-associated secretory phenotype: the dark side of tumor suppression," *Annual Review of Pathology: Mechanisms of Disease*, vol. 5, pp. 99–118, 2010.

[55] C. Niehrs, "Function and biological roles of the Dickkopf family of Wnt modulators," *Oncogene*, vol. 25, no. 57, pp. 7469–7481, 2006.

[56] Y. Kawano, M. Kitaoka, Y. Hamada, M. M. Walker, J. Waxman, and R. M. Kypta, "Regulation of prostate cell growth and morphogenesis by Dickkopf-3," *Oncogene*, vol. 25, no. 49, pp. 6528–6537, 2006.

[57] D.-T. Yin, W. Wu, M. Li et al., "DKK3 is a potential tumor suppressor gene in papillary thyroid carcinoma," *Endocrine-Related Cancer*, vol. 20, no. 4, pp. 507–514, 2013.

[58] V. Kloten, B. Becker, K. Winner et al., "Promoter hypermethylation of the tumor-suppressor genes ITIH5, DKK3, and RASSF1A as novel biomarkers for blood-based breast cancer screening," *Breast Cancer Research*, vol. 15, no. 1, article R4, 2013.

[59] S. B. Baylin and J. E. Ohm, "Epigenetic gene silencing in cancer—a mechanism for early oncogenic pathway addiction?" *Nature Reviews Cancer*, vol. 6, no. 2, pp. 107–116, 2006.

[60] H. Kim, D. O. Lee, S.-Y. Ku, S. H. Kim, J. H. Kim, and J. G. Kim, "The association between polymorphisms in Wnt antagonist genes and bone response to hormone therapy in postmenopausal Korean women," *Menopause*, vol. 19, no. 9, pp. 1008–1014, 2012.

[61] C. E. Ford, E. Jary, S. S. Q. Ma, S. Nixdorf, V. A. Heinzelmann-Schwarz, and R. L. Ward, "The Wnt gatekeeper SFRP4 modulates EMT, cell migration and downstream Wnt signalling in serous ovarian cancer cells," *PLoS ONE*, vol. 8, no. 1, Article ID e54362, 2013.

[62] T. Efferth, "Adenosine triphosphate-binding cassette transporter genes in ageing and age-related diseases," *Ageing Research Reviews*, vol. 2, no. 1, pp. 11–24, 2003.

[63] J.-H. Cho, D. K. Saini, W. K. A. Karunarathne, V. Kalyanaraman, and N. Gautam, "Alteration of Golgi structure in senescent cells and its regulation by a G protein γ subunit," *Cellular Signalling*, vol. 23, no. 5, pp. 785–793, 2011.

[64] A. R. J. Young and M. Narita, "SASP reflects senescence," *EMBO Reports*, vol. 10, no. 3, pp. 228–230, 2009.

[65] B. M. Kessler, "Ubiquitin—omics reveals novel networks and associations with human disease," *Current Opinion in Chemical Biology*, vol. 17, no. 1, pp. 59–65, 2013.

[66] C. Zhao, J. Qi, and A. Meng, "Characterization and expression pattern of two zebrafish atf7 genes," *Developmental Dynamics*, vol. 233, no. 3, pp. 1157–1162, 2005.

[67] C. Pouponnot, K. Sii-Felice, I. Hmitou et al., "Cell context reveals a dual role for Maf in oncogenesis," *Oncogene*, vol. 25, no. 9, pp. 1299–1310, 2006.

[68] R. Upadhyay, S. Sanduja, V. Kaza, and D. A. Dixon, "Genetic polymorphisms in RNA binding proteins contribute to breast cancer survival," *International Journal of Cancer*, vol. 132, no. 3, pp. E128–E138, 2013.

[69] A. C. Lossie, C.-L. Lo, K. M. Baumgarner, M. J. Cramer, J. P. Garner, and M. J. Justice, "ENU mutagenesis reveals that Notchless homolog 1 (*Drosophila*) affects Cdkn1a and several members of the Wnt pathway during murine pre-implantation development," *BMC Genetics*, vol. 13, article 106, 2012.

[70] S. Tan, K. Ding, R. Li et al., "Identification of miR-26 as a key mediator of estrogen stimulated cell proliferation by targeting CHD1, GREB1 and KPNA2," *Breast Cancer Research*, vol. 16, no. 2, article no. R40, 2014.

[71] F. Vitelli, M. Piccini, F. Caroli et al., "Identification and characterization of a highly conserved protein absent in the Alport syndrome (A), mental retardation (M), midface hypoplasia (M), and elliptocytosis (E) contiguous gene deletion syndrome (AMME)," *Genomics*, vol. 55, no. 3, pp. 335–340, 1999.

[72] S. M. Abdel-Rahman and B. L. Preuett, "Genetic predictors of susceptibility to cutaneous fungal infections: a pilot genome wide association study to refine a candidate gene search," *Journal of Dermatological Science*, vol. 67, no. 2, pp. 147–152, 2012.

[73] A. Schlosser, T. Thomsen, J. B. Moeller et al., "Characterization of FIBCD1 as an acetyl group-binding receptor that binds chitin," *The Journal of Immunology*, vol. 183, no. 6, pp. 3800–3809, 2009.

[74] T. Thomsen, J. B. Moeller, A. Schlosser et al., "The recognition unit of FIBCD1 organizes into a noncovalently linked tetrameric structure and uses a hydrophobic funnel (S1) for acetyl group recognition," *The Journal of Biological Chemistry*, vol. 285, no. 2, pp. 1229–1238, 2010.

[75] B. Castellana, D. Escuin, G. Peiró et al., "ASPN and GJB2 are implicated in the mechanisms of invasion of ductal breast carcinomas," *Journal of Cancer*, vol. 3, no. 1, pp. 175–183, 2012.

[76] T. Popli, J. Lee, and E. Sherr, "Hi $Tmtc_4$ interacts with C_3G, Wnt-less and $Zfhx_4$: a yeast two-hybrid trap for proteins associated with Temtamy Syndrome," *The FASEB Journal*, vol. 25, pp. 963–967, 2011.

[77] D. G. Wilkinson, "Multiple roles of Eph receptors and ephrins in neural development," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 155–164, 2001.

[78] R. Pasqualini, E. Koivunen, R. Kain et al., "Aminopeptidase N is a receptor for tumor-homing peptides and a target for inhibiting angiogenesis," *Cancer Research*, vol. 60, no. 3, pp. 722–727, 2000.

[79] A. Kehlen, U. Lendeckel, H. Dralle, J. Langner, and C. Hoang-Vu, "Biological significance of aminopeptidase N/CD13 in thyroid carcinomas," *Cancer Research*, vol. 63, no. 23, pp. 8500–8506, 2003.

[80] E. J. Weinstein, M. Bourner, R. Head, H. Zakeri, C. Bauer, and R. Mazzarella, "URP1: a member of a novel family of PH and FERM domain-containing membrane-associated proteins is significantly over-expressed in lung and colon carcinomas," *Biochimica et Biophysica Acta*, vol. 1637, no. 3, pp. 207–216, 2003.

[81] H. Land, L. F. Parada, and R. A. Weinberg, "Tumorigenic conversion of primary embryo fibroblasts requires at least two cooperating oncogenes," *Nature*, vol. 304, no. 5927, pp. 596–602, 1983.

[82] E. Hatzi, C. Murphy, A. Zoephel et al., "N-myc oncogene overexpression down-regulates leukemia inhibitory factor in neuroblastoma," *European Journal of Biochemistry*, vol. 269, no. 15, pp. 3732–3741, 2002.

[83] I. Taniuchi, D. Kitamura, Y. Maekawa, I. Fukuda, H. Kishi, and T. Watanabe, "Antigen-receptor induced clonal expansion and deletion of lymphocytes are impaired in mice lacking HS1 protein, a substrate of the antigen-receptor-coupled tyrosine kinases," *The EMBO Journal*, vol. 14, no. 15, pp. 3664–3678, 1995.

[84] P. Madsen, H. H. Rasmussen, H. Leffers, B. Honoré, and J. E. Celis, "Molecular cloning and expression of a novel keratinocyte protein (psoriasis-associated fatty acid-binding protein [PA-FABP]) that is highly up-regulated in psoriatic skin and that shares similarity to fatty acid-binding proteins," *Journal of Investigative Dermatology*, vol. 99, no. 3, pp. 299–305, 1992.

[85] Z. Wang, X. Min, S.-H. Xiao et al., "Molecular basis of sphingosine kinase 1 substrate recognition and catalysis," *Structure*, vol. 21, no. 5, pp. 798–809, 2013.

[86] V. Albinet, M.-L. Bats, A. Huwiler et al., "Dual role of sphingosine kinase-1 in promoting the differentiation of dermal fibroblasts and the dissemination of melanoma cells," *Oncogene*, vol. 33, no. 26, pp. 3364–3373, 2014.

[87] T. S. Nielsen, U. Kampmann, R. R. Nielsen et al., "Reduced mRNA and protein expression of perilipin A and G0/G 1 switch gene 2 (G0S2) in human adipose tissue in poorly controlled type 2 diabetes," *Journal of Clinical Endocrinology and Metabolism*, vol. 97, no. 7, pp. E1348–E1352, 2012.

[88] C. Welch, M. K. Santra, W. El-Assaad et al., "Identification of a protein, G0S2, that lacks Bcl-2 homology domains and interacts with and antagonizes Bcl-2," *Cancer Research*, vol. 69, no. 17, pp. 6782–6789, 2009.

[89] Y.-H. Jeong, M. Sekiya, M. Hirata et al., "The low-density lipoprotein receptor-related protein 10 is a negative regulator of the canonical Wnt/β-catenin signaling pathway," *Biochemical and Biophysical Research Communications*, vol. 392, no. 4, pp. 495–499, 2010.

[90] S. Lee, H. Kwon, K. Jeong, and Y. Pak, "Regulation of cancer cell proliferation by caveolin-2 down-regulation and re-expression," *International Journal of Oncology*, vol. 38, no. 5, pp. 1395–1402, 2011.

[91] J. Velasco, J. Li, L. DiPietro, M. A. Stepp, J. D. Sandy, and A. Plaas, "Adamts5 deletion blocks murine dermal repair through CD44-mediated aggrecan accumulation and modulation of transforming growth factor β1 (TGFβ1) signaling," *The Journal of Biological Chemistry*, vol. 286, no. 29, pp. 26016–26027, 2011.

[92] Y. Vistoropsky, M. Keter, I. Malkin, S. Trofimov, E. Kobyliansky, and G. Livshits, "Contribution of the putative genetic factors and ANKH gene polymorphisms to variation of circulating calciotropic molecules, PTH and BGP," *Human Molecular Genetics*, vol. 16, no. 10, pp. 1233–1240, 2007.

[93] A. Biernacka and N. G. Frangogiannis, "Aging and cardiac fibrosis," *Aging and Disease*, vol. 2, no. 2, pp. 158–173, 2011.

[94] X. D. Ruiz, L. R. Mlakar, Y. Yamaguchi et al., "Syndecan-2 is a novel target of insulin-like growth factor binding protein-3 and is over-expressed in fibrosis," *PLoS ONE*, vol. 7, no. 8, Article ID e43049, 2012.

[95] R. Ricciarelli, C. D'Abramo, S. Massone, U. M. Marinari, M. A. Pronzato, and M. Tabaton, "Microarray analysis in Alzheimer's disease and normal aging," *IUBMB Life*, vol. 56, no. 6, pp. 349–354, 2004.

[96] Å. Schiött, H. O. Sjögren, and M. Lindvall, "Monocyte-dependent costimulatory effect of TGF-β1 on rat T-cell activation," *Scandinavian Journal of Immunology*, vol. 44, no. 3, pp. 252–260, 1996.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 836929, 8 pages
http://dx.doi.org/10.1155/2015/836929

Research Article

Module Based Differential Coexpression Analysis Method for Type 2 Diabetes

Lin Yuan,[1] Chun-Hou Zheng,[2] Jun-Feng Xia,[3] and De-Shuang Huang[1]

[1]*School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China*
[2]*College of Electrical Engineering and Automation, Anhui University, Hefei 230601, China*
[3]*Institute of Health Sciences, Anhui University, Hefei 230601, China*

Correspondence should be addressed to De-Shuang Huang; dshuang@tongji.edu.cn

Received 4 December 2014; Accepted 29 December 2014

Academic Editor: Fang-Xiang Wu

More and more studies have shown that many complex diseases are contributed jointly by alterations of numerous genes. Genes often coordinate together as a functional biological pathway or network and are highly correlated. Differential coexpression analysis, as a more comprehensive technique to the differential expression analysis, was raised to research gene regulatory networks and biological pathways of phenotypic changes through measuring gene correlation changes between disease and normal conditions. In this paper, we propose a gene differential coexpression analysis algorithm in the level of gene sets and apply the algorithm to a publicly available type 2 diabetes (T2D) expression dataset. Firstly, we calculate coexpression biweight midcorrelation coefficients between all gene pairs. Then, we select informative correlation pairs using the "differential coexpression threshold" strategy. Finally, we identify the differential coexpression gene modules using maximum clique concept and k-clique algorithm. We apply the proposed differential coexpression analysis method on simulated data and T2D data. Two differential coexpression gene modules about T2D were detected, which should be useful for exploring the biological function of the related genes.

1. Introduction

DNA microarray has been widely used as measurement tools in gene expression data analysis [1–4]. Gene expression profiling data from DNA microarray can detect the expression levels of thousands of genes simultaneously, providing an effective way for mining disease-related genes and revealing information of the regulatory networks and biological pathways of genes. Currently, the analysis of gene expression data can be divided into three levels: first, analysis of the expression level of individual genes, determining its function based on gene expression level changes under different experimental conditions: for example, the tumor type specific genes are identified according to the significance of difference in gene expression using the statistical hypothesis testing analysis method; second, study of gene interaction and coregulation through the combination of genes and grouping; and, third, an attempt to deduce the potential gene regulatory networks mechanism and explain the observed gene expression data.

Among the microarray data analysis methods, gene differential expression analysis is one of the most widely used types of analysis for disease research. Gene differential expression analysis method selects differentially expressed genes according to expression change value of a single gene. In fact, gene expression value change between normal samples and disease samples can be used to present the possibility of the relation between gene and disease. However, the traditional pathogenicity genes selection methods based on gene expression data treat each gene individually and interaction between them is not considered. Actually, genes and their protein products do not perform their functions in isolation [5, 6], but in cooperation. Functional changes such as alteration in tumor cell growth process, energy metabolism, and immune activity were accompanied with coexpression changes. Differentially expressed genes selection methods often focus only on the size of the single genes and

the relationship of individual genes and disease, ignoring a plurality of pathogenic genes of the complex disease as a gene module with disease related, as well as within the module gene [7].

Differential coexpression analysis, as a more comprehensive technique to the differential expression analysis, was raised to research gene regulatory networks and biological pathways of phenotypic changes through measure gene correlation changes between disease and normal conditions. Differential coexpression genes are defined as genes whose correlated expression pattern differs between classes [8]. The gene coexpression changes between different conditions indicate gene regulatory pathways and networks associated with disease. In gene differential coexpression analysis, a pair of gene expression datasets under disease and normal conditions is transformed to a pair of coexpression matrix in which links represent transcriptionally correlated gene pairs [5]. Until now, methods for differential coexpression analysis of gene expression data have been extensively researched, and multiple algorithms have been developed and tested [9–12]. In those gene differential coexpression analysis methods, the most common choice of similarity measurement is Pearson's correlation coefficients. However, Pearson's correlation is sensitive to outliers. So biweight midcorrelation (bicor) is considered to be a good alternative to Pearson's correlation since it is more robust to outliers [13].

In biomedical research, many complex diseases are contributed jointly by alterations of numerous genes; they often coordinate together as a functional biological pathway or network and are highly correlated. With recent interest of gene differential coexpression analysis in the gene network or module, gene module analysis has emerged as a novel holistic approach for microarray analysis. Somewhat large units, made up of genes, are more densely connected to each other than to the rest of the network, are often referred to as modules, and have been considered to be the essential structural units of real gene networks. There exists overlap among gene modules in large real networks.

Until now, there are many methods to find gene modules. For example, Butte and Kohane [14] proposed a systems-based approach called Entropy Minimization and Boolean Parsimony (EMBP) that identifies, directly from gene expression data, modules of genes that are jointly associated with disease. Kostka and Spang [15] used additive model to find differential coexpression gene modules. Prieto et al. [16] used altered expression based on improved additive model, optimal residual ratio, and minimum F-distribution to find differential coexpression gene modules. However, the microarray data contains a large number of genes; those methods need to search all gene expression data resulting in a large amount of computation; the process is very time-consuming even using optimized search algorithm.

The maximum clique analysis can avoid exhaustive search and quickly find maximum gene module with biological significance. The maximum clique problem (MCP) is a classical combinatorial optimization problem in graph theory. In 1957, Ross and Harary [17] first proposed the deterministic algorithm to solve the maximum clique problem. Since then some researchers had presented a variety of algorithms to solve this problem. The maximum clique problem is widely used in different areas, such as signal transmission, computer vision, and biological research. In this study, a gene coexpression network can be treated as a graph; gene is represented by vertex and coexpression relationship is represented by edge. We will use k-clique algorithm [18], which is an effective and deterministic method for uniquely identifying overlapping modules in large real networks. We first show some basic definitions. k-cliques, the central objects of k-clique algorithm investigation, are defined as complete (fully connected) subgraphs of k vertices. k-clique adjacency is as follows: two k-cliques are adjacent if they share some vertices. k-clique chain is as follows: a subgraph, which is the union of a sequence of adjacent k-cliques. We use k-clique algorithm to find gene cliques, and maximum clique concept is used to quickly find large gene modules which are made of k-clique chain. For the sake of convenience, we use the terms graph and community or network interchangeably, the former stressing the mathematical concept and the latter the application.

In this paper, we proposed a new approach for gene differential coexpression analysis in gene modules level based on combining biweight midcorrelation, differential coexpression threshold strategy, and maximum clique concept and k-clique analysis. Biweight midcorrelation measures the coexpression relationship between genes and the k-clique analysis with maximum clique concept quickly finds maximum disease-related module with biological significance. We use the approach to further investigate the gene module in order to gain insight into coexpression relationship between genes. The algorithm can find differential coexpression disease genes modules and global coexpression patterns are determined for type 2 diabetes expression dataset. As far as we know, no one has done this experiment.

The rest of the paper is organized as follows. Section 2 describes the methods proposed in this study. The biweight midcorrelation coefficients, "gene differential coexpression threshold" strategy, and threshold selection strategy are first presented, and the algorithm of k-clique is consequently given. Section 3 presents the experiment on simulated data and type 2 diabetes (T2D) in rats dataset. Section 4 concludes the paper and outlines directions of future work.

2. Methods

2.1. Biweight Midcorrelation for Differential Coexpression. Differential coexpression analysis usually requires the definition of "distance" or "similarity" between measured datasets, the most common choice being Pearson's correlation coefficients. However, Pearson's correlation coefficient is sensitive to outliers [13]. Biweight midcorrelation is considered to be a good alternative to Pearson's correlation since it is more robust to outliers. Example of a gene expression matrix is as follows:

$$\begin{bmatrix} & \text{Gene1} & \text{Gene2} & \cdots & \text{Gene}p \\ Z_1 & X_{11} & X_{12} & \cdots & X_{1p} \\ Z_2 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_n & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}. \quad (1)$$

For each sample Z_i, we measure expression levels for a set of genes, so X_{ij} is the measurement of the expression level of the jth gene for the ith sample, where $j = 1, \ldots, p$. The xth column vector of matrix represents gene expression profile of gene X. In order to define the biweight midcorrelation (bicor) [13] of two numeric vectors $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$, we first define u_i, v_i with $i = 1, \ldots, m$:

$$u_i = \frac{x_i - \operatorname{med}(x)}{9\operatorname{mad}(x)}, \qquad v_i = \frac{y_i - \operatorname{med}(y)}{9\operatorname{mad}(y)}, \tag{2}$$

where $\operatorname{med}(x)$ is the median of vector x, $\operatorname{mad}(x)$ is the median absolute deviation of vector x, $\operatorname{mad}(x)$ is the median of new numeric vector in which each number is absolute difference between original vector value and $\operatorname{med}(x)$; this leads us to the definition of $\operatorname{mad}(x)$ and weight w_i for x_i, which are

$$\operatorname{mad}(x) = \operatorname{med}\left(\left|x_i - \operatorname{med}(X)\right|\right), \qquad w_i^{(x)} = \left(1 - u_i^2\right)^2 I\left(1 - \left|u_i\right|\right), \tag{3}$$

where the indicator $I(1 - |u_i|)$ takes 1 if $1 - |u_i| > 0$ and 0 otherwise. Thus, the weight $w_i^{(x)}$ is close to 1 if x_i is close to $\operatorname{med}(x)$, approaches 0 when x_i differs from by nearly $9\operatorname{mad}(x)$, and is 0 if x_i differs from $\operatorname{med}(x)$ by more than $9\operatorname{mad}(x)$. An analogous weight $w_i^{(x)}$ can be defined for y_i. Given the weights, we can define biweight midcorrelation of x and y as

$$\operatorname{bicor}(x, y) = \sum_{i=1}^{m} \left(x_i - \operatorname{med}(x)\right) w_i^{(x)} \left(y_i - \operatorname{med}(y)\right) w_i^{(y)} \cdot \left(\sqrt{\sum_{j=1}^{m} \left[\left(x_j - \operatorname{med}(x)\right) w_j^{(x)}\right]^2} \cdot \sqrt{\sum_{k=1}^{m} \left[\left(y_k - \operatorname{med}(y)\right) w_k^{(y)}\right]^2} \right)^{-1}. \tag{4}$$

It should be noted that the equations of biweight midcorrelation do not involve an explicit identification of outliers, and all elements whose weight $w_i = 0$ can be considered outliers. The user can also set up the maximum allowed proportion of outliers using the argument "maxPOutliers"; the "maxPOutliers" is interpreted as the maximum proportion of low and high outliers separately. For the value of bicor from –1 to 1, –1 represents the maximum negative correlation and 1 represents the maximum positive correlation. Zero represents irrelevant correlation.

2.2. The "Differential Coexpression Threshold" Strategy. We used biweight midcorrelation to measure every pair of genes in the gene expression dataset and get a gene coexpression matrix. The gene coexpression matrix is a square and symmetric matrix P whose rows and columns correspond to the genes and whose element P_{ij} denotes the coexpression relationship between genes i and j. In this paper, we use A_{GN} which represents gene coexpression adjacency matrix in normal conditions and A_{GD} which represents gene coexpression adjacency matrix in disease condition. To find differential coexpression gene modules which are coexpressed in normal condition and not related to disease condition, we set two thresholds T_1 for adjacency matrix A_{GN} in normal condition and T_2 for adjacency matrix A_{GD} in disease condition. $A_{GN}(i, j)$ is set to 1 if value of $A_{GN}(i, j)$ is greater than or equal to T_1; otherwise, $A_{GN}(i, j)$ is set to 0 and $A_{GD}(i, j)$ is set to 1 if value of $A_{GD}(i, j)$ is less than or equal to T_2; otherwise, $A_{GD}(i, j)$ is set to 0. We integrated A_{GN} and A_{GD} into a matrix A_G after we had intersection of the corresponding elements of A_{GN} and A_{GD}. $A_G(i, j) = 1$ means coexpression value of gene i and gene j in A_{GN} is greater than or equal to T_1, and coexpression value of genes i and j in A_{GD} is less than or equal to T_2. $A_{GN}(i, j)$ also can be set to 1 if value of $A_{GN}(i, j)$ is less than or equal to T_1 and $A_{GD}(i, j)$ is set to 1 if value of $A_{GD}(i, j)$ is greater than or equal to T_2. The method is shown in (5).

With the above mentioned strategy, we also set $A_G(i, j) = 1$ if the absolute value of $A_{GN}(i, j)$ subtracting $A_{GD}(i, j)$ is greater than or equal to T_3 and the absolute value of $A_{GN}(i, j)$ is greater than or equal to the absolute value of $A_{GD}(i, j)$ simultaneously. This is a special type of coexpression change. In reality, coexpression reversal probably has biological significance. The coexpression reversal between normal condition and disease condition has advantage in disease. For example, the coexpression of *p53* and *Klf4* recently reported that the positive or negative correlation between these two genes determines the outcome of DNA damage, DNA repair, or apoptosis [19]. We believe that our attention to this special coexpression change will help to explore subtle mechanisms involved in genes transcriptional regulation. We excavated maximum cliques which have biological significance from A_G adjacency matrix to further investigate gene regulatory networks. Consider the following:

$$\begin{aligned}
&\text{if } A_{GN}(i, j) \geq T_1, \text{ then } A_{GN}(i, j) = 1,\\
&\text{else } A_{GN}(i, j) = 0;\\
&\text{if } A_{GD}(i, j) \leq T_2, \text{ then } A_{GD}(i, j) = 1,\\
&\text{else } A_{GD}(i, j) = 0;\\
&A_G(i, j) = A_{GN}(i, j) \cap A_{GD}(i, j)\\
&\text{if } A_{GN}(i, j) \leq T_1, \text{ then } A_{GN}(i, j) = 1,\\
&\text{else } A_{GN}(i, j) = 0;\\
&\text{if } A_{GD}(i, j) \geq T_2, \text{ then } A_{GD}(i, j) = 1,\\
&\text{else } A_{GD}(i, j) = 0;\\
&A_G(i, j) = A_{GN}(i, j) \cap A_{GD}(i, j).
\end{aligned} \tag{5}$$

2.3. The Threshold Selection Strategy. The two real value adjacency matrixes are transformed into a binary matrix which contains two elements 0 and 1 only. Choosing different thresholds will lead to different results; too large T_1 threshold or too small T_2 threshold will lead to small link number, low density clique, and lost biological significance cliques. On the other hand, too small T_1 or too large T_2 will lead to many overlapping cliques. They are not helpful for finding biological significance differential coexpression gene disease-related modules. In fact, how to choose a reasonable threshold in conversion process is a problem which needs to be further studied. Generally, the selection of the threshold can be based on the proportion of outliers in the figure or the density of graph. The outlier is the point which is not connected to any edges. The density is defined as the ratio of number of edges to the maximum possible number of edges in the graph. The density of clique is 1.

For gene expression data analysis, closely linked functional module is not the strict sense of maximum clique due to the lack of certain section. In this paper, we use density to measure approximation degree of functional module with gene differential coexpression clique, which may be having more biological significance.

2.4. The Maximum Clique Concept and k-Clique Algorithm. Graph theoretical concepts are useful for the description and analysis of interactions and relationships in biological systems. In gene coexpression graph, gene is represented by vertex and coexpression relationship by edge. $G = (V, E)$ is an arbitrary undirected and weighted graph unless otherwise specified in graph theoretical concepts. $V = \{1, 2, \ldots, n\}$ is the vertex set of G, and E is the edge set of G. For each vertex $i \in V$, a positive weight w_i is associated with i. $A_G = (a_{ij})_{n\times n}$ is the adjacency matrix of G, where $a_{ij} = 1$ if $(i, j) \in E$ is an edge of G, and $a_{ij} = 0$ if $(i, j) \notin E$. Genes and relationship between genes are represented by vertex and edge, respectively.

A graph $G = (V, E)$ is complete if all its vertices are pairwise adjacent; that is, for all $i, j \in V$, $(i, j) \in E$. A clique C is a subset of V such that $G(C)$ is complete. The maximum clique problem asks for a clique of maximum weight. An independent set (stable set and vertex packing) is a subset of V, whose elements are pairwise nonadjacent. The maximum independent set problem asks for an independent set of maximum cardinality. The size of a maximum independent set is the stability number of G (denoted by $\alpha(G)$). The maximum weight independent set problem asks for an independent set of maximum weight. A maximum clique means a clique which is a subset of the nodes in V in which every pair of nodes in the subset is joined by an edge and is not a proper subset of any other cliques [20].

In application, the identification of maximal cliques is often of limited interest since the requirement of complete connectivity is so restrictive. When dealing with imperfect systems or with experimental data, we may need to consider more general notions of cohesive subgroups. In this paper, we consider different notions of cohesive subgroups that include n-cliques, k-plexes, and λ-sets [18]. It is well known that the nodes of large real networks have a power law degree distribution [21]. Most real networks typically contain parts in which the nodes (units) are more highly connected to each other compared to the rest of the network. The sets of such nodes are usually called clusters, communities, cohesive groups, or modules [22–26], which have no widely accepted unique definition. The basic observation on which our modules definition relies is that a typical gene differential coexpression module consists of several complete (fully connected) subcliques that tend to share many of their nodes. To find meaningful communities, several basic requirements should be satisfied: it cannot be too restrictive, should be based on the density of links, is required to be local, should not yield any cut-node or cut link (whose removal would disjoin the community), and, of course, should allow overlaps. We employ the community definition specified above because none of the others in the literature satisfy all these requirements simultaneously [27–29].

k-clique algorithm for detecting gene differential coexpression modules in a network has been published in the paper [26]. k-clique algorithm is also named clique percolation method. The existing divisive and agglomerative methods recently used for large real networks have some disadvantages. Divisive methods cut the network into smaller and smaller pieces; each node is forced to remain in only one community and be separated from its other communities, most of which then necessarily fall apart and disappear [27, 30]. The agglomerative [31] method has the same problem. The k-clique algorithm has demonstrated the advantages over the divisive method and agglomerative method. In the algorithm, although the numerical determination of the full set of k-clique communities is a polynomial problem, the algorithm is exponential and significantly more efficient for the graphs corresponding to actual data. The k-clique algorithm first locates all cliques (maximal complete subgraphs) of the network and then identifies the communities by carrying out a standard component analysis of the clique-clique overlap matrix [28]. The k-clique algorithm uses the threshold probability $d(k)$ (critical point) of k-clique percolation to find all maximal complete subgraphs. The critical point is shown in (6), where N is the number of genes or vertex of graph:

$$d(k) = \frac{1}{[(k-1)N]^{1/(k-1)}}. \tag{6}$$

The k-clique algorithm gives two plausible choices to measure the size of the largest k-clique percolation cluster in (7) and (8). The most natural one, which we denote by N^*, is the number of vertices belonging to this cluster. ϕ is an order parameter associated with this choice as the relative size of that cluster:

$$\phi = \frac{N^*}{N}. \tag{7}$$

The other choice is the number L^* of k-cliques of the largest k-clique percolation cluster. The associated order parameter is again the relative size of this cluster:

$$\varphi = \frac{L^*}{L}, \tag{8}$$

where L denotes the total number of k-cliques in the graph. L can be estimated as

$$L \approx \binom{N}{k} d^{k(k-1)/2} \approx \frac{N^k}{k!} d^{k(k-1)/2}. \quad (9)$$

In this paper, we use the biweight midcorrelation for constructing binary networks. Two-condition coexpression adjacency networks can always be transformed into a binary one by ignoring any directionality in the links and keeping only those stronger than a threshold weight. Then, the concept of maximum clique and k-clique algorithm were used to find gene differential coexpression modules. We named the proposed method "BMKC" (biweight midcorrelation and k-clique algorithm) method. Changing the threshold is like changing the resolution with which the community structure is investigated: by increasing, the communities start to shrink and fall apart. A very similar effect can be observed by changing the value of k as well: increasing k makes the communities smaller and more disintegrated but, at the same time, also more cohesive. More details about k-clique algorithm can be found in [28, 32].

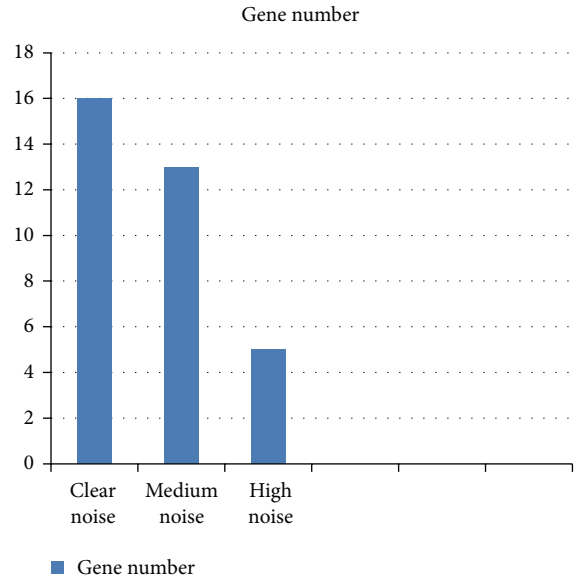


Figure 1: The column bar graph shows the effect of the noise parameter σ on the size of the gene group found by our algorithm.

3. Results

3.1. Experiment Result on Simulated Datasets. We first evaluate the algorithm in a supervised setting. We generate a control group of 30 samples and a disease group of another 30 samples, both consisting of 120 genes. For the control group, 20 coexpressed genes are sampled directly from the biweight midcorrelation. We focus on whether k-clique algorithm can find coexpression gene modules from the background of noise. We first draw a vector with 20 rows and a vector with 30 columns from a standard normal distribution. The actual expression levels are obtained by adding independent errors sampled from a normal distribution with mean zero and standard deviation (SD) σ. These 20 genes form the target pattern. We then hide them in 100 additional noise genes, which are sampled independent and identically distributed (i.i.d.) from a standard normal distribution. The disease group is simulated by 120 independent noise genes drawn from a standard normal only.

In the above setting, we use SD σ to tune the strength of the signal resulting from the 20 coexpressed genes. To observe its effect in detail, we use three different values: for a clear signal, $\sigma = 1/10$, for medium noise, $\sigma = 1/4$, and, for high noise, $\sigma = 1$. To guard for sampling effects, we repeat each procedure 50 times and average the results, which are displayed in Figure 1. One can see that, for the clear and medium signal, the algorithm can recover the differentially coexpressed genes modules reliably. Also, depending on the prominence of the signal, the influence of σ is more or less pronounced. In an exploratory analysis setting with several hidden patterns, we could use T_1, T_2, and T_3 to control the size of target patterns.

3.2. Analyzing a Type 2 Diabetes (T2D) in Rats. As a real-world application, we apply the BMKC method to a pair of type 2 diabetes (T2D) rats datasets (dataset pair T), which has been published in study [33]. Dataset pair T is from dataset GSE3068 of Gene Expression Omnibus (GEO) database. Yu et al. preprocessed dataset GSE3068. Dataset pair T includes 4765 genes in 10 disease samples and 10 normal samples. We use our algorithm to find differential coexpression modules in the type 2 diabetes.

For computational efficiency, we calculate the sum of each row or column of adjacency matrix; the sum means the number of genes related to the gene. The gene is outlier if the sum is zero. First, we calculate the sum of each row or column of the adjacency matrix and delete the outlier. Second, we calculate the sum of each row or column of the adjacency matrix and discard the lower 50% of them. We set $T_3 = 1.3$ and the minimum number of each clique to four. Finally, we apply our algorithm to the remainder genes and excavate two differential coexpression modules. Tables 1 and 2 list each gene symbol in the clique. The adjacency graphs of each differential coexpression module are shown in Figures 2 and 3. From these two figures, we can see that the cliques in each of the differential coexpression modules are overlapping, forming a closely related module. In normal condition, the absolute bicor value of total of 24 genes in modules distributes from 0.78 to 0.97. Yet, in disease condition, the absolute bicor value of genes distributes from 0.21 to 0.09. In the results of our study, the gene differential coexpression modules included quite a number of previously reported T2D-related genes: *Hifla* and *Sirt2* [34], *Smarca4* [35], *Sh2b2* [36], *Madd* [37], and *Rxrb* [38]. Despite not being previously reported to be related with T2D, other genes in the modules should receive adequate attention for their distinct traits from the perspective of differential coexpression. Further studies on the transcriptional mechanisms and functional consequences could pay more attention to these genes.

Table 1: Genes in each clique.

Clique number	Gene symbol				
1	***Hifla****	*Ifngr1*	*RGD1305094*	*Tenc1*	***Sirt2***
2	*Clcn1*	***Smarca4***	*Zkscan17*	*Rpl27a*	***Sirt2***
3	***Hifla***	*Ifngr1*	*Pfkfb3*	*Tenc1*	***Sirt2***
4	***Sh2b2***	*Pcsk5*	*Lamc1*	*Rpl27a*	***Sirt2***
5	*Lamc1*	***Smarca4***	*Zkscan17*	***Sirt2***	*Rpl27a*
6	***Hifla***	*RGD130504*	*Mxd4*	***Sirt2***	
7	*Tra1*	***Smarca4***	*Zkscan17*	***Sirt2***	

*Bold genes refer to the previously reported T2D-related genes. The other genes are identified in the differential coexpression modules.

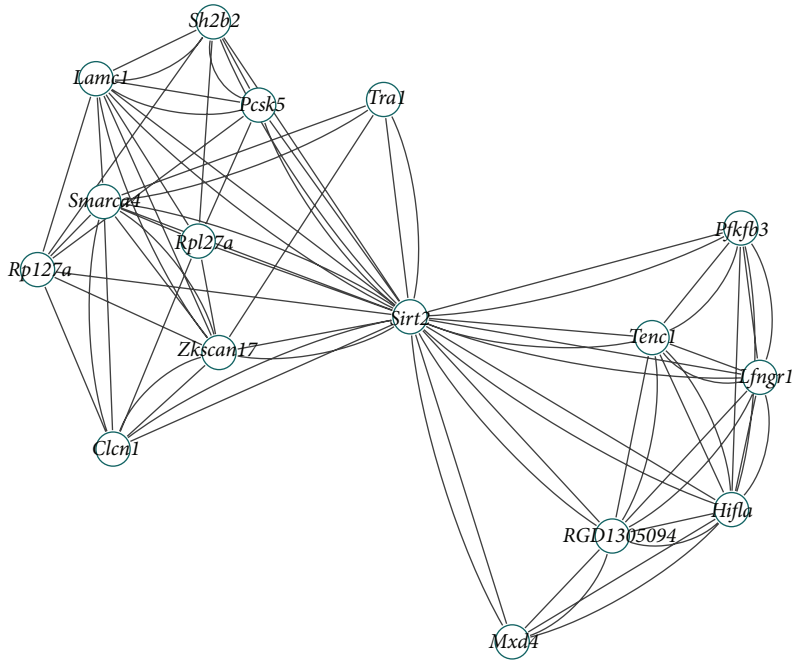


Figure 2: The adjacency graph of first gene differential coexpression module.

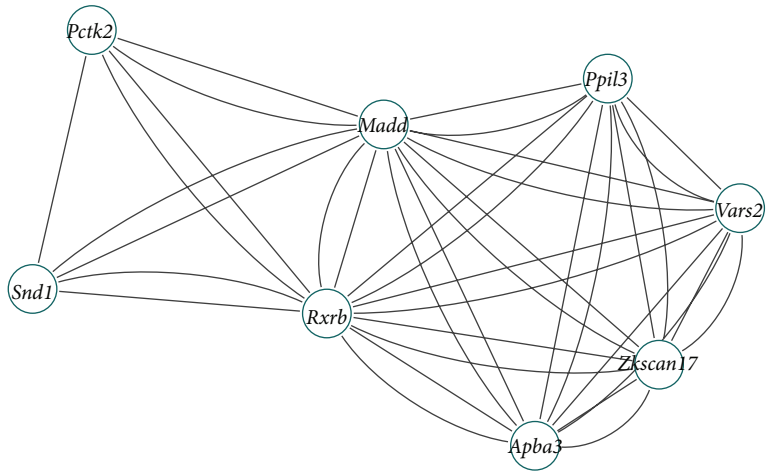


Figure 3: The adjacency graph of second gene differential coexpression module.

Table 2: Genes in each clique.

Clique number	Gene symbol					
1	*Vars2*	*Apba3*	***Madd***	*Zkscan17*	*Ppil3*	***Rxrb***
2	*Snd1*	***Madd***	*Pctk2*	***Rxrb***		

3.3. Significance Analysis of the BMKC Method. Naturally, the question of whether our findings are artifacts of the high dimensionality of the data arises. To assess this question, we apply a permutation procedure. Under the null hypothesis, we assume that all genes are mutually independent in both conditions groups. We heuristically sample from the null hypothesis by (group-wise) shuffling the expression values for each gene independently. Thus, random expression data are generated where all covariance structures are removed. Applying our algorithm to the randomized data yields one random score. We repeat the procedure 1000 times. Using the empirical distribution of the simulated scores, the simulated score means the global total sum of differential coexpression change of each gene in modules. We calculate P values for the observed scores in the nonpermuted data. For each of the patterns in the type 2 diabetes example, we only observe one random score smaller than the biological one. This corresponds to an empirical P value of 0.001. Hence, it is unlikely that the observed differential coexpression is a chance artifact.

4. Conclusions

In this paper, we proposed a new approach in gene sets level for differential coexpression analysis, which combine biweight midcorrelation and threshold selection strategy and also applied maximum clique concept with k-clique algorithm to the specific gene set to further investigate gene regulatory networks. Biweight midcorrelation is more robust for outliers and threshold selection strategy is an effective preprocess step of the proposed method. Experimental results on simulated datasets show that our method had good performance. We apply the proposed BMHT method to real dataset designed for T2D study, and two differential coexpression gene modules were detected, which should be a useful resource for T2D study and could be used for exploring the biological function of the related genes. In the future, we will focus on how to quickly excavate gene differential coexpression module from gene coexpression adjacency matrix.

Conflict of Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the National Science Foundation of China under Grants nos. 61272339, 61271098, and 31301101 and the Key Project of Anhui Educational Committee, under Grant no. KJ2012A005.

References

[1] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature Reviews Genetics*, vol. 7, no. 1, pp. 55–65, 2006.

[2] 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.

[3] M. P. S. Brown, W. N. Grundy, D. Lin et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 1, pp. 262–267, 2000.

[4] A. Sturn, J. Quackenbush, and Z. Trajanoski, "Genesis: cluster analysis of microarray data," *Bioinformatics*, vol. 18, no. 1, pp. 207–208, 2002.

[5] J. K. Choi, U. Yu, O. J. Yoo, and S. Kim, "Differential coexpression analysis using microarray data and its application to human cancer," *Bioinformatics*, vol. 21, no. 24, pp. 4348–4355, 2005.

[6] J. Rachlin, D. D. Cohen, C. Cantor, and S. Kasif, "Biological context networks: a mosaic view of the interactome," *Molecular Systems Biology*, vol. 2, article 66, 2006.

[7] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, "Coexpresion analysis of human genes across many microarray data sets," *Genome Research*, vol. 14, no. 6, pp. 1085–1094, 2004.

[8] A. Reverter, A. Ingham, S. A. Lehnert et al., "Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer," *Bioinformatics*, vol. 22, no. 19, pp. 2396–2404, 2006.

[9] S. L. Carter, C. M. Brechbühler, M. Griffin, and A. T. Bond, "Gene co-expression network topology provides a framework for molecular characterization of cellular state," *Bioinformatics*, vol. 20, no. 14, pp. 2242–2250, 2004.

[10] M. J. Mason, G. Fan, K. Plath, Q. Zhou, and S. Horvath, "Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells," *BMC Genomics*, vol. 10, article 327, 2009.

[11] T. F. Fuller, A. Ghazalpour, J. E. Aten, T. A. Drake, A. J. Lusis, and S. Horvath, "Weighted gene coexpression network analysis strategies applied to mouse weight," *Mammalian Genome*, vol. 18, no. 6-7, pp. 463–472, 2007.

[12] J. M. Freudenberg, S. Sivaganesan, M. Wagner, and M. Medvedovic, "A semi-parametric Bayesian model for unsupervised differential co-expression analysis," *BMC Bioinformatics*, vol. 11, article 234, 2010.

[13] R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, San Diego, Calif, USA, 1997.

[14] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Proceedings of the Pacific Symposium on Biocomputing*, vol. 5, pp. 415–426, 2000.

[15] D. Kostka and R. Spang, "Finding disease specific alterations in the co-expression of genes," *Bioinformatics*, vol. 20, supplement 1, pp. i194–i199, 2004.

[16] C. Prieto, M. J. Rivas, J. M. Sánchez, J. López-Fidalgo, and J. de Las Rivas, "Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes," *Bioinformatics*, vol. 22, no. 9, pp. 1103–1110, 2006.

[17] I. C. Ross and F. Harary, "On the determination of redundancies in sociometric chains," *Psychometrika*, vol. 17, no. 2, pp. 195–208, 1952.

[18] S. Wasserman and K. Faust, *Social Network Analysis, Methods and Applications*, Cambridge University Press, Cambridge,UK, 1994.

[19] Z. Qibing, H. Yuan, Z. Qimin, S. Yan, and L. Zhihua, "Role for Krüppel-like factor 4 in determining the outcome of p53 response to DNA damage," *Cancer Research*, vol. 69, no. 21, pp. 8284–8292, 2009.

[20] I. M. Bomze, M. Budinich, P. M. Pardalos, and M. Pelillo, "The maximum clique problem," in *Handbook of Combinatorial Optimization*, pp. 1–74, 1999.

[21] B. Albert-László and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[22] J. Scott, *Social Network Analysis: A Handbook*, Sage Publications, London, UK, 2nd edition, 2000.

[23] R. M. Shiffrin and K. Borner, "Mapping knowledge domains," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, supplement 1, pp. 5183–5185, 2004.

[24] B. S. Everitt, *Cluster Analysis*, Edward Arnold, London, UK, 3th edition, 1993.

[25] S. Knudsen, *A Guide to Analysis of DNA Microarray Data*, Wiley-Liss, New York, NY, USA, 2nd edition, 2004.

[26] M. E. J. Newman, "Detecting community structure in networks," *The European Physical Journal B*, vol. 38, no. 2, pp. 321–330, 2004.

[27] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[28] M. G. Everett and S. P. Borgatti, "Analyzing clique overlap," *Connections*, vol. 21, pp. 49–61, 1998.

[29] S. Kosub, "Local density," in *Network Analysis*, pp. 112–142, Springer, Berlin, Germany, 2005.

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

[31] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, Article ID 066133, 5 pages, 2004.

[32] I. Derényi, G. Palla, and T. Vicsek, "Clique percolation in random networks," *Physical Review Letters*, vol. 94, no. 16, Article ID 160202, 2005.

[33] H. Yu, B.-H. Liu, Z.-Q. Ye, C. Li, Y.-X. Li, and Y.-Y. Li, "Link-based quantitative methods to identify differentially coexpressed genes and gene Pairs," *BMC Bioinformatics*, vol. 12, article 315, 2011.

[34] J. C. Milne, P. D. Lambert, S. Schenk et al., "Small molecule activators of SIRT1 as therapeutics for the treatment of type 2 diabetes," *Nature*, vol. 450, no. 7170, pp. 712–716, 2007.

[35] L. L. Nguyen, A. D. Kriketos, D. P. Hancock, I. D. Caterson, and G. S. Denyer, "Insulin resistance does not influence gene expression in skeletal muscle," *Journal of Biochemistry and Molecular Biology*, vol. 39, no. 4, pp. 457–463, 2006.

[36] M. Li, Z. Li, D. L. Morris, and L. Rui, "Identification of SH2B2β as an inhibitor for SH2B1- and SH2B2α-promoted Janus kinase-2 activation and insulin signaling," *Endocrinology*, vol. 148, no. 4, pp. 1615–1621, 2007.

[37] J. Dupuis, C. Langenberg, I. Prokopenko, and R. Saxena, "The genetics of type 2 diabetes: what have we learned from GWAS?" *Nature*, vol. 1212, pp. 59–77, 2010.

[38] S. Sookoian and C. J. Pirola, "Metabolic syndrome: from the genetics to the pathophysiology," *Current Hypertension Reports*, vol. 13, no. 2, pp. 149–157, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 678764, 10 pages
http://dx.doi.org/10.1155/2015/678764

Research Article

AcconPred: Predicting Solvent Accessibility and Contact Number Simultaneously by a Multitask Learning Framework under the Conditional Neural Fields Model

Jianzhu Ma[1] and Sheng Wang[1,2]

[1] *Toyota Technological Institute at Chicago, 6045 S. Kenwood Avenue, Chicago, IL 60637, USA*
[2] *Department of Human Genetics, University of Chicago, E. 58th Street, Chicago, IL 60637, USA*

Correspondence should be addressed to Sheng Wang; wangsheng@ttic.edu

Received 27 December 2014; Accepted 11 March 2015

Academic Editor: Min Li

Motivation. The solvent accessibility of protein residues is one of the driving forces of protein folding, while the contact number of protein residues limits the possibilities of protein conformations. The de novo prediction of these properties from protein sequence is important for the study of protein structure and function. Although these two properties are certainly related with each other, it is challenging to exploit this dependency for the prediction. *Method*. We present a method AcconPred for predicting solvent accessibility and contact number simultaneously, which is based on a shared weight multitask learning framework under the CNF (conditional neural fields) model. The multitask learning framework on a collection of related tasks provides more accurate prediction than the framework trained only on a single task. The CNF method not only models the complex relationship between the input features and the predicted labels, but also exploits the interdependency among adjacent labels. *Results*. Trained on 5729 monomeric soluble globular protein datasets, AcconPred could reach 0.68 three-state accuracy for solvent accessibility and 0.75 correlation for contact number. Tested on the 105 CASP11 domain datasets for solvent accessibility, AcconPred could reach 0.64 accuracy, which outperforms existing methods.

1. Introduction

The solvent accessibility of a protein residue is the surface area of the residue that is accessible to a solvent, which was first described by Lee and Richards [1] in 1971. During the process of protein folding, the residue solvent accessibility plays a very important role as it is related to the spatial arrangement and packing of the protein [2], which is depicted as the hydrophobic effect [3]. Specifically, the trends of the hydrophobic residues to be buried in the interior of the protein and the hydrophilic residues to be exposed to the solvent form the hydrophobic effect that functions as the driving force for the folding of monomeric soluble globular proteins [4–6].

Solvent accessibility can help protein structure prediction in two aspects. (1) Since solvent accessibility is calculated on all-atom protein structure coordinates, it encodes the global information of the 3D protein structure into a 1D feature, which makes solvent accessibility as an excellent piece of complementary information to the other local 1D features such as secondary structure [7–9], structural alphabet [10–12], or backbone torsion angles [13–15]. (2) Compared to the other global information such as the contact map [16, 17] or the distance map [18, 19], solvent accessibility shares similar property of the other local 1D feature that it could be predicted into a relatively accurate level [20]. Therefore, the predicted solvent accessibility has been widely utilized for detection as well as threading of remote homologous proteins [21–23] and quality assessment of protein models [24, 25].

The contact number is yet another kind of 1D feature that encodes the 3D information, which is related to, but different from, solvent accessibility [26]. The contact number of a protein residue is actually the result of protein folding. It has been suggested that, given the contact number for

each residue, the possibilities of protein conformations that satisfy the contact number constraints are very limited [27]. Thus, the predicted contact numbers of a protein may serve as useful restraints for de novo structure prediction [26] or contact map prediction [28].

To predict the protein solvent accessibility, most methods first discretize it into two- or three-state labels based on the continuous relative solvent accessibility value [20]. Then these methods apply a variety of learning approaches for the prediction, such as neural networks [29–34], SVM (support vector machine) [35–37], Bayesian statistics [38], and nearest neighbor [20, 39]. Some other methods also attempt to directly predict the continuous absolute or relative solvent accessibility value [14, 34, 40–42].

Comparing with the solvent accessibility prediction, there are much fewer methods that deal with the prediction of contact number. For example, Kinjo et al. [26] employ linear regression analysis, Pollastri et al. [32] use neural networks, and Yuan [43] applies SVM.

Since a high dependency between the adjacent labels for both solvent accessibility and contact number exists [44], it is hard to utilize this information based on the previous proposed computational methods. For instance, neural network methods usually do not take the interdependency relationship among the labels of adjacent residues into consideration. Similarly, it is also challenging for SVM to deal with this dependency information [45]. Although hidden Markov model (HMM) [44] is capable of describing this dependency, it is challenging for HMM to model the complex nonlinear relationship between input protein features and the predicted solvent accessibility labels, especially when a large amount of heterogeneous protein features is available [45].

Recently, ACCpro5 [46] could reach almost perfect prediction of protein solvent accessibility by the aid of structural similarity in the protein template database. However, such approach might not perform well on those de novo folds or the sequences that cannot find any similar proteins in the database.

Although solvent accessibility and contact number are two different quantities, they are certainly related with each other, both reflecting the hydrophobic or hydrophilic atmosphere of each residue in the protein structure [26]. For example, a residue with a large contact number would probably be buried inside the core, whereas a residue with a small contact number would probably be exposed to the solvent. Therefore, a learning approach that could utilize this relationship to extract the universal representation of the features would be beneficial.

Here we present AcconPred (solvent accessibility and contact number prediction), available at http://ttic.uchicago.edu/~majianzhu/AcconPred_package_v1.00.tar.gz based on a shared weight multitask learning framework under the CNF (conditional neural fields) model. As a recently invented probabilistic graphical model, CNF [47] has been used for a variety of bioinformatics tasks [21–23, 45, 48–52]. Specifically, CNF is a perfect integration of CRF (Conditional Random Fields) [53] and neural networks. Besides modeling the nonlinear relationship between the input protein features and the predicted labels as what neural network does, CNF can also model the interdependency among adjacent labels as what CRF does.

It has been shown that a unified neural network architecture, trained simultaneously on a collection of related tasks, provides more accurate labelings than a network trained only on a single task [54]. A study by Caruana thus demonstrates the power of multitask learning that could extract the universal representation of the input features [55]. In AcconPred, we integrate multitask learning framework under the CNF model by sharing the weight of the neuron functions between the two tasks, followed by a stochastic gradient descent for training the parameters.

Last but not least, AcconPred can provide a probability distribution over all the possible labels. That is, instead of predicting a single label at each residue, AcconPred will generate the label probability distribution for solvent accessibility and contact number. Our testing data shows that AcconPred achieves better accuracy on solvent accessibility prediction and higher correlation on contact number prediction than the other methods.

2. Method

2.1. Preliminary Definition

2.1.1. Calculating Solvent Accessibility from Native Protein Structure. We applied DSSP [7] to calculate the absolute accessible surface area for each residue in a protein. The relative solvent accessibility (RSA) of the residue X is calculated through dividing the absolute accessible surface area by the maximum solvent accessibility which uses Gly-X-Gly extended tripeptides [56]. In particular, these values are 210 (Phe), 175 (Ile), 170 (Leu), 155 (Val), 145 (Pro), 115 (Ala), 75 (Gly), 185 (Met), 135 (Cys), 255 (Trp), 230 (Tyr), 140 (Thr), 115 (Ser), 180 (Gln), 160 (Asn), 190 (Glu), 150 (Asp), 195 (His), 200 (Lys), and 225 (Arg), in units of $Å^2$.

With the relative solvent accessibility value, the classification was divided into three states, say, buried (B), intermediate (I), and exposed (E), as in the literatures [14, 20]. In this work, the usage of 10% for B/I and 40% for I/E in the 3-state definition is based on the following two facts: (1) such division is close to the definition of previous method [20]; (2) at this cutoff, the background distribution for the three states in our training data is close to 1 : 1 : 1. A more comprehensive interpretation for this 10%/40% threshold is described in Results and shown in Figure 2.

2.1.2. Calculating Contact Number from Native Protein Structure. To calculate the contact number for each residue, we followed similar definition from previous works [26, 43]. Basically, the contact number (CN) of the ith residue in a protein structure is the number of C-beta atoms from the other residues (excluding 5 nearest-neighbor residues) within the sphere of the radius 7.5 Å centered at the C-beta atom of the ith residue. We also limit the maximal contact number as 14 if the observed contact number is above 14, because such cases are rare in our training data. So for each residue, there are 15 states of contact number in total.

2.2. Datasets

2.2.1. Training and Validation Data. Training and validation data were extracted from all monomeric, globular, and nonmembrane protein structures. They were downloaded from Protein Data Bank (PDB) [57] dated before May 1, 2014. The monomeric proteins were extracted according to the "Remark 350 Author Determined Biological Unit: Monomeric" recorded in the PDB file. To exclude those nonglobular proteins, we calculated the buried residue ratio (i.e., the percentage of the residues in buried state) for each protein and removed those proteins with <10% buried residue ratio. To exclude those membrane proteins, the PDBTM database [58] was employed.

The reason for using monomeric protein to predict solvent accessibility is based on the fact that the patterns in the surface of the monomeric proteins are different from those in the interface of the oligomeric proteins [59]. Again, the reason why we exclude the membrane proteins is that they have the opposite solvent accessibility pattern to those monomeric, globular soluble proteins. Furthermore, the 10% buried residue ratio cutoff was derived from statistics for the globular protein database [60].

Finally, we excluded proteins with length less than 50, having chain-breaks in the middle, and the 40% sequence identity was applied to remove redundancy. So in total we have 5729 monomeric, globular, and nonmembrane protein structures as our training and validation dataset (5-cross validation). The 5729 PDB IDs included in the training and validation datasets could be found in the Supplementary Material available online at http://dx.doi.org/10.1155/2015/678764.

2.2.2. Testing Data. The testing data were collected from the CASP11 [61] targets containing 105 domains. Note that all CASP11 targets were released after May 1, 2014. The PDB structures for the 105 CASP11 testing datasets could be found in the Supplementary Files.

In order to compare with the existing programs, we further included the dataset from Yuan [43] as the testing data for contact number prediction. The 945 PDB IDs included in the Yuan dataset could be found in the Supplementary Files.

2.3. Protein Features. A variety of protein features have been studied by [14, 29–32, 41, 62, 63] to predict the solvent accessibility or the contact number. They could be categorized into three classes: evolution related, structure related, and amino acid related features, which will form our feature vector $F(i)$ for residue i. Furthermore, since the solvent accessibility or the contact number for a certain residue could be influenced by its nearby residues in sequence, we then introduce a windows size k to capture this information. That is, we take the feature vectors from $F(i-k), F(i-k+1), \ldots, F(i), \ldots, F(i+k-1), F(i+k)$ as the final input features for residue i. In this work we set the windows size $k = 5$.

2.3.1. Evolution Related Features. Solvent accessibility as well as contact number of a residue has a strong relationship with the residue's substitution and evolution. Residues in the buried core and residues on the solvent-exposed surfaces were shown to have different substitution patterns due to different selection pressure [64]. Evolution information such as PSSM (position specific scoring matrix) and PSFM (position specific frequency matrix) generated by PSI-BLAST [65] has been used and proved to enhance the prediction performance. Here we use different evolution information from the HHM file generated by HHpred [66]. In particular, it first invokes PSI-BLAST with five iterations and E-value 0.001 and then computes the homology information for each residue combined with a context-specific background probability [67]. Overall, for each residue, we have $40 = 20 + 20$ evolution related features.

2.3.2. Structure Related Features. Local structural features are also very useful in predicting solvent accessibility, as indicated in [41]. Here we use the predicted secondary structure elements (SSEs) probability as the structure related features for each residue position. In particular, we use both 3-class and 8-class SSEs. The 3-class SSE is predicted by PSIPRED [8] which is more accurate but contains less information, while the 8-class secondary structure element is predicted by RaptorX-SS8 [45] which is less accurate but contains more information. Overall, for each residue, we have $11 = 8 + 3$ structure related features.

2.3.3. Amino Acid Related Features. Besides using position dependent evolutionary and structural features, we also use position independent features such as (a) physicochemical property, (b) specific propensity of being endpoints of an SS segment, and (c) correlated contact potential, for each amino acid. Specifically, physicochemical property has 7 values for each amino acid (shown in Table 1 from [68]); specific propensity of being endpoints of an SS segment has 11 values for each amino acid (shown in Table 1 from [69]); correlated contact potential has 40 values for each amino acid (shown in Table 3 from [70]). All these features have been studied in [45] for secondary structure elements prediction and in [21–23] for homology detection. Overall, for each residue, we have $58 = 7 + 11 + 40$ amino acid dependent features.

2.4. Prediction Method

2.4.1. CNF Model. Conditional neural fields (CNF) [47] are probabilistic graphical models that have been extensively used in modeling sequential data [45, 49]. Given features on each residue on a protein sequence, we could compute the probability of each label for one residue and the transition probability for neighboring residues. Formally, for a given protein with length L, we denote its predicted labels (say, 3-state solvent accessibility or 15-state contact number) as $\mathbf{Y}(= (Y_1, \ldots, Y_L))$, where $Y_i \in \{1, 2, \ldots, M\}$, $M = 3$ for solvent accessibility prediction, and $M = 15$ for contact number prediction. We also represent the input features of a given protein by an $n \times L$ matrix $\mathbf{X}(= (F(1), \ldots, F(L)))$, where n represents the number of hidden neurons and the ith column vector $F(i)$ represents the protein feature vector associated with the ith residue, defined in the previous section. Then

we can formulize the conditional probability of the predicted labels $\mathbf{Y}$ on protein feature matrix $\mathbf{X}$ as follows:

$$P(\mathbf{Y} \mid \mathbf{X}) \propto \exp\left(\sum_{i=1}^{L-1}\psi(Y_i, Y_{i+1}) + \sum_{i=1}^{L}\sum_{j=1}^{n}\phi\left(Y_i, N_j(F(i-k), \ldots, F(i+k))\right)\right), \quad (1)$$

where $\psi(Y_i, Y_{i+1})$ is the potential function defined on an edge connecting two nodes; $\phi(Y_i, N_j(F(i-k), \ldots, F(i+k)))$ is the potential function defined at the position i; $N_j()$ is a hidden neuron function that does nonlinear transformation of input protein features; k is the window size. Formally, $\psi()$ and $\phi()$ are defined as follows:

$$\psi(Y_i, Y_{i+1}) = \sum_{a,b} t_{a,b}\delta(Y_i = a)\delta(Y_{i+1} = b),$$
$$\phi\left(Y_i, N_j\right) = \sum_{a} u_{a,j}N_j\left(w_j^T f(i)\right)\delta(Y_i = a), \quad (2)$$

where $\delta()$ is an indicator function; $f(i)$ represents the final input features $F(i-k), \ldots, F(i+k)$ for residue i; W, U, and T are model parameters to be trained. Specifically, W is the parameter from the input features to hidden neuron nodes, U from neuron to label, and T from label to label, respectively; a and b represent predicted labels (see Figure 1). The details for the training and prediction of the CNF model could be found in [45]. One beneficial result of CNF is the probability output for each label at a position through a MAP (maximum a posteriori) procedure. These probabilities, generated by CNF models trained by different combinations of feature classes, could be further utilized as features for training a consensus CNF model.

2.4.2. Multitask Learning Framework. Multitask learning (MTL) has recently attracted extensive research interest in the data mining and machine learning community [71–74]. It has been observed that learning multiple related tasks simultaneously often improves predicted accuracy [54]. Inspired by [75], a variety of functionally important protein properties, such as secondary structure and solvent accessibility, can be encoded as a labeling of amino acids and trained in multitask simultaneously under a deep neural network framework [75]. Here we propose a similar procedure for learning two tasks, say solvent accessibility and contact number, under a weight sharing CNF framework.

Specifically, assuming we have T related tasks, the "weight sharing" strategy implies that the parameters for the $N_j()$ function are shared between tasks. That is to say, the hidden neuron function that does nonlinear transformation of input protein features is shared for predicting solvent accessibility and contact number. The whole CNF framework includes the parameters $\theta_t = \{W, U_t, T_t\}$ for each task t. With this setup (i.e., only the neuron to label function U and the label to label function T are task-specific), the CNF framework automatically learns an embedding that generalizes across tasks in the first hidden neuron layers and learns features specific for the desired tasks in the second layers.

When using stochastic gradient descent to train the model parameters, we could carry out the following three steps: (a) select a task at random, (b) select a random training example for this task, and (c) compute the gradients of the CNF attributed to this task with respect to this example and update the parameters. Again, the probabilities generated by CNF models trained for different task could be utilized as features for training a consensus CNF model for a single task.

3. Results

We evaluate our program AcconPred on two prediction tasks, say solvent accessibility prediction and contact number prediction, on our own training data and CASP11 testing data. For contact number prediction, in order to compare with the existing programs, we further include the Yuan [43] dataset as the testing data. Besides using accuracy as the measurement for both solvent accessibility and contact number, we also use the following evaluation metrics for solvent accessibility, which includes precision (defined as TP/(TP+FN)), recall (defined as TP/(TP+FN)), and *F1* score (defined as 2TP/(2TP + FP + FN)), where TP, TN, FP, and FN are the numbers of the true positives, true negatives, false positives, and false negatives for a given dataset, respectively. To evaluate the performance of contact number, we also calculate the Pearson correlation between the predicted and the observed values.

In the following sections, we first give an interpretation of the 10%/40% threshold that defines the 3-state solvent accessibility. Then we evaluate the performance of AcconPred on the training data. Followed by briefly describing the programs to be compared, we show the outperformance of AcconPred with the existing programs on the testing data, which includes CASP11 and Yuan dataset.

3.1. Interpretation of the 10%/40% Threshold That Defines the 3-State Solvent Accessibility. Traditionally, predicting solvent accessibility using machine learning models is regarded as either 2, 3, or 10 labels of classification problem or a real value regression problem. There is no widely accepted criterion on how to classify the real value solvent accessibility into a finite number of discrete states such as buried, intermediate, and exposed. The reason is that, in a classification problem, with fewer labels we could get a more accurate prediction but at the same time lose lots of information by merging adjacent classes. This fact still holds between classification and regression because regression could be recognized as a kind of infinite labels prediction task with lower accuracy comparing with classification under the same situation.

Therefore, it is a tradeoff between using fewer labels of less information and using more labels less accurate. In addition, even for the same number of labels in the classification problem, the boundary for each label still needs to be finely determined. Remember that solvent accessibility represents the relative buried degree of one residue in the whole 3D

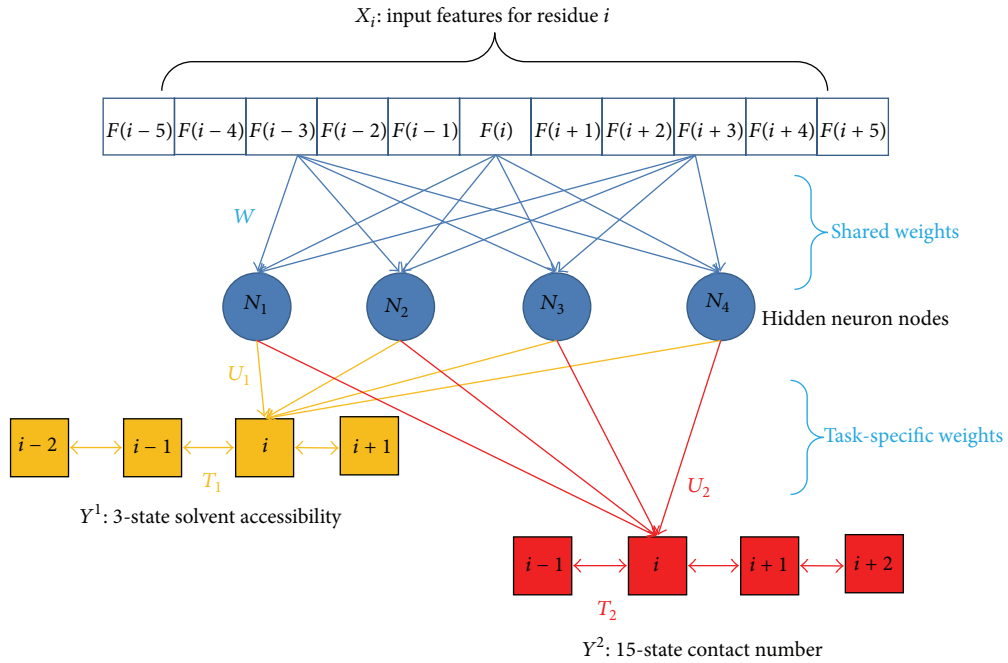


Figure 1: The shared weight multitask learning framework under the CNF (conditional neural fields) model for 3-state solvent accessibility and 15-state contact number prediction. CNF could model the relationship between input features X and label Y through a hidden layer of neuron nodes, which conduct nonlinear transformation of X. Note that the weight W from the input features to hidden neuron nodes is fixed for all tasks, while the weight U from neuron to label and the weight T from label to label are task-specific.

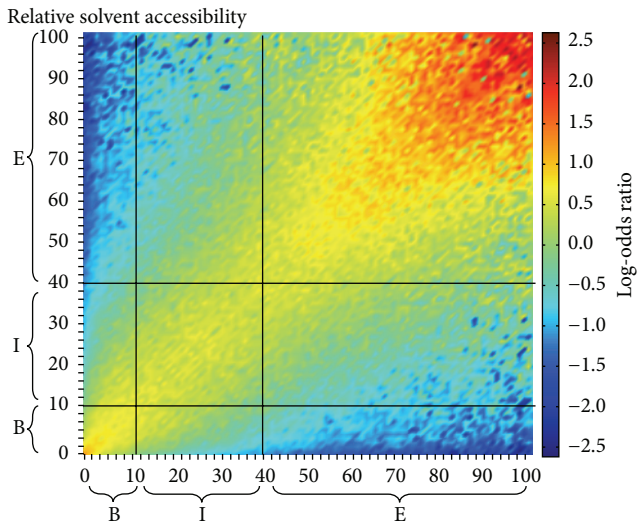


Figure 2: Log-odds ratio between the pair frequencies in the structure alignments and the background frequencies, with respect to the relative solvent accessibility in 1% unit. The thick black line indicates the boundaries at 10% and 40% to define the 3-label solvent accessibility, say buried (B), intermediate (I), and exposed (E).

protein so it is possible for two aligned residues on two structural related proteins to have different real value of accessibility in some range. To decide the range of each label is equal to giving a standard to judge if two residues with different solvent accessibilities can be aligned together.

Table 1: Precision, recall, and $F1$ score for different evaluation dataset of 3-state solvent accessibility prediction.

Evaluation dataset	Precision	Recall	$F1$ score
†Buried overall	0.76	0.78	0.77
‡Buried >0.9	0.96	0.31	0.47
Buried >0.8	0.92	0.45	0.60
Buried >0.7	0.88	0.57	0.69
Buried >0.6	0.84	0.66	0.74
Buried >0.5	0.79	0.74	0.76
Buried >0.4	0.75	0.82	0.78
Intermediate overall	0.56	0.50	0.53
Intermediate >0.9	1.00	0.0001	0.002
Intermediate >0.8	0.82	0.006	0.01
Intermediate >0.7	0.74	0.06	0.11
Intermediate >0.6	0.67	0.19	0.30
Intermediate >0.5	0.61	0.38	0.47
Intermediate >0.4	0.55	0.61	0.58
Exposed overall	0.71	0.76	0.73
Exposed >0.9	0.94	0.11	0.20
Exposed >0.8	0.88	0.31	0.46
Exposed >0.7	0.83	0.47	0.60
Exposed >0.6	0.78	0.61	0.68
Exposed >0.5	0.74	0.72	0.73
Exposed >0.4	0.69	0.81	0.75

†Overall indicates the whole set of the predicted labels.
‡ >0.9 indicates that the set of the predicted labels is chosen according to the predicted probability which is larger than 0.9.

Table 2: Prediction accuracy of different feature class and learning model for 3-state solvent accessibility.

Features	Evolution	Structure	Amino acid	†Combined single	‡Combined MTL
Q3 accuracy	0.64	0.59	0.55	0.66	**0.68**

†Combined single indicates that all classes of features, including evolution, structure, and amino acid, are used for training a single task model.
‡Combined MTL indicates that all classes of features are used for training a multitask learning model.

Table 3: Prediction accuracy of different feature class and learning models for 15-state contact number (with the same explanation as in Table 2).

Features	Evolution	Structure	Amino acid	Combined single	Combined MTL
Q15 accuracy	0.26	0.24	0.19	0.28	**0.30**

In this work, the three discrete states on relative solvent accessibility with boundaries at 10% and 40% are used (see Figure 2). We could give an interpretation for such boundaries by all-against-all protein pairwise structure alignments [76–79] on our training data. Followed by filtering out the pairs with TM-score [80] lower than 0.65 which indicates that the two proteins have no obvious biological relevance [81], we calculate the log-odds ratio between the pair frequencies in the remaining structure alignments and the background frequencies, with respect to the relative solvent accessibility in 1% unit. As shown in Figure 2, the area with more red color means that the corresponding two relative solvent accessibilities on two aligned proteins have more chance to coappear in the structure alignments, while the area with more blue color is vice versa. As a result, it can be concluded that, under such boundaries, the within-class distance is low (with more yellow or red points), while the between-class distance is high (with more cyan or blue area).

3.2. Performance on Training Data

3.2.1. Results for 3-State Solvent Accessibility Prediction

(1) Precision, Recall, and F1 Score for Each Predicted Label. Table 1 gives detailed results for each label of solvent accessibility prediction, say buried, intermediate, and exposed. Besides the overall analysis in terms of precision, recall, and *F1* score, we also provide the subset analysis of the predicted label which is chosen according to the predicted probability. From this table, we observe that when predicted probability is above 0.8, both the predicted buried label and exposed label could reach about 0.9 accuracy. However, the prediction of the intermediate label is least accurate, which can be probably expected from the arbitrariness of the threshold between the three states [20].

(2) Relative Importance of the Three Classes of Features. As mentioned in the previous section, the features used in the training process consist of three classes: evolution related, structure related, and amino acid related, respectively. In order to estimate the impact of each class on 3-state solvent accessibility prediction, we apply each of them to train the model and perform the prediction. Table 2 illustrates the prediction accuracy of different feature classes and different learning models, including single task learning model and multitask learning model. It could be observed that using amino acid related feature alone could reach 0.55 Q3 accuracy, and this accuracy could be largely increased by using the evolution related feature alone. It is interesting that although the structure related features are actually derived from the evolutionary information, the combination of all these three classes of features could reach 0.66 Q3 accuracy. Finally, we show that the performance improvement could be gained by performing multitask learning for 2% accuracy.

Table 4: Prediction accuracy of different tolerance values for 15-state contact number.

Tolerance	0	1	2	3
Accuracy	0.30	0.63	0.83	0.93

3.2.2. Results for 15-State Contact Number Prediction. Table 3 illustrates the prediction accuracy of different feature classes and different learning models for 15-state contact number, with the same trend in Table 2 in 3-state solvent accessibility perdition. It should be noted that if the difference between the predicted contact number and the observed value is only 1 or 2, we still could tolerate the result. Table 4 shows the prediction accuracy of different tolerance values, ranging from 0 to 3. If 1, 2, or 3 differences between the predicted contact number and the observed value are tolerated, the accuracy could reach 0.63, 0.83, and 0.93, respectively. The Pearson correlation score of AcconPred on the training data is 0.75.

3.3. Performance on Testing Data

3.3.1. The Existing Programs to Be Compared. We compare AcconPred with three popular solvent accessibility prediction programs, say SPINE-X [14], SANN [20], and ACCpro5 [46], as well as two contact number prediction programs, say Kinjo's method [26] and Yuan's method [43]. For solvent accessibility prediction, SPINE-X is a neural network based method, whereas SANN is based on nearest neighbor. In contrast to these two methods that rely on protein sequence information alone, ACCpro5 exploits the additional structural information derived from PDB. For contact number prediction, both Kinjo's and Yuan's methods extract features from protein sequence information. However, Kinjo's method applies linear regression for the prediction, while Yuan's method employs SVM method.

Table 5: Comparison results of the prediction accuracy of AcconPred with existing programs for 3-state solvent accessibility on the CASP11 dataset.

Method	SPINE-X	SANN	ACCpro5	AcconPred
Q3 accuracy	0.57	0.61	0.58	**0.64**

Table 6: Comparison results of the Pearson correlation score of AcconPred with existing programs for contact number prediction on the Yuan dataset.

Method	Kinjo	Yuan	AcconPred
Correlation	0.63	0.64	**0.72**

3.3.2. Results on CASP11 Data. Table 5 summarizes the results of three existing and well-known methods (say, SPINE-X, SANN, and ACCpro5, resp.) for predicting the 3-state solvent accessibility on the CASP11 105 domain cases. It should be noted that the original 3-state output of SPINE-X is based on the 25%/75% threshold, while SANN is 9%/36%. However, besides the discretized output, both SPINE-X and SANN also output predicted continuous relative solvent accessibility that ranges from 0 to 100%. So we use the same 10%/40% threshold as AcconPred to relabel the output from SPINE-X and SANN. Furthermore, the original output of ACCpro5 is 2-state which cut at 25%. Nonetheless, ACCpro5 also generates 20-state relative solvent accessibility at all thresholds between 0% and 95% at 5% increments. So in this case we could also easily transform the output of ACCpro5 into the 3-state at 10%/40% threshold. We observe that AcconPred could reach 0.65 Q3 accuracy, which is higher than SPINE-X, SANN, and ACCpro5 whose Q3 accuracies are 0.57, 0.61, and 0.58, respectively. All detailed results from SPINE-X, SANN, and ACCpro5 could be found in Supplementary Files.

We also calculate the Q15 prediction accuracy and correlation of AcconPred for 15-state contact number on CASP11 data. The results are 0.28 for Q15 and 0.71 for correlation, which is quite consistent with the results from the training data (0.3 for Q15 and 0.74 for correlation) and the Yuan data (0.28 for Q15 and 0.72 for correlation).

3.3.3. Results on Yuan Data. Since the software of both Kinjo's method and Yuan's method is not available, we perform AcconPred on the training set from Yuan. It should be noted that the Yuan data (containing 945 PDB chains) were also the training data for Kinjo's method [26]. Because the same dataset is used for contact number prediction, we could directly extract the results of Kinjo's method and Yuan's method from their paper for the comparison analysis. Table 6 summarizes the correlation results for Kinjo's method, Yuan's method, and AcconPred. We observe that our proposed method AcconPred outperforms the other methods significantly. The correlation score of AcconPred is 0.72, which is better than Kinjo's method (correlation score is 0.63) and Yuan's method (correlation score is 0.64).

4. Discussion and Future Work

In this work, we have presented AcconPred for predicting the 3-state solvent accessibility as well as the 15-state contact number for a given protein sequence. The method is based on a shared weight multitask learning framework under the CNF model. The overall performance of AcconPred for both solvent accessibility and contact number prediction is significantly better than the state-of-the-art methods.

There are two reasons why AcconPred could achieve this performance. (1) The CNF model not only captures the complex nonlinear relationship between the input protein features and the predicted labels, but also exploits interdependence among adjacent labels [45, 47]. (2) The shared weight multitask learning framework could incorporate the information of both solvent accessibility and contact number simultaneously during training [75].

Furthermore, the CNF model defines a probability distribution over the label space. The probability distribution, generated by CNF models trained on different combinations of feature classes (shown in Tables 2 and 3) for both solvent accessibility and contact number, could be further applied as the input feature to train a regression neural network model for predicting the continuous relative solvent accessibility. Meanwhile, the predicted contact number probability alone could be applied as topology constraints for the contact map prediction. It is suggested that the same framework of AcconPred could be applied to predict 10-state relative solvent accessibility, with 10% at each interval. Similar as in Table 4, we could also measure the prediction accuracy of different tolerance values for 10-state solvent accessibility.

Another uniqueness of our work is the training data, which excludes those "outlier" cases for solvent accessibility training, such as oligomer, membrane, and nonglobular proteins. This is because of the fact that these proteins have quite different solvent accessibility patterns with the monomeric soluble globular proteins. Recently, [82] pointed out that there were preferred chemical patterns of closely packed residues at the protein-protein interface. It implies that our training data that contains monomeric soluble globular proteins could serve as a control set for protein-protein interface prediction.

Conflict of Interests

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

References

[1] B. Lee and F. M. Richards, "The interpretation of protein structures: estimation of static accessibility," *Journal of Molecular Biology*, vol. 55, no. 3, pp. 379–400, 1971.

[2] W. Kauzmann, "Some factors in the interpretation of protein denaturation," in *Advances in Protein Chemistry*, vol. 14, pp. 1–63, 1959.

[3] K. A. Dill, "Dominant forces in protein folding," *Biochemistry*, vol. 29, no. 31, pp. 7133–7155, 1990.

[4] C. Chothia, "Structural invariants in protein folding," *Nature*, vol. 254, no. 5498, pp. 304–308, 1975.

[5] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus, "Hydrophobicity of amino acid residues in globular proteins," *Science*, vol. 229, no. 4716, pp. 834–838, 1985.

[6] K. A. Sharp, "Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models," *Biochemistry*, vol. 30, no. 40, pp. 9686–9697, 1991.

[7] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers—Peptide Science Section*, vol. 22, no. 12, pp. 2577–2637, 1983.

[8] L. J. McGuffin, K. Bryson, and D. T. Jones, "The PSIPRED protein structure prediction server," *Bioinformatics*, vol. 16, no. 4, pp. 404–405, 2000.

[9] D. Frishman and P. Argos, "Knowledge–based protein secondary structure assignment," *Proteins: Structure, Function, and Bioinformatics*, vol. 23, no. 4, pp. 566–579, 1995.

[10] A. G. de Brevern, C. Etchebest, and S. Hazout, "Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks," *Proteins: Structure, Function, and Bioinformatics*, vol. 41, no. 3, pp. 271–287, 2000.

[11] W.-M. Zheng and X. Liu, "A protein structural alphabet and its substitution matrix CLESUM," in *Transactions on Computational Systems Biology II*, vol. 3680 of *Lecture Notes in Comput. Sci.*, pp. 59–67, Springer, Berlin, Germany, 2005.

[12] I. Budowski-Tal, Y. Nov, and R. Kolodny, "FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 8, pp. 3481–3486, 2010.

[13] G. J. Kleywegt and T. A. Jones, "Phi/Psi-chology: ramachandran revisited," *Structure*, vol. 4, no. 12, pp. 1395–1400, 1996.

[14] E. Faraggi, B. Xue, and Y. Zhou, "Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network," *Proteins: Structure, Function and Bioinformatics*, vol. 74, no. 4, pp. 847–856, 2009.

[15] Y. Shen, F. Delaglio, G. Cornilescu, and A. Bax, "TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts," *Journal of Biomolecular NMR*, vol. 44, no. 4, pp. 213–223, 2009.

[16] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments," *Bioinformatics*, vol. 28, no. 2, Article ID btr638, pp. 184–190, 2012.

[17] M. Vendruscolo, R. Najmanovich, and E. Domany, "Protein folding in contact map space," *Physical Review Letters*, vol. 82, no. 3, pp. 656–659, 1999.

[18] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," *Journal of Molecular Biology*, vol. 233, no. 1, pp. 123–138, 1993.

[19] F. Zhao and J. Xu, "A position-specific distance-dependent statistical potential for protein structure and functional study," *Structure*, vol. 20, no. 6, pp. 1118–1126, 2012.

[20] K. Joo, S. J. Lee, and J. Lee, "Sann: solvent accessibility prediction of proteins by nearest neighbor method," *Proteins: Structure, Function and Bioinformatics*, vol. 80, no. 7, pp. 1791–1797, 2012.

[21] J. Ma, S. Wang, F. Zhao, and J. Xu, "Protein threading using context-specific alignment potential," *Bioinformatics*, vol. 29, no. 13, pp. i257–i265, 2013.

[22] J. Ma, J. Peng, S. Wang, and J. Xu, "A conditional neural fields model for protein threading," *Bioinformatics*, vol. 28, no. 12, pp. i59–i66, 2012.

[23] J. Ma, S. Wang, Z. Wang, and J. Xu, "MRFalign: protein homology detection through alignment of markov random fields," *PLoS Computational Biology*, vol. 10, no. 3, Article ID e1003500, 2014.

[24] P. Benkert, M. Künzli, and T. Schwede, "QMEAN server for protein model quality estimation," *Nucleic Acids Research*, vol. 37, no. 2, pp. W510–W514, 2009.

[25] J. Cheng, Z. Wang, A. N. Tegge, and J. Eickholt, "Prediction of global and local quality of CASP8 models by MULTICOM series," *Proteins: Structure, Function and Bioinformatics*, vol. 77, no. 9, pp. 181–184, 2009.

[26] A. R. Kinjo, K. Horimoto, and K. Nishikawa, "Predicting absolute contact numbers of native protein structure from amino acid sequence," *Proteins: Structure, Function and Genetics*, vol. 58, no. 1, pp. 158–165, 2005.

[27] A. Kabakçioğlu, I. Kanter, M. Vendruscolo, and E. Domany, "Statistical properties of contact vectors," *Physical Review E*, vol. 65, no. 4, Article ID 041904, 2002.

[28] A. N. Tegge, Z. Wang, J. Eickholt, and J. Cheng, "NNcon: improved protein contact map prediction using 2D-recursive neural networks," *Nucleic Acids Research*, vol. 37, supplement 2, pp. W515–W518, 2009.

[29] S. R. Holbrook, S. M. Muskal, and S.-H. Kim, "Predicting surface exposure of amino acids from protein sequence," *Protein Engineering*, vol. 3, no. 8, pp. 659–665, 1990.

[30] B. Rost and C. Sander, "Conservation and prediction of solvent accessibility in protein families," *Proteins: Structure, Function and Genetics*, vol. 20, no. 3, pp. 216–226, 1994.

[31] L. Ehrlich, M. Reczko, H. Bohr, and R. C. Wade, "Prediction of protein hydration sites from sequence by modular neural networks," *Protein Engineering*, vol. 11, no. 1, pp. 11–19, 1998.

[32] G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio, "Prediction of coordination number and relative solvent accessibility in proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 47, no. 2, pp. 142–153, 2002.

[33] S. Ahmad and M. M. Gromiha, "NETASA: neural network based prediction of solvent accessibility," *Bioinformatics*, vol. 18, no. 6, pp. 819–824, 2002.

[34] R. Adamczak, A. Porollo, and J. Meller, "Accurate prediction of solvent accessibility using neural networks-based regression," *Proteins: Structure, Function and Genetics*, vol. 56, no. 4, pp. 753–767, 2004.

[35] Z. Yuan, K. Burrage, and J. S. Mattick, "Prediction of protein solvent accessibility using support vector machines," *Proteins: Structure, Function and Genetics*, vol. 48, no. 3, pp. 566–570, 2002.

[36] H. Kim and H. Park, "Prediction of protein relative solvent accessibility with support vector machines and long–range interaction 3D local descriptor," *Proteins: Structure, Function and Genetics*, vol. 54, no. 3, pp. 557–562, 2004.

[37] M. N. Nguyen and J. C. Rajapakse, "Prediction of protein relative solvent accessibility with a two-stage SVM approach," *Proteins: Structure, Function and Genetics*, vol. 59, no. 1, pp. 30–37, 2005.

[38] M. J. Thompson and R. A. Goldstein, "Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes," *Proteins: Structure, Function, and Genetics*, vol. 25, no. 1, pp. 38–47, 1996.

[39] J. Sim, S.-Y. Kim, and J. Lee, "Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method," *Bioinformatics*, vol. 21, no. 12, pp. 2844–2849, 2005.

[40] S. Ahmad, M. M. Gromiha, and A. Sarai, "Real value prediction of solvent accessibility from amino acid sequence," *Proteins: Structure, Function and Genetics*, vol. 50, no. 4, pp. 629–635, 2003.

[41] A. Garg, H. Kaur, and G. P. S. Raghava, "Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure," *Proteins: Structure, Function and Genetics*, vol. 61, no. 2, pp. 318–324, 2005.

[42] Z. Yuan and B. Huang, "Prediction of protein accessible surface areas by support vector regression," *Proteins: Structure, Function, and Bioinformatics*, vol. 57, no. 3, pp. 558–564, 2004.

[43] Z. Yuan, "Better prediction of protein contact number using a support vector regression analysis of amino acid sequence," *BMC Bioinformatics*, vol. 6, article 248, 2005.

[44] N. Goldman, J. L. Thorne, and D. T. Jones, "Assessing the impact of secondary structure and solvent accessibility on protein evolution," *Genetics*, vol. 149, no. 1, pp. 445–458, 1998.

[45] Z. Wang, F. Zhao, J. Peng, and J. Xu, "Protein 8-class secondary structure prediction using conditional neural fields," *Proteomics*, vol. 11, no. 19, pp. 3786–3792, 2011.

[46] C. N. Magnan and P. Baldi, "SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity," *Bioinformatics*, vol. 30, no. 18, pp. 2592–2597, 2014.

[47] J. Peng, L. Bo, and J. Xu, "Conditional neural fields," in *Advances in Neural Information Processing Systems*, 2009.

[48] S. Wang, J. Peng, and J. Xu, "Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling," *Bioinformatics*, vol. 27, no. 18, pp. 2537–2545, 2011.

[49] F. Zhao, J. Peng, and J. Xu, "Fragment-free approach to protein folding using conditional neural fields," *Bioinformatics*, vol. 26, no. 12, Article ID btq193, pp. i310–i317, 2010.

[50] M. Källberg, H. Wang, S. Wang et al., "Template-based protein structure modeling using the RaptorX web server," *Nature Protocols*, vol. 7, no. 8, pp. 1511–1522, 2012.

[51] M. Källberg, G. Margaryan, S. Wang, J. Ma, and J. Xu, "RaptorX server: a resource for template-based protein structure modeling," in *Protein Structure Prediction*, vol. 1137 of *Methods in Molecular Biology*, pp. 17–27, Springer, 2014.

[52] I. Dubchak, S. Balasubramanian, S. Wang et al., "An integrative computational approach for prioritization of genomic variants," *PLoS ONE*, vol. 9, no. 12, Article ID e114903, 2014.

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

[54] R. Collobert and J. Weston, "A unified architecture for natural language processing: deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*, pp. 160–167, ACM, July 2008.

[55] R. Caruana, *Multitask Learning*, Springer, Berlin, Germany, 1998.

[56] C. Chothia, "The nature of the accessible and buried surfaces in proteins," *Journal of Molecular Biology*, vol. 105, no. 1, pp. 1–12, 1976.

[57] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[58] D. Kozma, I. Simon, and G. E. Tusnády, "PDBTM: protein data bank of transmembrane proteins after 8 years," *Nucleic Acids Research*, vol. 41, no. 1, pp. D524–D529, 2013.

[59] R. A. Jordan, Y. El-Manzalawy, D. Dobbs, and V. Honavar, "Predicting protein-protein interface residues using local surface structural similarity," *BMC Bioinformatics*, vol. 13, article 41, 2012.

[60] R. Sowdhamini, S. D. Rufino, and T. L. Blundell, "A database of globular protein structural domains: clustering of representative family members into similar folds," *Folding and Design*, vol. 1, no. 3, pp. 209–220, 1996.

[61] J. Moult, "A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction," *Current Opinion in Structural Biology*, vol. 15, no. 3, pp. 285–289, 2005.

[62] R. Adamczak, A. Porollo, and J. Meller, "Combining prediction of secondary structure and solvent accessibility in proteins," *Proteins: Structure, Function and Genetics*, vol. 59, no. 3, pp. 467–475, 2005.

[63] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi, "SCRATCH: a protein structure and structural feature prediction server," *Nucleic Acids Research*, vol. 33, supplement 2, pp. W72–W76, 2005.

[64] Y. Y. Tseng and J. Liang, "Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach," *Molecular Biology and Evolution*, vol. 23, no. 2, pp. 421–436, 2006.

[65] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[66] J. Söding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics*, vol. 21, no. 7, pp. 951–960, 2005.

[67] A. Biegert and J. Söding, "Sequence context-specific profiles for homology searching," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 10, pp. 3770–3775, 2009.

[68] J. Meiler, M. Müller, A. Zeidler, and F. Schmäschke, "Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks," *Journal of Molecular Modeling*, vol. 7, no. 9, pp. 360–369, 2001.

[69] M. Duan, M. Huang, C. Ma, L. Li, and Y. Zhou, "Position-specific residue preference features around the ends of helices and strands and a novel strategy for the prediction of secondary structures," *Protein Science*, vol. 17, no. 9, pp. 1505–1512, 2008.

[70] Y. H. Tan, H. Huang, and D. Kihara, "Statistical potential-based amino acid similarity matrices for aligning distantly related protein sequences," *Proteins: Structure, Function and Genetics*, vol. 64, no. 3, pp. 587–600, 2006.

[71] H. Fei and J. Huan, "Structured feature selection and task relationship inference for multi-task learning," *Knowledge and Information Systems*, vol. 35, no. 2, pp. 345–364, 2013.

[72] O. Chapelle, P. Shivaswamy, S. Vadrevu, K. Weinberger, Z. Ya, and B. Tseng, "Multi-task learning for boosting with application to web search ranking," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*, pp. 1189–1197, ACM, July 2010.

[73] J. Chen, J. Liu, and J. Ye, "Learning incoherent sparse and low-rank patterns from multiple tasks," *ACM Transactions on Knowledge Discovery from Data*, vol. 5, no. 4, article 22, 2012.

[74] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient $l_{2,1}$-norm minimization," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pp. 339–348, AUAI Press, 2009.

[75] Y. Qi, M. Oja, J. Weston, and W. S. Noble, "A unified multitask architecture for predicting local protein properties," *PLoS ONE*, vol. 7, no. 3, Article ID e32235, 2012.

[76] S. Wang, J. Ma, J. Peng, and J. Xu, "Protein structure alignment beyond spatial proximity," *Scientific Reports*, vol. 3, article 1448, 2013.

[77] S. Wang and W.-M. Zheng, "CLePAPS: fast pair alignment of protein structures based on conformational letters," *Journal of Bioinformatics and Computational Biology*, vol. 6, no. 2, pp. 347–366, 2008.

[78] S. Wang and W.-M. Zheng, "Fast multiple alignment of protein structures using conformational letter blocks," *The Open Bioinformatics Journal*, vol. 3, pp. 69–83, 2009.

[79] J. Ma and S. Wang, "Algorithms, applications, and challenges of protein structure alignment," *Advances in Protein Chemistry and Structural Biology*, vol. 94, pp. 121–175, 2014.

[80] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins: Structure, Function and Genetics*, vol. 57, no. 4, pp. 702–710, 2004.

[81] J. Xu and Y. Zhang, "How significant is a protein structure similarity with TM-score = 0.5?" *Bioinformatics*, vol. 26, no. 7, pp. 889–895, 2010.

[82] Q. Luo, R. Hamer, G. Reinert, and C. M. Deane, "Local network patterns in protein-protein interfaces," *PLoS ONE*, vol. 8, no. 3, Article ID e57031, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 751646, 10 pages
http://dx.doi.org/10.1155/2015/751646

Research Article

Improving Classification of Protein Interaction Articles Using Context Similarity-Based Feature Selection

Yifei Chen, Yuxing Sun, and Bing-Qing Han

School of Technology, Nanjing Audit University, 86 W. Yushan Road, Nanjing 211815, China

Correspondence should be addressed to Yifei Chen; yifeichen91@nau.edu.cn

Received 20 October 2014; Revised 13 December 2014; Accepted 14 December 2014

Academic Editor: Fang-Xiang Wu

Protein interaction article classification is a text classification task in the biological domain to determine which articles describe protein-protein interactions. Since the feature space in text classification is high-dimensional, feature selection is widely used for reducing the dimensionality of features to speed up computation without sacrificing classification performance. Many existing feature selection methods are based on the statistical measure of document frequency and term frequency. One potential drawback of these methods is that they treat features separately. Hence, first we design a similarity measure between the context information to take word cooccurrences and phrase chunks around the features into account. Then we introduce the similarity of context information to the importance measure of the features to substitute the document and term frequency. Hence we propose new context similarity-based feature selection methods. Their performance is evaluated on two protein interaction article collections and compared against the frequency-based methods. The experimental results reveal that the context similarity-based methods perform better in terms of the $F1$ measure and the dimension reduction rate. Benefiting from the context information surrounding the features, the proposed methods can select distinctive features effectively for protein interaction article classification.

1. Introduction

An overwhelming number of biological articles are published daily online as a result of growing interest in biological research, especially relating to the study of protein-protein interactions (PPIs). It is essential to classify which articles describe PPIs, that is, to filter out those irrelevant articles from the whole collection of the biological literature. This allows a more efficient extraction of PPIs from the large amount of biological literature. Automated text classification is a key technology to rapidly find relevant articles. Text classification has been successfully applied to various domains such as text sentinel classification [1], spam e-mail filtering [2, 3], author identification [4], and web page classification [5]. Research on protein interaction article classification (IAC) is a text classification task with practical significance in the biological domain.

In the classic text classification framework, a feature extraction mechanism extracts features from raw articles, including all distinct terms (words). This is also known as bag-of-words (BOW) representation for text documents. Hence each article is represented by a multidimensional feature vector where each dimension corresponds to a term (feature) within the literature collection. Even a small literature collection would contain tens of thousands of features [6, 7]. The high dimensionality of the feature space not only increases computational time but also degrades classification performance. Hence, automated feature selection plays an essential role in making the text classification more efficient and accurate by selecting a subset of the most important features [8, 9]. Feature selection is an active research area in many fields such as data mining, machine learning, and rough sets [10–13].

The process of feature selection typically involves certain metrics that are designed for measuring the importance level of features, and the most important features are selected to help in efficient utilization of resources for large scale problems [14]. The existing feature selection methods are mostly based on the statistical information in documents, including term frequency and document frequency [7, 14–18]. Term frequency is the number of times a particular term appears in a document while document frequency

is the number of documents containing that term within the literature collection. One potential drawback of most of these frequency-based feature selection methods is that they treat each feature separately [19]. In other words, these approaches are context independent: they do not utilize the context information around the terms when judging their importance, such as word order, word cooccurrence, multiword chunks, and semantic relationships. However, this information is important for classifying which articles are PPI relevant or nonrelevant. For example, protein names exist in both PPI relevant and nonrelevant documents. So they could have great document frequency or term frequency. However, obviously they are not distinctive terms for the purpose of classification. Hence, it is difficult to measure the importance of all the terms just according to the document frequency or term frequency. After in-depth research we have noticed that, in the PPI relevant documents, the fact that proteins interact with each other is described through the context of those proteins. Meanwhile in nonrelevant documents, the fact that there are no interactions between the particular proteins is also depicted within the context of the documents. The above observation leads us to an interesting issue which is that the context of features in biological articles can be utilized to measure feature importance and to improve the feature selection process. Hence we propose context similarity-based feature selection methods.

This paper is organized as follows: we provide an overview of the existing frequency-based feature selection methods for text classification in Section 2, and this is followed by a definition of the proposed context similarity-based feature selection methods. Then in order to examine the two kinds of methods carefully, the experimental results and discussion are presented in Section 3 to find which one is more useful in the IAC task. This is followed by a conclusion in Section 4.

2. Materials and Methods

2.1. Existing Feature Selection Methods for Text Classification. Feature selection is a process which selects a subset of the most important features. Such selection can help in building effective and efficient models for text classification. Normally, feature selection techniques can be divided into three categories: filters, wrappers, and embedded methods [19]. Filters measure feature importance using various scoring metrics that are independent of a learning model or classifier and select top-N features attaining the highest scores. Univariate filter techniques are computationally fast. However, they do not take feature dependencies into consideration, which was discussed as the motivation of this paper in Section 1. In addition, multivariate filter techniques incorporate feature dependencies to some degree, while they are slower and less scalable than univariate techniques. Wrappers evaluate features using a certain search algorithm together with a specific learning model or classifier. Wrapper techniques consider feature dependencies and provide interaction between features during the subset search processing but are computationally expensive compared with filters. Embedded methods integrate feature selection into the model learning phase. Therefore, they merge with the model or classifier much further than the wrappers. Nevertheless, they are also computationally more intensive than filters.

Considering the high dimensionality of the feature space for text classification tasks, the most frequently used approach for feature selection is the univariate filter method [7]. And among them four document frequency-based methods and two term frequency-based methods that will be discussed in the paper are illustrated as follows, where $P(t_k \mid c_i)$ is the percentage of documents belonging to a category c_i in which the term t_k occurs and $P(t_k \mid \bar{c}_i)$ is the percentage of documents not belonging to a category c_i in which the term t_k occurs. $|c|$ is the number of categories, which is 2 for the IAC task.

(1) Document Frequency (DF). Document frequency (DF) is a simple and effective feature selection method which is based on the assumption that infrequent terms are not reliable in text classification and may degrade the performance [7]. Hence, if the document frequency in which a term occurs is the largest, the term is retained [20]. The DF metrics of the term t_k can be computed as follows:

$$\mathrm{DF}(t_k) = \sum_{i=1}^{|c|} \mathrm{DF}(t_k, c_i) = \sum_{i=1}^{|c|} P(t_k \mid c_i), \tag{1}$$

where $\mathrm{DF}(t_k, c_i)$ is the DF measure of the term t_k in a category c_i and $\mathrm{DF}(t_k)$ is the sum of $\mathrm{DF}(t_k, c_i)$ across all the categories.

(2) Gini Index (GI). Gini Index (GI) was originally used to find the best attributes in decision trees. Shang et al. [15] proposed an improved version of the GI method to apply it directly to text feature selection. The $\mathrm{GI}(t_k, c_i)$ measures the purity of the feature t_k towards a category c_i. Its sum across categories, $\mathrm{GI}(t_k)$, is given as

$$\mathrm{GI}(t_k) = \sum_{i=1}^{|c|} \mathrm{GI}(t_k, c_i) = \sum_{i=1}^{|c|} P(t_k \mid c_i)^2 P(c_i \mid t_k)^2, \tag{2}$$

where $P(c_i \mid t_k)$ is the conditional probability of the feature t_k belonging to a category c_i given presence of the feature t_k.

(3) Class Discriminating Measure (CDM). Class discriminating measure (CDM) is a derivation of the odds ration introduced by Chen et al. [16]. The results in their paper indicate that CDM is a better feature selection approach than information gain (IG). The CDM calculates the effectiveness of the term t_k as follows:

$$\mathrm{CDM}(t_k) = \sum_{i=1}^{|c|} \mathrm{CDM}(t_k, c_i) = \sum_{i=1}^{|c|} \left| \log \frac{P(t_k \mid c_i)}{P(t_k \mid \bar{c}_i)} \right|, \tag{3}$$

where $\mathrm{CDM}(t_k, c_i)$ is the CDM measure of the term t_k in a category c_i and $\mathrm{CDM}(t_k)$ is the sum of $\mathrm{CDM}(t_k, c_i)$ across all the categories.

(4) Accuracy Balanced (Acc2). Accuracy balanced (Acc2) is a two-side metric (it selects both negative and positive features) that is based on the difference of the distributions of a term belonging to a category and not belonging to that category

in the documents. In Forman [14], the Acc2 is studied and claimed to have a performance comparable to the IG and chi-square statistical metrics. The Acc2 of the term t_k can be computed as follows:

$$\mathrm{Acc2}(t_k) = \sum_{i=1}^{|c|} \mathrm{Acc2}(t_k, c_i) = \sum_{i=1}^{|c|} \left| P(t_k \mid c_i) - P(t_k \mid \bar{c}_i) \right|, \tag{4}$$

where $\mathrm{Acc2}(t_k, c_i)$ is the Acc2 measure of the term t_k in a category c_i and $\mathrm{Acc2}(t_k)$ is the sum of $\mathrm{Acc2}(t_k, c_i)$ across all the categories.

(5) Term Frequency Inverse Document Frequency (TFIDF). Term frequency inverse document frequency (TFIDF) is a numerical statistic that is intended to reflect how important a term is to a document in a collection or corpus. One of the simplest filter metrics is computed by summing the TFIDF. Wei et al. [21] introduced category information to TFIDF, which can be reformed using a notation of term frequency $\mathrm{tf}(t_k, c_i)$ that is the number of occurrences of a term t_k in documents from a category c_i. Consider

$$\mathrm{TFIDF}(t_k) = \sum_{i=1}^{|c|} \mathrm{tf}(t_k, c_i) \times \log\left(\frac{1}{P(t_k \mid c_i)} \right). \tag{5}$$

(6) Normalized Term Frequency-Based Gini Index ($GINI_{NTF}$). Normalized term frequency-based Gini Index ($\mathrm{GINI_{NTF}}$) revised the document frequency in the Gini Index metric with the term frequency by Azam and Yao [17]. Experimental results revealed that the term frequency-based metric was useful in feature selection. We reform the formula of $\mathrm{GINI_{NTF}}$ as follows:

$$\mathrm{GINI_{NTF}}(t_k) = \sum_{i=1}^{|c|} \left(\frac{\mathrm{tf_{norm}}(t_k, c_i)}{\mathrm{doc}(c_i)} \right)^2 \times \left(\frac{\mathrm{tf_{norm}}(t_k, c_i)}{\mathrm{tf_{norm}}(t_k, c_i) + \mathrm{tf_{norm}}(t_k, \bar{c}_i)} \right)^2, \tag{6}$$

where $\mathrm{tf_{norm}}(t_k, c_i)$ is the normalized term frequency of t_k in documents from a category c_i and $\mathrm{tf_{norm}}(t_k, \bar{c}_i)$ is the normalized term frequency of t_k in documents not from a category c_i. The normalized values of term frequency are used in the metric so that term frequencies are not influenced by varying lengths of documents.

2.2. Context Similarity-Based Feature Selection Methods. According to the bag-of-words document representation, each raw document in the article collection is transformed into a high-dimensional vector before the process of text classification. In order to address the issues of high dimensionality, the feature filter methods, such as the DF, GI, CDM, and Acc2, are utilized to select the most important features based on document frequency. One potential problem of these frequency-based methods is that they ignore the context relationships between features. As we have discussed in Section 1, context information is essential for the IAC task. When attempting to judge the importance levels of features, it may be advantageous to explicitly compare the similarity shared among contexts in PPI relevant articles or nonrelevant articles. Hence when building the feature selection metrics, we take the significance of context information of each feature into account through the context similarity.

Context Similarity Measure. $\mathrm{sim_{context}}(t_k, c_i)$ is designed to explicitly express the similarity shared by contexts of the term t_k in a certain category c_i. The measure is based on the word cooccurrences and chunks of a pair of context strings $\mathrm{context}_d(t_k, w)$ and $\mathrm{context}_{d'}(t_k, w)$ containing the term t_k within a category c_i. $\mathrm{context}_d(t_k, w)$ denotes a document d containing a term t_k within a context string $\{t_{-wk}, \ldots, t_{-1k}, t_k, t_{1k}, \ldots, t_{wk}\}$, where w is a window size that takes into account w terms before and after the term t_k. The term t_k is contained in another context string of document d', $\mathrm{context}_{d'}(t_k, w)$, which is $\{t'_{-wk}, \ldots, t'_{-1k}, t_k, t'_{1k}, \ldots, t'_{wk}\}$ with the window size w. Using $\mathrm{context}_d$, a multiword phrase chunk containing t_k and its word cooccurrence can be considered to measure the importance of t_k.

First $\mathrm{sim}(\mathrm{context}_d(t_k, w), \mathrm{context}_{d'}(t_k, w))$ is defined to measure the similarity between the context string pair $(\mathrm{context}_d, \mathrm{context}_{d'})$ as follows:

$$\mathrm{sim}\left(\mathrm{context}_d(t_k, w), \mathrm{context}_{d'}(t_k, w)\right) = \frac{\sum_{w=0}^{|w|} \mathrm{dis}\left(\mathrm{context}_d(t_k, w), \mathrm{context}_{d'}(t_k, w)\right)}{|w| + 1}. \tag{7}$$

The sum of all the context strings from 0 to maximum window size $|w|$ is utilized to incorporate word cooccurrence and phrase similarity comprehensively. $|w|$ is used to control the scope of the local information of term t_k involved in the measurement, and trials on the training data show that $|w| = 3$ is the optimal value. In this paper, Jaro-Winkler [22] distance is employed as the distance function of two context strings, $\mathrm{dis}(\mathrm{context}_d(t_k, w), \mathrm{context}_{d'}(t_k, w))$, because it was designed and best suited for short strings. The Jaro-Winkler distance is a measure of similarity between two strings, and it is a variant of the Jaro distance metric [23, 24]. The higher the Jaro-Winkler distance for two strings is, the more similar the strings are. The score is normalized such that 0 equates to no similarity and 1 is an exact match.

Then, $\mathrm{sim_{context}}(t_k, c_i)$ is defined to measure the similarity of context in the documents containing the term t_k belonging to a category c_i as follows:

$$\mathrm{sim_{context}}(t_k, c_i) = \sum_{\mathrm{context}_d, \mathrm{context}_{d'} \in c_k} \mathrm{sim}\left(\mathrm{context}_d(t_k, w), \mathrm{context}_{d'}(t_k, w)\right). \tag{8}$$

Context Similarity-Based Feature Selection Methods. In order to elaborate the context similarity-based feature selection metrics, the class discriminating measure (CDM) is considered as an example, which was very useful in reducing the feature set in some application domains. The metric of CDM has been defined in Section 2.1 based on $P(t_k \mid c_i)$ and $P(t_k \mid \bar{c}_i)$. Here $P(t_k \mid c_i)$, the percentage of documents with the term t_k belonging to the category c_i, can also be represented as $\mathrm{doc}(t_k, c_i)/\mathrm{doc}(c_i)$, where $\mathrm{doc}(t_k, c_i)$ is the document

frequency containing the term t_k in the category c_i and $\mathrm{doc}(c_i)$ is the total number of articles in the category c_i. $P(t_k \mid \bar{c}_i)$, the percentage of documents with the term t_k not belonging to the category c_i, can be represented as $\mathrm{doc}(t_k, \bar{c}_i)/\mathrm{doc}(\bar{c}_i)$, where $\mathrm{doc}(t_k, \bar{c}_i)$ is the document frequency containing the term t_k not in the category c_i and $\mathrm{doc}(\bar{c}_i)$ is the total number of articles not in the category c_i. Hence, we can have the following CDM metric:

$$\begin{aligned}\mathrm{CDM}(t_k) &= \sum_{i=1}^{|c|} \left| \log \frac{P(t_k \mid c_i)}{P(t_k \mid \bar{c}_i)} \right| \\ &= \sum_{i=1}^{|c|} \left| \log \left(\frac{\mathrm{doc}(t_k, c_i)}{\mathrm{doc}(c_i)} \cdot \frac{\mathrm{doc}(\bar{c}_i)}{\mathrm{doc}(t_k, \bar{c}_i)} \right) \right|.\end{aligned} \tag{9}$$

In order to make use of the context information of terms and not just the document frequency, we substitute the context similarity measure $\mathrm{sim}_{\mathrm{context}}(t_k, c_i)$ for the document frequency $\mathrm{doc}(t_k, c_i)$. Then the obtained metric with reformed definition is referred to as $\mathrm{CDM}_{\mathrm{cs}}$, class discriminating measure based on context similarity. If the context similarity of a term within a certain text category is greater, the term is more important for text classification. The definition of $\mathrm{CDM}_{\mathrm{cs}}$ is as follows:

$$\mathrm{CDM}_{\mathrm{cs}}(t_k) = \sum_{i=1}^{|c|} \left| \log \left(\frac{\mathrm{sim}_{\mathrm{context}}(t_k, c_i)}{\mathrm{doc}(c_i)} \cdot \frac{\mathrm{doc}(\bar{c}_i)}{\mathrm{sim}_{\mathrm{context}}(t_k, \bar{c}_i)} \right) \right|. \tag{10}$$

The other three document frequency-based metrics defined in Section 2.1 can also be reformed in the same way based on the context similarity to $\mathrm{Acc2}_{\mathrm{cs}}$, $\mathrm{GI}_{\mathrm{cs}}$, and $\mathrm{DF}_{\mathrm{cs}}$:

$$\begin{aligned}\mathrm{Acc2}_{\mathrm{cs}}(t_k) &= \sum_{i=1}^{|c|} \left| \frac{\mathrm{sim}_{\mathrm{context}}(t_k, c_i)}{\mathrm{doc}(c_i)} - \frac{\mathrm{sim}_{\mathrm{context}}(t_k, \bar{c}_i)}{\mathrm{doc}(\bar{c}_i)} \right|, \\ \mathrm{GI}_{\mathrm{cs}}(t_k) &= \sum_{i=1}^{|c|} \left| \left(\frac{\mathrm{sim}_{\mathrm{context}}(t_k, c_i)}{\mathrm{doc}(c_i)} \right)^2 \left(\frac{\mathrm{sim}_{\mathrm{context}}(t_k, c_i)}{\mathrm{doc}(t_k)} \right)^2 \right|, \\ \mathrm{DF}_{\mathrm{cs}}(t_k) &= \sum_{i=1}^{|c|} \frac{\mathrm{sim}_{\mathrm{context}}(t_k, c_i)}{\mathrm{doc}(c_i)},\end{aligned} \tag{11}$$

where $\mathrm{doc}(t_k)$ is the number of documents containing the term t_k in all the text categories.

3. Results and Discussion

3.1. Experimental Settings

Classification Model $\mathrm{Model}_{\mathrm{SVM\_poly}}$. Support vector machines (SVMs) pioneered by Vapnik [25] are suitable for complex classification problems. Their power comes from the combination of the kernel trick and maximum margin hyperplane separation. SVMs are one of the most successful approaches for classification in text mining [26, 27]. Hence, in this paper, we employ the SVMs with a polynomial kernel as a classification model, $\mathrm{Model}_{\mathrm{SVM\_poly}}$, which is trained and tested using the LIBSVM toolbox [28]. A 10-fold cross-validation is adopted to tune parameters.

Data Sets. An in-depth investigation will be carried out to compare the performances of the four proposed context similarity-based methods and the six existing frequency-based feature selection methods. Two data sets ($\mathrm{Data}_{\mathrm{BCII}}$ and $\mathrm{Data}_{\mathrm{BCIII}}$) are used in our experiments to evaluate the performance, which are both extracted from the BioCreAtIvE (the Critical Assessment of Information Extraction in Biology) challenges. The challenges were set up to evaluate the state of the art of text mining and information extraction in the biological domain.

In the data preprocessing step, all words are converted to lower case, punctuation marks and stop words are removed, and no stemming is used. Consider the following.

(1) $\mathrm{Data}_{\mathrm{BCII}}$: we obtain the $\mathrm{Data}_{\mathrm{BCII}}$ from the Protein Interaction Article Subtask (IAS) of the BioCreAtIvE II challenge [29]. The $\mathrm{Data}_{\mathrm{BCII}}$ is composed of abstracts of 6,172 articles in total, which are taken from a set of MEDLINE articles that are annotated as interaction articles or not according to the guidelines used by the MINT and IntAct databases. There are 5,495 abstracts used as training data and 677 ones as test data. And there are 3,536 and 338 interaction articles, that is, positive examples, in the training and test set, respectively.

(2) $\mathrm{Data}_{\mathrm{BCIII}}$: we obtain the $\mathrm{Data}_{\mathrm{BCIII}}$ from the PPI Article Classification Task (ACT) of the BioCreAtIvE III challenge [30]. The training set (TR) consists of a balanced collection of 2,280 articles classified through manual inspection, divided into PPI relevant and nonrelevant articles. The annotation guidelines for this task were refined iteratively based on the feedback from both annotation databases and specially trained domain experts. The development (DE) and test (TE) set take into account PPI relevant journals based on the current content of collaborating PPI databases. Random samples of abstracts from these journals were taken to generate a development set of 4,000 abstracts (628 PPI relevant and 3,318 nonrelevant abstracts) in total and a test set of 6,000 abstracts (918 PPI relevant and 5,090 nonrelevant abstracts). These two disjointed sets were drawn from the same sample collection.

Performance Measures. Since the applications are restricted to IAC, which is a binary classification task, we measure the performance in terms of $F1$ measure [20]. The $F1$ is determined by a combination of precision and recall. Precision is the percentage of documents that are correctly classified as being positive. Recall is the percentage of positive documents that are correctly classified. The precision, recall, and $F1$ are obtained as

$$\begin{aligned}\mathrm{Precision} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \\ \mathrm{Recall} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\ F1 &= \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}},\end{aligned} \tag{12}$$

where TP is the number of positive documents that are correctly classified as positive ones, FP is the number of negative documents that are misclassified as positive ones, TN is the number of negative documents that are correctly classified as negative ones, and FN is the number of positive documents that are misclassified as negative ones.

3.2. Experimental Results on the $Data_{BCII}$. First, we test all the feature selection methods when $Model_{SVM\_poly}$ is applied on the $Data_{BCII}$ data set, where there are 29,979 total features extracted using the bag-of-words document representation. The proposed context similarity-based methods, GI_{cs}, DF_{cs}, CDM_{cs}, and $Acc2_{cs}$, are compared with the frequency-based methods, GI, DF, CDM, Acc2, TFIDF, and $GINI_{NTF}$, when the number of the selected features is the top 0.5%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%. Figure 1 shows the trend curves of all the feature selection methods, and the optimal parameter value of the window size of context information is 3, which is tuned through 10-fold cross-validation.

Figure 1 indicates that all these feature selection methods have a similar trend on the $Data_{BCII}$, and the proposed methods are more effective. The context similarity-based methods and the term frequency-based methods achieve the best performance when around 4% top important features are selected, while the document frequency-based methods obtain the best performance when around 7-8% features are used. Moreover, the proposed methods outperform the other methods on selecting the top important features to achieve the best $F1$ measure. Among the context similarity-based feature selection methods, when the top 1300 features (4.3% of total number of features) are selected, GI_{cs} acquires the highest $F1$ measure 77.07, which effectively improves the $F1$ measure of the $Model_{SVM\_poly}$ when all the features are used (73.55) by 3.52.

Further, in order to study the performance of all these feature selection methods in more detail, a small feature set in the scope of the top 2000 is used. The corresponding $F1$ measure results are shown in Table 1 when the top 100, 300, 500, 700, 900, 1100, 1300, 1500, 1700, and 1900 features are selected. The best result for each feature set is shown in bold. It can be seen from Table 1 that the context similarity-based methods outperform those methods based on the document frequency or term frequency. The last column of Table 1 presents the best performance of the $Model_{SVM\_poly}$ that various feature selection methods can achieve, and the size of selected features when the best performance is achieved is illustrated in the parentheses. It can be seen that, compared with the four document frequency-based methods, the TFIDF and the $GINI_{NTF}$ perform better, which shows that term frequency is a relatively more important factor than document frequency. Moreover, all the context similarity-based methods achieve better performance with fewer selected features, and among them the GI_{cs} performs the best on the $Data_{BCII}$. Hence, the proposed method can extract more effective information from context similarity measure of term cooccurrences and chunks than just calculating the document frequency or term frequency. This context information is helpful when measuring the importance of features to boost the performance.

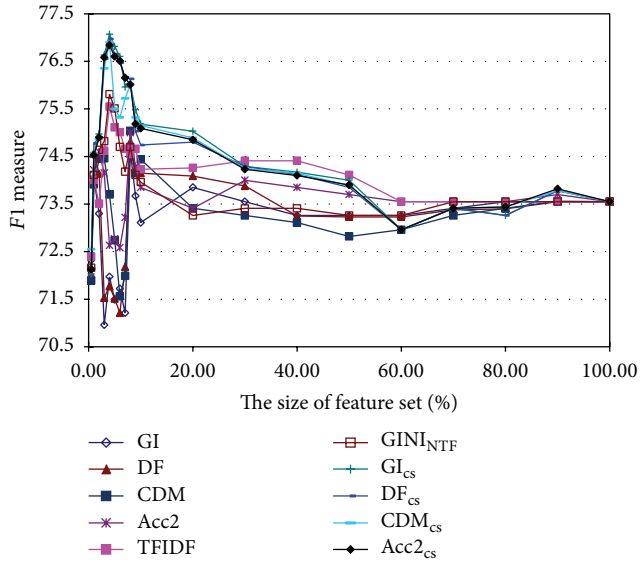


Figure 1: The $F1$ performance curves of all the feature selection methods on the $Data_{BCII}$ when the number of the selected features is the top 0.5%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%.

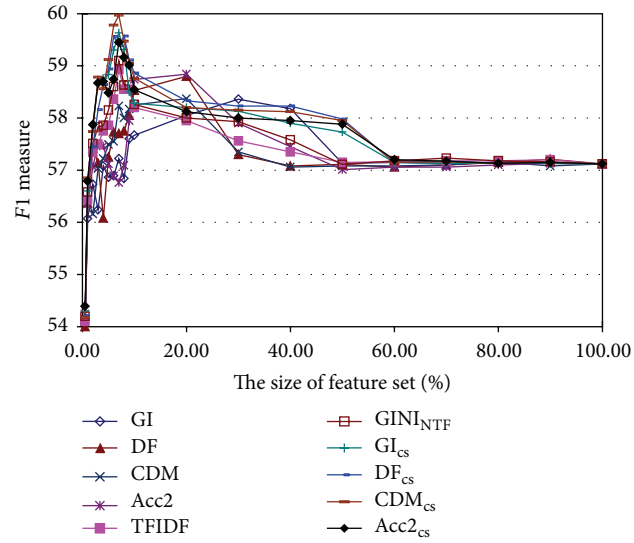


Figure 2: The $F1$ performance curves of all the feature selection methods on the $Data_{BCIII}$ when the number of selected features is the top 0.5%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%.

3.3. Experimental Results on the $Data_{BCIII}$. Then, we test the proposed feature selection methods on the $Data_{BCIII}$ when the number of selected features is the top 0.5%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%, where there are 23,084 features extracted using the bag-of-words representation in total. Figure 2 shows the trend curves of the $F1$ measure versus different sizes of selected features. From Figure 2 we can see that when around 7% top important features are used, the

Table 1: The $F1$ measure results when the $\text{Model}_{\text{SVM\_poly}}$ is applied to the $\text{Data}_{\text{BCII}}$ when the top 100, 300, 500, 700, 900, 1100, 1300, 1500, 1700, and 1900 features are selected. In each column, the bold value indicates the best performance for each feature set when various feature selection methods are used, respectively. The "best" column presents the best performance that various feature selection methods can achieve, and the numbers in the parentheses are the corresponding sizes of feature sets.

Number of features	100	300	500	700	900	1100	1300	1500	1700	1900	best
GI	72.32	74.04	73.30	73.34	70.96	71.79	71.97	71.49	71.27	71.21	74.60(2300)
DF	72.02	74.56	74.14	72.50	71.53	72.21	71.77	71.51	71.21	72.18	74.80(2300)
CDM	71.89	73.91	74.45	74.49	74.46	74.25	73.70	72.73	71.56	71.99	75.04(2100)
Acc2	72.40	74.22	74.60	73.49	74.15	74.01	72.63	72.77	72.58	73.22	75.00(2100)
TFIDF	72.40	74.01	73.51	73.61	74.62	75.11	75.55	75.11	75.01	74.66	75.55(1300)
GINI_{NTF}	72.16	74.11	74.64	74.70	74.82	75.41	75.81	75.51	74.70	74.18	75.81(1300)
GI_{cs}	72.03	74.52	**74.97**	**76.14**	**76.63**	**76.66**	**77.07**	**76.81**	**76.60**	75.96	**77.07**(1300)
DF_{cs}	72.15	**74.77**	74.90	76.09	76.55	76.60	76.97	76.65	76.47	**76.16**	76.97(1300)
CDM_{cs}	**72.55**	74.57	74.87	75.99	76.35	76.47	76.90	75.50	75.38	75.81	76.90(1300)
Acc2_{cs}	72.13	74.53	74.90	76.06	76.58	76.61	76.84	76.60	76.51	76.15	76.84(1300)

Table 2: The $F1$ measure results when the $\text{Model}_{\text{SVM\_poly}}$ is used on the $\text{Data}_{\text{BCIII}}$ when the top 100, 300, 500, 700, 900, 1100, 1300, 1500, 1700, and 1900 features are selected. In each column, the bold value indicates the best performance for each feature set when various feature selection methods are used, respectively. The "best" column presents the best performance that various feature selection methods can achieve, and the numbers in the parentheses are the corresponding sizes of feature sets.

Number of features	100	300	500	700	900	1100	1300	1500	1700	1900	best
GI	52.16	56.07	56.73	56.24	57.08	56.86	56.89	57.34	57.22	56.84	58.36(5900)
DF	50.91	56.37	57.42	57.14	56.98	57.25	57.73	57.09	57.70	57.75	58.80(4300)
CDM	52.13	56.43	56.16	57.03	57.22	57.49	57.55	58.27	58.23	58.00	58.37(4500)
Acc2	52.24	56.35	57.07	57.48	57.46	57.12	56.89	57.14	56.76	57.09	58.84(3900)
TFIDF	52.10	56.43	57.34	57.50	57.75	57.86	58.26	58.51	58.93	58.55	58.93(1700)
GINI_{NTF}	52.20	56.58	57.51	57.83	57.86	58.05	58.60	58.83	59.10	58.63	59.10(1700)
GI_{cs}	52.30	56.62	57.80	58.66	58.76	58.83	59.30	59.51	59.63	59.31	59.63(1700)
DF_{cs}	52.21	56.80	57.45	58.16	58.64	58.94	59.56	59.39	59.39	**59.57**	59.57(1900)
CDM_{cs}	52.17	**56.85**	57.74	**58.78**	**59.06**	**59.12**	**59.78**	**59.81**	**59.97**	59.47	**59.97**(1700)
Acc2_{cs}	**52.39**	56.79	**57.87**	58.67	58.70	58.48	58.74	59.06	59.45	59.16	59.45(1700)

proposed methods and term frequency-based methods can achieve the best performance, while document frequency-based methods need to utilize more than 15% top features to achieve their best performance, which is less effective.

Then, for the purpose of more detailed study on a small feature set, Table 2 shows the $F1$ measure results when the size of the selected features is 100, 300, 500, 700, 900, 1100, 1300, 1500, 1700, and 1900. The best result for each feature set is shown in bold. It can be seen that on the $\text{Data}_{\text{BCIII}}$ the performance of the context similarity-based methods is also better than that of their corresponding frequency-based methods. And when the size of the feature set is 1700 (7.4% of the total number of features), CDM_{cs} acquires the highest $F1$ measure value 59.97, which improves the $F1$ measure of the $\text{Model}_{\text{SVM\_poly}}$ when all the features are used (57.12) by 2.85. Hence the context information of terms is helpful for the feature selection in IAC applications.

We notice that there is a significant drop in performance from the $\text{Data}_{\text{BCII}}$ to $\text{Data}_{\text{BCIII}}$, which suffered from the fact that the training article collection is extracted from different online article sources compared with the test data sets, and that the test data sets have the high class skew problem [30].

3.4. Analysis and Discussion

Comparison of the Selected Features. Besides the $F1$ measure results, we also analyze the effectiveness of feature selection methods through studying the profile of the selected features. The sorted lists of the top-10 features picked by each method are given in Tables 3 and 4 on the $\text{Data}_{\text{BCII}}$ and $\text{Data}_{\text{BCIII}}$, respectively. The features that are selected commonly by all the methods are indicated in bold. These common features make the same contribution to the classification performance, such as "interact" in Table 3 and "interaction" in Table 4. Hence we compare the special features selected by different methods. We note that there are two categories of special selected features according to two different feature selection principals. The first category features are the ones

Table 3: The top 10 features on the $Data_{BCII}$ selected by various feature selection methods. The terms that are selected commonly by all the methods are indicated in bold.

Number	GI	DF	CDM	Acc2	TFIDF	$GINI_{NTF}$	GI_{cs}	DF_{cs}	CDM_{cs}	$Acc2_{cs}$
1	protein	protein	hybrid	bind	proteins	protein	**interact**	**interact**	bind	**interact**
2	bind	bind	**interact**	interaction	cell	bind	bind	hybrid	**interact**	hybrid
3	interaction	proteins	protein	domain	receptor	domain	hybrid	bind	hybrid	bind
4	domain	cell	cell	complex	cells	proteins	binds	interaction	binds	interaction
5	proteins	**interact**	proteins	**interact**	kinase	**interact**	identified	binds	analyse	binds
6	complex	cells	spots	proteins	domain	cell	analyse	analyse	complex	analyse
7	cell	domain	binds	cell	bind	complex	interaction	domain	interaction	expression
8	terminal	analyse	cells	hybrid	beta	cells	activation	human	activity	human
9	**interact**	complex	spot	protein	protein	kinase	function	activity	domain	activity
10	cells	interaction	domains	interacts	**interact**	receptor	activity	identified	identified	identified

Table 4: The top 10 features on the $Data_{BCIII}$ selected by various feature selection methods. The terms that are selected commonly by all the methods are indicated in bold.

Number	GI	DF	CDM	Acc2	TFIDF	$GINI_{NTF}$	GI_{cs}	DF_{cs}	CDM_{cs}	$Acc2_{cs}$
1	protein	protein	interacts	protein	cells	protein	hybrid	interact	binds	binds
2	bind	cell	**interaction**	bind	cell	cells	interact	binds	interact	interact
3	results	cells	interact	**interaction**	expression	cell	binds	study	bind	bind
4	cell	bind	binds	domain	**interaction**	**interaction**	expression	bind	expression	expression
5	cells	results	gene	complex	bind	expression	bind	expression	**interaction**	study
6	study	**interaction**	domain	proteins	protein	proteins	study	**interaction**	activity	**interaction**
7	**interaction**	activity	cell	kinase	proteins	complex	subunit	activity	complex	complex
8	using	proteins	terminal	gene	gene	domain	**interaction**	subunit	domain	activity
9	gene	study	interacting	cell	genes	gene	activity	results	results	terminal
10	use	function	ubiquitin	interacts	human	human	increase	complex	activity	results

selected based on the statistical frequency. These features obtain higher scores because more documents contain them or they occur more. However, the term cooccurrences and chunks within the document are ignored. For example, the terms "protein" and "cell" are selected by all the frequency-based methods but the context similarity-based methods on both $Data_{BCII}$ and $Data_{BCIII}$. Considering "protein," it is just used to describe different protein names, which can appear anywhere in biological articles with the result of high document frequency or term frequency. However, it is not a distinctive feature to classify PPI relevant or nonrelevant articles. If such irrelevant features are assigned higher scores by a feature selection method, the performance obtained by those features would be degraded. On the contrary, these features are assigned lower values by our proposed methods, because their context dissimilarity between the PPI relevant and nonrelevant articles depresses their scores. The second category features are shared by the context similarity-based methods, such as the terms "activate" in Table 3 and "activity" in Table 4. Their evaluation scores are raised by the context similarity within the PPI relevant articles, which is important for the classification purpose.

In order to further study the proposed methods on common and special selected features, the top 1000 features are selected on both data sets, respectively. We perform experiments on the pairs of one context similarity-based method and one frequency-based feature selection method. First, the common features selected for each pair by both feature selection methods are fed into the $Model_{SVM\_poly}$. Then the performance of this $Model_{SVM\_poly}$ based on the common features is compared with the performance achieved based on all the top 1000 features selected by the context similarity-based method and the frequency-based method, respectively. Our purpose is to reveal which kind of feature selection methods can increase the performance more with their special selected features. The results are listed in Tables 5 and 6 on the $Data_{BCII}$ and $Data_{BCIII}$, respectively. It can be seen that the increments of context similarity-based methods are higher than the frequency-based methods, so the special features selected through context similarity-based methods can bring more distinctive information for the classifier on both data sets.

Dimension Reduction Rate. In addition to $F1$ measure, dimension reduction rate is another important aspect of feature selection. Therefore, a dimension reduction is also studied during the experiments. To compute a dimension reduction rate together with the $F1$ measure, a scoring scheme from Gunal and Edizkan [31] is defined as follows:

$$\text{Score} = \frac{1}{k}\sum_{i=1}^{k}\frac{\dim_N}{\dim_i}F1_i, \tag{13}$$

Table 5: The comparison of common and special selected features on the $Data_{BCII}$. *C* denotes the *F*1 measure of the $Model_{SVM\_poly}$ based on the common features selected by the frequency-based method and the context similarity-based method. The integer in parentheses is the number of the common features; CS denotes the *F*1 measure obtained by the context similarity-based method based on the top 1000 features. The number in parentheses is the increments compared with *C*; *F* denotes the *F*1 measure obtained by the frequency-based method based on the top 1000 features. The number in parentheses is the increments compared with *C*.

		GI_{cs}	DF_{cs}	CDM_{cs}	$Acc2_{cs}$
GI	*C*	71.62(714)	71.49(734)	71.40(702)	71.47(730)
	CS	76.64(+5.02)	76.57(+5.08)	76.81(+5.41)	76.60(+5.13)
	F	71.76(+0.14)	71.76(+0.27)	71.76(+0.36)	71.76(+0.29)
DF	*C*	72.31(625)	71.23(644)	71.06(613)	71.21(643)
	CS	76.64(+4.33)	76.57(+5.34)	76.81(+5.75)	76.60(+5.39)
	F	72.48(+0.17)	72.48(+1.25)	74.48(+3.42)	72.48(+1.27)
CDM	*C*	73.54(534)	73.17(552)	72.96(537)	73.16(552)
	CS	76.64(+3.10)	76.57(+3.40)	76.81(+3.85)	76.60(+3.44)
	F	74.92(+1.38)	74.92(+1.75)	74.92(+1.96)	74.92(+1.76)
Acc2	*C*	72.87(635)	73.40(650)	73.88(643)	73.40(650)
	CS	76.64(+3.77)	76.57(+3.17)	76.81(+2.93)	76.60(+3.20)
	F	74.04(+1.17)	74.04(+0.64)	74.04(+0.16)	74.04(+0.64)
TFIDF	*C*	72.90(668)	73.49(740)	73.40(692)	73.47(693)
	CS	76.64(+3.74)	76.57(+3.08)	76.81(+3.41)	76.60(+3.13)
	F	74.87(+1.97)	74.87(+1.38)	74.87(+1.47)	74.87(+1.40)
$GINI_{NTF}$	*C*	73.67(720)	73.77(754)	73.49(710)	73.60(730)
	CS	76.64(+2.97)	76.57(+2.80)	76.81(+3.32)	76.60(+3.00)
	F	75.10(+1.43)	75.10(+1.33)	75.10(+1.61)	75.10(+1.50)

Table 6: The comparison of common and special selected features on the $Data_{BCIII}$. *C* denotes the *F*1 measure of the $Model_{SVM\_poly}$ based on the common features selected by the frequency-based method and the context similarity-based method. The integer in parentheses is the number of the common features; CS denotes the *F*1 measure obtained by the context similarity-based method based on the top 1000 features. The number in parentheses is the increments compared with *C*; *F* denotes the *F*1 measure obtained by the frequency-based method based on the top 1000 features. The number in parentheses is the increments compared with *C*.

		GI_{cs}	DF_{cs}	CDM_{cs}	$Acc2_{cs}$
GI	*C*	55.22(740)	55.04(781)	55.54(764)	56.37(745)
	CS	58.78(+3.56)	58.76(+3.72)	59.09(+3.55)	58.57(+2.00)
	F	56.38(+1.16)	56.38(+1.34)	56.38(+0.84)	56.38(+0.01)
DF	*C*	57.02(579)	56.85(593)	56.70(658)	57.20(609)
	CS	58.78(+1.76)	58.76(+1.91)	59.09(+2.39)	58.57(+1.37)
	F	57.61(+0.59)	57.61(+0.76)	57.61(+0.91)	57.61(+0.41)
CDM	*C*	57.22(544)	57.11(545)	57.50(550)	57.18(560)
	CS	58.78(+1.56)	58.76(+1.65)	59.09(+1.59)	58.57(+1.39)
	F	57.09(−0.13)	57.09(−0.02)	57.09(−0.41)	57.09(−0.09)
Acc2	*C*	56.17(656)	56.13(678)	57.00(671)	56.51(673)
	CS	58.78(+2.61)	58.76(+2.63)	59.09(+2.09)	58.57(+2.06)
	F	57.09(+0.92)	57.09(+0.96)	57.09(+0.09)	57.09(+0.58)
TFIDF	*C*	57.20(668)	57.14(701)	57.05(692)	57.09(690)
	CS	58.78(+1.58)	58.76(+1.62)	59.09(+2.04)	58.57(+1.48)
	F	57.80(+0.60)	57.80(+0.66)	57.80(+0.75)	57.80(+0.71)
$GINI_{NTF}$	*C*	57.35(720)	57.19(754)	57.17(715)	57.30(698)
	CS	58.78(+1.43)	58.76(+1.57)	59.09(+1.92)	58.57(+1.27)
	F	57.96(+0.61)	57.96(+0.77)	57.96(+0.79)	57.96(+0.66)

Table 7: Rate scores of dimension reduction on the $Data_{BCII}$ and $Data_{BCIII}$, respectively.

	GI	DF	CDM	Acc2	TFIDF	$GINI_{NTF}$	GI_{cs}	DF_{cs}	CDM_{cs}	$Acc2_{cs}$
$Data_{BCII}$	4640	4642	4664	4679	4693	4700	4729	4734	4738	4731
$Data_{BCIII}$	2684	2667	2693	2695	2705	2712	2727	2723	2729	2728

where k is the number of trails, $\dim_N$ is the maximum feature size, $\dim_i$ is the feature size at the ith trail, and $F1_i$ is the $F1$ measure of the ith trail. Here, $\dim_i$ is a set of sequences, 100, 300, 500, 700, 900, 1100, 1300, 1500, 1700, and 1900, and so k is 10. The results of dimension reduction analysis using the described scoring scheme are presented in Table 7. It is apparent from this table that the context similarity-based feature selection methods provide better performance than the frequency-based methods.

4. Conclusions

In this paper, novel context similarity-based feature selection methods were introduced for text classification in the biological domain to classify protein interaction articles. They assign importance scores to features based on their similarity measure of context information within certain text categories. Using two different data sets, the performance of the proposed methods was investigated and compared against four document frequency-based and two term frequency-based methods. The effectiveness of the proposed methods was demonstrated and analyzed on the $F1$ measure, the profile of selected features, and dimension reduction rate for the IAC tasks. Since IAC is a binary text classification task in biological domain, we also want to know the performance of the proposed methods when they are extended to multiclass problems. Hence, an adaptation of the context similarity-based selection method to multiclassification problems remains an interesting future task.

Conflict of Interests

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

Acknowledgments

The authors want to thank the anonymous reviewers for their helpful comments and suggestions. This work is supported by the National Natural Science Foundation of China (nos. 61202135 and 61402231), the Natural Science Foundation of Jiangsu Province (nos. BK2012472 and BK2011692), the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (no. 12KJD520005), and the Qing Lan Project.

References

[1] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, "A feature selection method based on improved fisher's discriminant ratio for text sentiment classification," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8696–8702, 2011.

[2] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to Spam filtering," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10206–10222, 2009.

[3] B. Zhou, Y. Y. Yao, and J. Lou, "A three-way decision approach to email spam filtering," in *Proceedings of the 23rd Canadian Conference on Artificial Intelligence (Canadain AI '10)*, vol. 6085 of *Lecture Notes in Computer Science*, pp. 28–39, 2010.

[4] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi, "Author gender identification from text," *Digital Investigation*, vol. 8, no. 1, pp. 78–88, 2011.

[5] S. A. Özel, "A web page classification system based on a genetic algorithm using tagged-terms as features," *Expert Systems with Applications*, vol. 38, no. 4, pp. 3407–3415, 2011.

[6] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, 2007.

[7] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the 14th International Conference on Machine Learning (ICML '97)*, pp. 412–420, 1997.

[8] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[9] N. Azam and J. T. Yao, "Incorporating game theory in feature selection for text categorization," in *Proceedings of the 13th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC '11)*, vol. 6743 of *Lecture Notes in Computer Sience*, pp. 215–222, Springer, 2011.

[10] H. Liang, J. Wang, and Y. Yao, "User-oriented feature selection for machine learning," *The Computer Journal*, vol. 50, no. 4, pp. 421–434, 2007.

[11] S. Piramuthu, "The protein-protein interaction tasks of biocreative III: evaluating feature selection methods for learning in data mining applications," *European Journal of Operational Research*, vol. 156, no. 2, pp. 483–494, 2004.

[12] Y. Y. Yao and Y. Zhao, "Attribute reduction in decision-theoretic rough set models," *Information Sciences*, vol. 178, no. 17, pp. 3356–3373, 2008.

[13] Y. Y. Yao, Y. Zhao, and J. Wang, "On reduct construction algorithms," *Transactions on Computational Science*, vol. 2, pp. 100–117, 2008.

[14] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[15] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A novel feature selection algorithm for text categorization," *Expert Systems with Applications*, vol. 33, no. 1, pp. 1–5, 2007.

[16] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5432–5435, 2009.

[17] N. Azam and J. T. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," *Expert Systems with Applications*, vol. 39, no. 5, pp. 4760–4768, 2012.

[18] J. Yang, Y. Liu, X. Zhu, Z. Liu, and X. Zhang, "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization," *Information Processing and Management*, vol. 48, no. 4, pp. 741–754, 2012.

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

[20] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[21] Y.-Q. Wei, P.-Y. Liu, and Z.-F. Zhu, "A feature selection method based on improved TFIDF," in *Proceedings of the 3rd International Conference on Pervasive Computing and Applications (ICPCA '08)*, pp. 94–97, Alexandria, Egypt, October 2008.

[22] W. E. Winkler, "tring comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage," in *Proceedings of the Section on Survey Research Methods (American Statistical Association)*, vol. 359, pp. 354–359, 1990.

[23] M. A. Jaro, "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414–420, 1989.

[24] M. A. Jaro, "Probabilistic linkage of large public health data files," *Statistics in Medicine*, vol. 14, no. 5–7, pp. 491–498, 1995.

[25] V. N. Vapnik, *Statistical Learning Theory*, Adaptive and Learning Systems for Signal Processing, Communications, and Control, John Wiley & Sons, New York, NY, USA, 1998.

[26] T. Xia and Y. Du, "Improve VSM text classification by title vector based document representation method," in *Proceedings of the 6th International Conference on Computer Science and Education (ICCSE '11)*, pp. 210–213, August 2011.

[27] M. Antunes, C. Silva, B. Ribeiro, and M. Correia, "A hybrid ais-svm ensemble approach for text classification," in *Proceedings of the 10th International Conference on Adaptive and Natural Computing Algorithms*, pp. 342–352, 2011.

[28] C.-C. Chang and C.-J. Lin, "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[29] M. Krallinger and A. Valencia, "Evaluating the detection and ranking of protein interaction relevant articles: the biocreative challenge interaction article sub-task (ias)," in *Proceedings of the 2nd BioCreative Challenge EvaluationWorkshop*, pp. 29–39, 2007.

[30] M. Krallinger, M. Vazquez, F. Leitner et al., "The protein-protein interaction tasks of bioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text," *BMC Bioinformatics*, vol. 12, supplement 8, article S3, 2011.

[31] S. Gunal and R. Edizkan, "Subspace based feature selection for pattern recognition," *Information Sciences*, vol. 178, no. 19, pp. 3716–3726, 2008.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 852070, 12 pages
http://dx.doi.org/10.1155/2015/852070

Research Article

Spatially Enhanced Differential RNA Methylation Analysis from Affinity-Based Sequencing Data with Hidden Markov Model

Yu-Chen Zhang,[1] Shao-Wu Zhang,[1] Lian Liu,[1] Hui Liu,[2] Lin Zhang,[2] Xiaodong Cui,[3] Yufei Huang,[3] and Jia Meng[4]

[1] *Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an 710072, China*
[2] *School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116, China*
[3] *Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX 78249, USA*
[4] *XJTLU-WTNC Research Institute, Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China*

Correspondence should be addressed to Shao-Wu Zhang; zhangsw@nwpu.edu.cn and Jia Meng; jia.meng@xjtlu.edu.cn

Received 12 February 2015; Accepted 25 March 2015

Academic Editor: Fang-Xiang Wu

With the development of new sequencing technology, the entire N6-methyl-adenosine (m^6A) RNA methylome can now be unbiased profiled with methylated RNA immune-precipitation sequencing technique (MeRIP-Seq), making it possible to detect differential methylation states of RNA between two conditions, for example, between normal and cancerous tissue. However, as an affinity-based method, MeRIP-Seq has yet provided base-pair resolution; that is, a single methylation site determined from MeRIP-Seq data can in practice contain multiple RNA methylation residuals, some of which can be regulated by different enzymes and thus differentially methylated between two conditions. Since existing peak-based methods could not effectively differentiate multiple methylation residuals located within a single methylation site, we propose a hidden Markov model (HMM) based approach to address this issue. Specifically, the detected RNA methylation site is further divided into multiple adjacent small bins and then scanned with higher resolution using a hidden Markov model to model the dependency between spatially adjacent bins for improved accuracy. We tested the proposed algorithm on both simulated data and real data. Result suggests that the proposed algorithm clearly outperforms existing peak-based approach on simulated systems and detects differential methylation regions with higher statistical significance on real dataset.

1. Introduction

Although the presence of posttranscriptional biochemical modifications to RNA has been established in 1960s [1], due to historical limitations, RNA epigenetics is largely uncharted territory until recently [2–4]. In 2012, a powerful sequencing protocol methylated RNA immune-precipitation sequencing (MeRIP-Seq or m^6A-Seq) was developed [5, 6], in which the fragmented mRNA fragments with N6-methyl-adnosine (m^6A) are pulled down with anti-m^6A antibody and then purified and passed to subsequent sequencing to generate the so-called "IP sample" for profiling the transcriptome-wide RNA m^6A methylome. Very often, a paired "input sample" is generated as well using all the RNA for measuring the entire transcriptome background (please refer to [7] for a more comprehensive protocol of this approach). This technique facilitates a number of research findings recently which includes the following: the role of RNA methylation in controlling the circadian clock [8], addiction [9], and stem cell [10], and [2, 3, 5, 6, 8–16]. It also enabled the construction of mammalian RNA methylation database [17] and systems biology approaches for decomposing the RNA methylome to unveil the latent enzymatic regulators of epitranscriptome [18]. Software tools for RNA methylation site detection [19, 20] and for differential RNA methylation analysis [21] from MeRIP-Seq data are now available in a rather user friendly

manner. Nevertheless, as a newly arising technique, MeRIP-Seq still poses computational challenges that call for novel and sophisticated approaches.

Differential methylation analysis is of crucial importance for epigenetics research. Differentially methylated regions (DMRs), that is, regions that exhibit different methylation levels between two experimental conditions, for example, normal and cancerous, can be as small as a single base or as large as an entire gene locus, depending on the biological question of interest and the bioinformatics methods used for their identification [22]. Differential methylation analysis from MeRIP-Seq seeks to identify the differences in RNA methylome in a case-control study (e.g., cancerous and normal), which usually involves at least four high-throughput sequencing (HTS) samples, including the IP and input samples under both the case and control conditions. For affinity-based methods developed for DNA epigenetics (such as MeDIP-Seq and ChIP-Seq), since the absolute amount of DNA is most likely to stay unchanged between two conditions, the percentage of modified DNA molecule is linearly correlated with the absolute amount; thus the difference in methylation is consistent when measured in relative (percentage) and absolute amount. However, in MeRIP-Seq, due to the change in transcriptional expression level between two conditions, it is possible that while the absolute amount of methylated RNA increases, the relative amount (percentage of methylated RNA) decreases as shown in Figure 1. From computational perspective, the differential methylation analysis of RNA is quite different from that of DNA, and DNA differential methylation approaches [23], such as MOABS [24] and DMAP [25], may not be directly applicable to RNA. Until now, methods aiming at the differential analysis of MeRIP-Seq data do not extensively appear in literature. exomePeak [19, 21] is dedicatedly developed for differential RNA methylation analysis from MeRIP-Seq data. The detection of DMRs is based on rhtest [26], which is an extended version of hypergeometric test, computing the statistical significance of the difference in the percentages of methylated fragments between the two conditions, which directly indicates the difference in enzymatic regulation. Before the detection of DMRs, peaks (methylated regions) are called firstly from the transcriptome by comparing the IP with input sample by relative enrichment [7, 19, 27]. Only with the detected methylation sites can we effectively estimate the methylation level.

Affinity-based approaches cannot provide single-base resolution. Since multiple RNA methylation residuals may locate in proximity and cannot be effectively differentiated with peak calling procedure, they can appear as a single broad methylation site in the peak calling result from MACS [27] or exomePeak [19]. In many cases, this discrepancy can be trivial and does not significantly affect relevant study; however, it can be disastrous in differential methylation analysis, because multiple RNA methylation residuals can be regulated by different enzyme complexes and thus may be differentially methylated. Failing to identify the precise location of each methylation residual can lead to large bias in the estimation

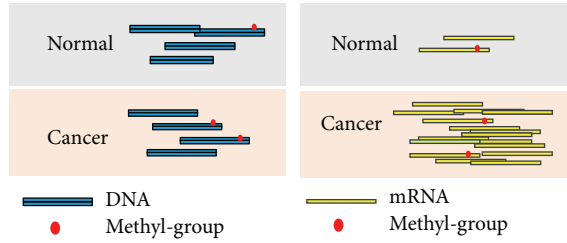


Figure 1: Comparison of the differential methylation analysis in DNA and RNA. The first column shows the DNA related differential analysis in ChIP-Seq or MeDIP-Seq, where the total DNA is often considered the same under two experimental conditions, so the differential analysis can be performed by directly comparing the absolute amount of methylated RNAs in the two IP samples. In contrast, for RNA (the second column), the background is total RNA, which can vary significantly under different conditions, and therefore, the absolute amount of methylated RNA for a specific site does not necessarily correlate with the degree of methylation. For the example shown in the above figure, while amount of methylated RNA increases under the cancer condition, the relative amount (percentage of methylated RNA) decreases, indicating a hypomethylation at RNA level. As a result, the differential analysis of RNA methylome in MeRIP-Seq should be performed by comparing the percentages of methylated RNA to reflect the influence of methylation enzymatic regulation.

of its methylation level and in the comparison to a different condition. Currently, all existing methods for RNA differential methylation from MeRIP-Seq data are peak-based. In this paper, based on the rhtest method developed in exomePeak package [21], we proposed FET-HMM, a novel strategy for spatially enhanced differential RNA methylation analysis using hidden Markov model (HMM). When applying to the RNA methylation site detected from a peak calling algorithm, FET-HMM breaks a single site into multiple adjacent small bins and evaluates whether a specific bin is differentially methylated or not between two experimental conditions with spatial dependency incorporated by HMM. Figure 2 shows the comparison between existing and our methods.

HMM is a statistical model that integrates multiple random processes and has been widely used in DNA-templated epigenetic analysis and in RNA methylation sites detection (or peak calling) [28–30], but so far it has not been applied for RNA differential methylation analysis. We applied the newly developed approach FET-HMM on both simulated and real datasets. The results on simulated data showed that FET-HMM can effectively improve the performance of rhtest in terms of the area under the curve (AUC) when detecting differential methylation sites. When applied to human MeRIP-Seq datasets, FET-HMM method returns more biological meaningful results than exomePeak method. The FET-HMM algorithm has been implemented in an open source R package for differential methylation analysis from MeRIP-Seq data and is freely available from GitHub. The method is detailed in the following section.

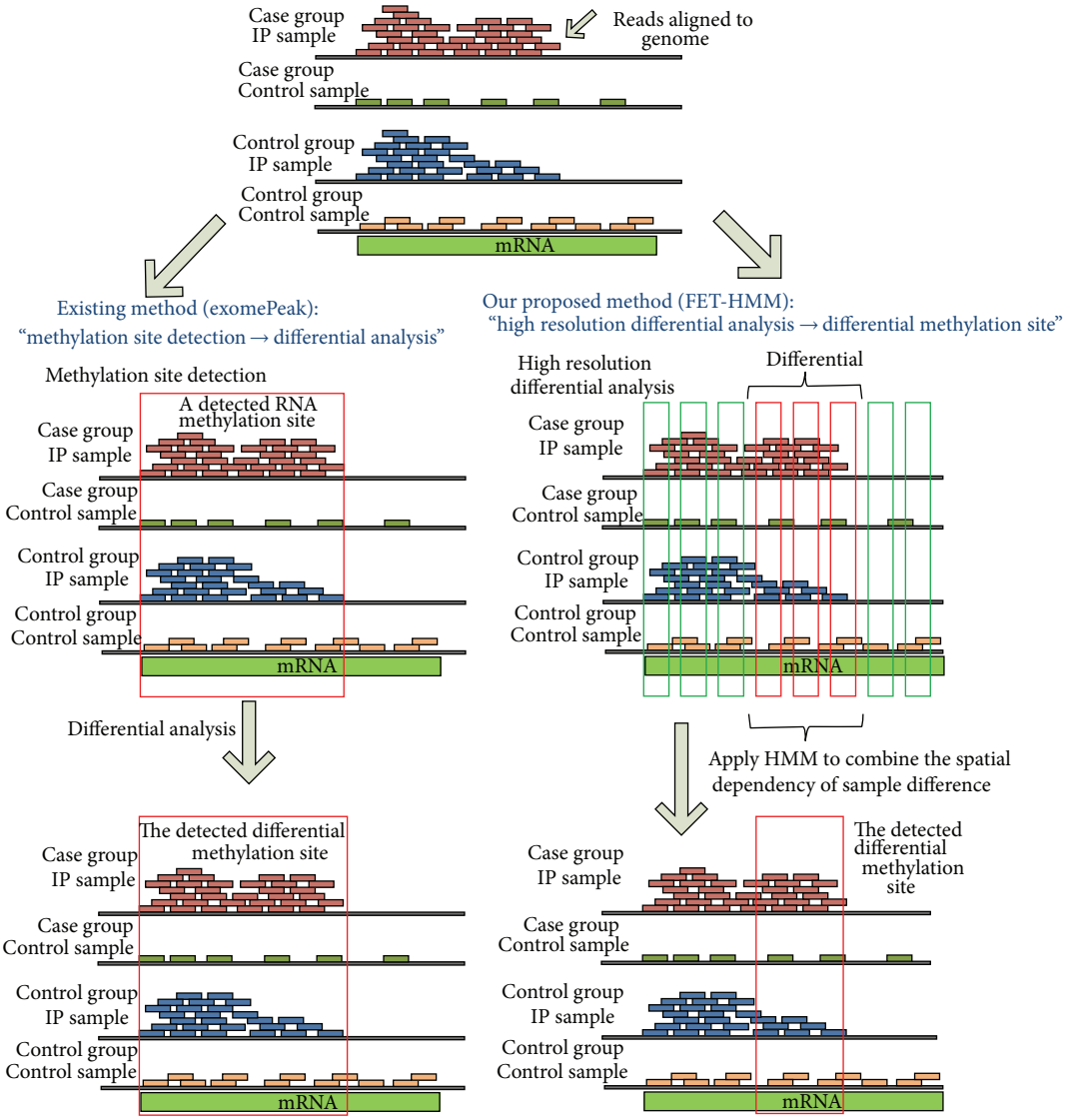


FIGURE 2: Comparison of differential methylation analysis methods. This figure shows the difference between existing peak-based differential analysis method and the proposed method. Started from aligned reads, the left part of this figure shows how exomePeak conducts differential analysis. It firstly identifies a single methylation site and then decides whether the methylation site as a whole is differentially methylated or not. However, the newly proposed method will split the testing region into multiple adjacent small bins and then will integrate their dependency with HMM for more accurate identification of differential methylation site. In the above example, the RNA methylation site detected using exomePeak method may consist of two methylation residuals, and only the one on the right side is differentially methylated in this case-control study. The proposed FET-HMM method is likely to work better than peak-based exomePeak method under this scenario.

2. Methods

In this section, we firstly review the usage of rhtest, a modified version of Fisher's exact test (FET), for differential RNA methylation analysis and then introduce spatially enhanced approach FET-HMM.

2.1. Peak-Based Differential RNA Methylation Analysis with Rhtest. To conduct differential RNA methylation analysis in a case-control study, we should get four samples, that is, the IP and input samples from both groups. Consider that there are a number of RNA methylation sites detected with peak calling approaches [19, 20, 27] from MeRIP-Seq. Then we can assume that the number of reads within the gth RNA methylation sites follows the Poisson distribution, with

$$\begin{aligned} X_{0,g} &\sim \text{Poisson}\left(N_0\lambda_{0,g}\right), \\ X_{1,g} &\sim \text{Poisson}\left(N_1\lambda_{1,g}\right), \\ Y_{0,g} &\sim \text{Poisson}\left(M_0\overline{\lambda}_{0,g}\right), \\ Y_{1,g} &\sim \text{Poisson}\left(M_1\overline{\lambda}_{1,g}\right), \end{aligned} \tag{1}$$

where $X_{0,g}$ and $X_{1,g}$ are the reads counts of the input samples for untreated and treated condition and consistently, $Y_{0,g}$ and $Y_{1,g}$ are the reads counts of the IP samples for untreated and treated samples. Here, $g = 1, 2, \ldots, G$ indicates the gth RNA methylation site. (N_0, N_1, M_0, M_1) are the size (or the sequencing depth) of library, respectively; and the parameters $(\lambda_{0,g}, \lambda_{1,g}, \overline{\lambda}_{0,g}, \overline{\lambda}_{1,g})$ are the normalized Poisson means in a standard library, indicating the expectation of the reads counts within a bin. Following the formulation from previous study [26], we assume that $\overline{\lambda}_{0,g}$ and $\overline{\lambda}_{1,g}$ satisfy the following relationship with $\overline{\lambda}_{0,g} = \lambda_{0,g}\eta_{0,g}/f_0$ and $\overline{\lambda}_{1,g} = \lambda_{1,g}\eta_{1,g}/f_1$, where f_0 and f_1 indicate the percentage of the expressed RNA fragments that are modified in the untreated and treated samples, respectively. $\eta_{0,g}$ and $\eta_{1,g}$ indicate the percentage of RNA fragments mapped inside the RNA methylation site that carry the methylation mark. We would like to test whether $\eta_{1,g} = \eta_{0,g}$. According to the properties of the Poisson distributions [31, 32], given $X_{0,g} + Y_{0,g} = t_{0,g}$, $X_{1,g} + Y_{1,g} = t_{1,g}$, we should have $X_{0,g} \sim \text{Binomial}(p_{0,g}, t_{0,g})$ and $X_{1,g} \sim \text{Binomial}(p_{1,g}, t_{1,g})$, where $p_{1,g} = N_1 f_1/(N_1 f_1 + M_1\eta_{1,g})$ and $p_{0,g} = N_0 f_0/(N_0 f_0 + M_0\eta_{0,g})$. For different experimental conditions, if we assume that the total amount of modifications remains the same, only its distribution may change, then we can have $f_0 = f_1 = f$. We also notice that if $N_1 M_0 = N_0 M_1$, then $\eta_{1,g} = \eta_{0,g} \Leftrightarrow p_{1,g} = p_{0,g}$, and testing whether the two Binomial distributions have the same successful rate is equivalent to the classical problem of testing the independence in a 2×2 contingency table. In order to establish $N_1 M_0 = N_0 M_1$, only one of the 4 samples needs to be rescaled. When $N_1 M_0 = N_0 M_1$ is achieved after rescaling, under the null hypothesis $p_{1,g} = p_{0,g}$, $X_{0,n}$ follows a hypergeometric distribution as in (2), and we may use Fisher's exact test [33–36] with two tails to evaluate its significance. Consider

$$p\left(X_{0,g} = k\right) \sim \text{Hyper}\left(X_{0,g} \mid K, n, N\right) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \quad (2)$$

where $N = t_{0,g} + t_{1,g} = x_{0,g} + x_{1,g} + y_{0,g} + y_{1,g}$, $n = t_{0,g} = x_{0,g} + y_{0,g}$, and $K = x_{0,g} + x_{1,g}$. The smaller the p value is, the more likely the gth RNA methylation site is differentially methylated between two conditions.

2.2. Spatially Enhanced Differential RNA Methylation Analysis with FET-HMM. The method developed in the previous section could not effectively discriminate multiple RNA methylation residuals located within a single RNA methylation site (as shown in Figure 1). We seek to enhance the spatial resolution with hidden Markov model. Similar to various formulation, for a particular RNA methylation site, we firstly divided it into N mutually connected bins of length L. Then we can still assume that the number of reads within the nth bin follows the Poisson distribution, with

$$\begin{aligned} X_{0,n} &\sim \text{Poisson}\left(N_0\lambda_{0,n}\right), \\ X_{1,n} &\sim \text{Poisson}\left(N_1\lambda_{1,n}\right), \\ Y_{0,n} &\sim \text{Poisson}\left(M_0\overline{\lambda}_{0,n}\right), \\ Y_{1,n} &\sim \text{Poisson}\left(M_1\overline{\lambda}_{1,n}\right), \end{aligned} \quad (3)$$

where $X_{0,n}$ and $X_{1,n}$ are the reads counts of the input samples for untreated and treated condition and consistently, $Y_{0,n}$ and $Y_{1,n}$ are the reads counts of the IP samples for untreated and treated samples. Here, $n = 1, 2, \ldots, N$ indicates the nth bin. The parameters $(\lambda_{0,n}, \lambda_{1,n}, \overline{\lambda}_{0,n}, \overline{\lambda}_{1,n})$ are the normalized Poisson means in a standard library, indicating the expectation of the reads counts within a bin. Following the formulation from previous study [26], we assume that $\overline{\lambda}_{0,n}$ and $\overline{\lambda}_{1,n}$ satisfy the following relationship with $\overline{\lambda}_{0,n} = \lambda_{0,n}\eta_{0,n}/f_0$ and $\overline{\lambda}_{1,n} = \lambda_{1,n}\eta_{1,n}/f_1$, where f_0 and f_1 indicate the percentage of the expressed RNA fragments that are modified in the untreated and treated samples, respectively. $\eta_{0,n}$ and $\eta_{1,n}$ indicate the percentage of RNA fragments mapped inside the bin that carry the methylation mark. We can easily test whether $\eta_{1,n} = \eta_{0,n}$ (whether differential methylation is observed) for a specific bin; however, we should not neglect the dependencies between the reads counts of adjacent bins within an RNA methylation site; that is, if differential methylation is observed on a specific bin, it is likely that differential methylation can also be observed on bins adjacent to it and vice versa. The dependency can be effectively incorporated with an HMM formulation, and we thus developed a new strategy for the identification of differential methylation regions (DMRs) with improved spatial resolution.

To begin with, with respect to nth bin, the hidden true states of differential methylation are denoted as $S = \{s_1, s_2, \ldots, s_N\}$, where $s_n \in \{0, 1\}$ with 1 representing differential methylation state (DMS) and 0 otherwise. Considering that a differential methylation region may span multiple adjacent bins, we assume that the true hidden DMS S follows a first order Markov chain, whose transition matrix A contains entries defined as

$$A_{ij} = P\left(s_{n+1} = j \mid s_n = i\right), \quad i, j \in \{0, 1\}, \quad (4)$$

where A_{ij} denotes the probability for the hidden variable switching from DMS i at the nth bin to the DMS j at the $(n+1)$th bin. In addition, the initial probability $p(S_1 = 0) = u$ and $p(S_1 = 1) = 1 - u$, which can be denoted as $\pi = (u, 1 - u)$. Next, the result of rhtest [21, 26] was used as the observed variable of the HMM. However, the information acquired from rhtest is a statistical significance of differential methylation in terms of p values and FDRs (False Discovery Rates). We seek to enhance the differential methylation results by incorporating spatial dependency. Specifically, 3 different strategies are developed for this purpose with their own advantages and disadvantages, which are detailed in the following.

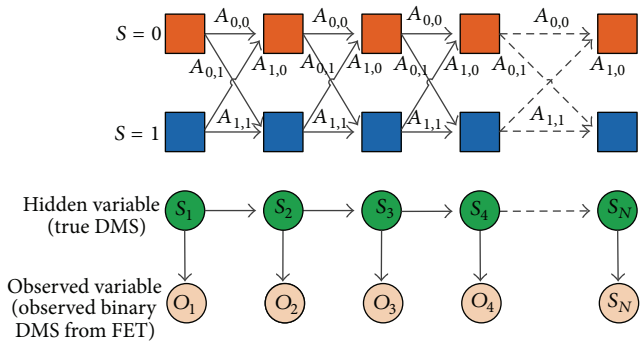


FIGURE 3: Hidden Markov model. In FHB strategy, the "observation" is a binary status reported from FET, and the emission probability is Bernoulli distribution.

2.3. FHB Strategy: Combine Fisher's Exact Test and HMM with Binary Observation. In FHB strategy, we use the binary decisions received from FET as the observation of hidden Markov model. The model essentially evaluates how likely a true differential methylation state can be detected by FET, or if FET reports a DMS with a significance level, how likely it is true after incorporating spatial dependency. We assume that a state can be correctly observed with probability p; and a mistake happens with probability $(1 - p)$. Since the observation from FET is considered as binary, a cut-off threshold should be used to switch the FDR (False Discovery Rate) value to generate the "observed" set of observed variable $O = (o_1, o_2, \ldots, o_n)$ with $o_n \in \{0, 1\}$. Then according to the standard HMM definition, these probabilities consist of an emission matrix B, whose entries are defined as

$$B_{ij} = P\left(o_n = j \mid s_n = i\right) = \begin{cases} p, & i, j \in \{0, 1\},\ i = j, \\ 1 - p, & i, j \in \{0, 1\},\ i \neq j. \end{cases} \quad (5)$$

The detailed structure of HMM is shown in Figure 3.

Finally, we applied the widely used Baum-Welch algorithm [37–39] to estimate the unknown parameters of the HMM. Baum-Welch algorithm applies the well-known Expectation and MSaximization (EM) strategy to conduct the process of estimation. The implementation steps of Baum-Welch algorithm are as follows.

The Proposed Algorithm

(1) Initialization. Given the initial value of A_{ij}, π_i, and B_{ij} randomly according to the conditions of probability, we hence get the initial model parameters $\lambda^{(0)} = (\pi^{(0)}, A^{(0)}, B^{(0)})$.

(2) EM Steps

E Step. Let $\gamma_n(i)$ denote the probability of the hidden DMS being at i at the nth bin, and let $\xi_n(i, j)$ denote the probability of the hidden DMS being at i at the nth bin and the DMS being at j at the $(n+1)$th bin. Also, we denote $ts_{ik}, k \in \{0, 1\}$, to represent the times of the transition from DMS i to any DMS k and ts_{ij} to represent the times of the transition from DMS i to the DMS j. $\gamma_n(i)$ and $\xi_n(i, j)$ can be computed through (6) and (7), and the expectation of ts_{ik} and ts_{ij} can be calculated by (8) and (9). $\lambda^{(m)} = (\pi^{(m)}, A^{(m)}, B^{(m)})$ represents the parameters of HMM after the mth iteration. Consider

$$\gamma_n(i) = P\left(s_n = i \mid O, \lambda^{(m)}\right) = \frac{P\left(s_n = i,\ O \mid \lambda^{(m)}\right)}{P\left(O \mid \lambda^{(m)}\right)}, \quad (6)$$

$$\xi_n(i, j) = P\left(s_n = i,\ s_{n+1} = j \mid O, \lambda^{(m)}\right) = \frac{P\left(s_n = i,\ s_{n+1} = j,\ O \mid \lambda^{(m)}\right)}{P\left(O \mid \lambda^{(m)}\right)}, \quad (7)$$

$$E\left[t_{s_{ik}}\right] = \sum_{n=1}^{N-1} \gamma_n(i), \quad (8)$$

$$E\left[t_{s_{ij}}\right] = \sum_{n=1}^{N-1} \xi_n(i, j). \quad (9)$$

M Step. After using (10), (11), and (12) to estimate π_i, A_{ij}, and B_{ij}, we get $\lambda^{(m+1)}$. One has

$$\pi_i^{(m+1)} = \gamma_1(i), \quad (10)$$

$$a_{ij}^{(m+1)} = \frac{E\left[t_{s_{ij}}\right]}{E\left[t_{s_{ik}}\right]} = \frac{\sum_{n=1}^{N-1} \xi_n(i, j)}{\sum_{n=1}^{N-1} \gamma_n(i)}, \quad (11)$$

$$b_i^{(m+1)}(k) = \frac{\sum_{n=1}^{N} \gamma_n(i)\, I_{\{o_n = k\}}}{\sum_{n=1}^{N} \gamma_n(i)}. \quad (12)$$

In (12),

$$I_{\{o_n = k\}} = \begin{cases} 1 & o_n = k \\ 0 & o_n \neq k \end{cases} \quad (13)$$

is the indicative function.

(3) Loop. Repeat the EM steps until the convergence of A_{ij}, π_i, and B_{ij}. After the procedures above, optimal model parameter $\lambda^{(\mathrm{op})}$ could be obtained. Let $u_{nk} = 1$ if we are absolutely sure $s_n = k$ and $u_{nk} = 0$ otherwise. What we focused on is the final expectation of $u_{nk}, k \in \{0, 1\}$, which can be calculated as

$$E\left[u_{nk} \mid O, \lambda^{(\mathrm{op})}\right] = P\left(s_n = k \mid O, \lambda^{(\mathrm{op})}\right). \quad (14)$$

Then we could obtain the posterior probability of a bin being at a specific state, and the performance of FET-HMM can be compared with that of exomePeak on simulated dataset when the true state is available.

2.4. FHC Strategy: Combine Fisher's Exact Test and HMM with Continuous Observation. In FHB strategy, we adopt a switching cut-off threshold to convert the statistical significance

(p value from differential analysis with rhtest) into binary states as the observation of HMM. This strategy has two limitations. Firstly, we could hardly find the most reasonable threshold for a dataset, and different threshold can lead to different results. Secondly, some information gets lost in the conversion from p value to binary states; for example, both p values 0.01 and 0.001 are converted as DMS state 1 after a binary conversion with significance level 0.05; however, the former is less confident. In addition, Bernoulli distribution may not be the most suitable distribution for the emission probability of observed variable. Therefore, a strategy seeking to directly smooth the continuous statistical significance without binary conversion may be superior. For this purpose, we use the p values from FET to approximate the likelihood of a bin with DMS state 0 and $(1 - p$ value) for its likelihood with DMS state 1. The p values generated from FET can be used to estimate the emission probability of HMM directly and then passed to HMM for smoothing purposes. It should be denoted as

$$B_{\mathrm{II}} = \begin{bmatrix} p \text{ value}_1 & 1 - p \text{ value}_1 \\ p \text{ value}_2 & 1 - p \text{ value}_2 \\ \vdots & \vdots \\ p \text{ value}_N & 1 - p \text{ value}_N \end{bmatrix}. \tag{15}$$

After getting the matrix B_{II} of size N by 2 constructed from FET p values, the Baum-Welch algorithm introduced in FHB can be applied to spatially enhance the local result, with formula (12) omitted because matrix B_{II} does not need to be reestimated every iteration. Please note that using p values to approximate directly the probability matrix B_{II} helps to avoid the binary conversion and information loss, and we will show in the Result section that this trick indeed improves the performance of algorithm.

2.5. FastFH Strategy: A High-Efficiency Strategy for Applying FET-HMM on Big Omics Data. When the proposed method is used in real MeRIP-Seq dataset, two problems would emerge. What comes first was some reads would be mapped into very short genes; thus the number of the bins would be quite small. In other words, the length of some Markov chains would be too short for accurate estimation of parameters and finally affects the results of DMRs detection. In addition, computational time was another important factor that we should take into consideration. Take the human hg19 data we were going to test as an example. If there were more than 30000 detected RNA methylation sites in total, the Baum-Welch algorithm would be performed more than 30000 times and the execution time might be too long. In order to solve these two limitations, we could combine the two strategies together. Firstly, the threshold used in FHB was used here again to switch the FDR into binary DMS. Then we could estimate transition matrix A_{III} directly from this DMS information as shown in

$$\begin{aligned} \pi_{\mathrm{III}} &= \left(1 - \frac{\sum_{i=1}^{N} \mathrm{DMS}_i}{N}, \frac{\sum_{i=1}^{N} \mathrm{DMS}_i}{N}\right), \\ A_{\mathrm{III}} &= \begin{bmatrix} P\left(S_{n+1} = 0 \mid S_n = 0\right) & P\left(S_{n+1} = 1 \mid S_n = 0\right) \\ P\left(S_{n+1} = 0 \mid S_n = 1\right) & P\left(S_{n+1} = 1 \mid S_n = 1\right) \end{bmatrix}, \end{aligned} \tag{16}$$

where $P(S_{n+1} \mid S_n)$ denotes the conditional probability for the transition from S_n to S_{n+1}, which can be conveniently estimated by scanning all the states of differential methylation $S = \{s_1, s_2, \ldots, s_N\}$ on all RNA methylation sites. For every single gene, the emission probability B_{III} has the same form as B_{II} in FHC strategy. By doing this, the A_{III} matrix can be estimated in a single step instead of an iterative manner so as to save computation load. This result should be also more robust on short RNA methylation sites with less number of bins than previous strategy. Secondly, we chose the Estep in FHB strategy to compute the final expectation defined in formula (14) for every single bin on every RNA methylation sites of real RNA epigenetics data. FastFHC strategy applied Estep after estimating transition matrix and initial probability for all genes. π_{III} and A_{III} are considered the same on different RNA methylation sites and are estimated like FHB with binary converted observation. Although some information can be lost in the conversion step, since tens of thousands of RNA methylation sites are pooled together for estimation of π_{III} and A_{III}, it should be still relatively accurate. The 3 strategies are summarized in Figure 4.

3. Result

3.1. Test on Simulated Data. For MeRIP-Seq, as the ground truth is not available for the differential RNA methylation status in real data, the performance of our proposed method (FHB and FHC strategy) was first validated on simulated datasets. Specifically, the reads counts for the IP and input samples under two experimental conditions were generated from model assumptions, respectively. In every set of data, 100 RNA methylation sites are generated, each with 1000 adjacent bins. The sequencing depths were all set 10^8, and the normalized Poisson mean λ_0 of untreated input was set to 10^{-6}, unless otherwise clarified. To simulate differential expression, reads counts of each gene in both the IP and the input control sample also vary in a certain range compared with the untreated condition, respectively; and we assume its log2 fold change follows a uniform distribution between $[-3, 3]$. To mimic differential methylation, the methylation reads counts log2 odds ratio follows a uniform distribution between $[-3, 3]$ for differential methylation bins and 0 for nondifferential bins. In order to impose dependency of adjacent bins on the simulated data, we applied a definite HMM to generate the labels used as the hidden DMS of the 1000 adjacent bins to indicate whether a bin is differential methylated or not. Then the label was used to generate the data and also used as the ground truth for evaluating

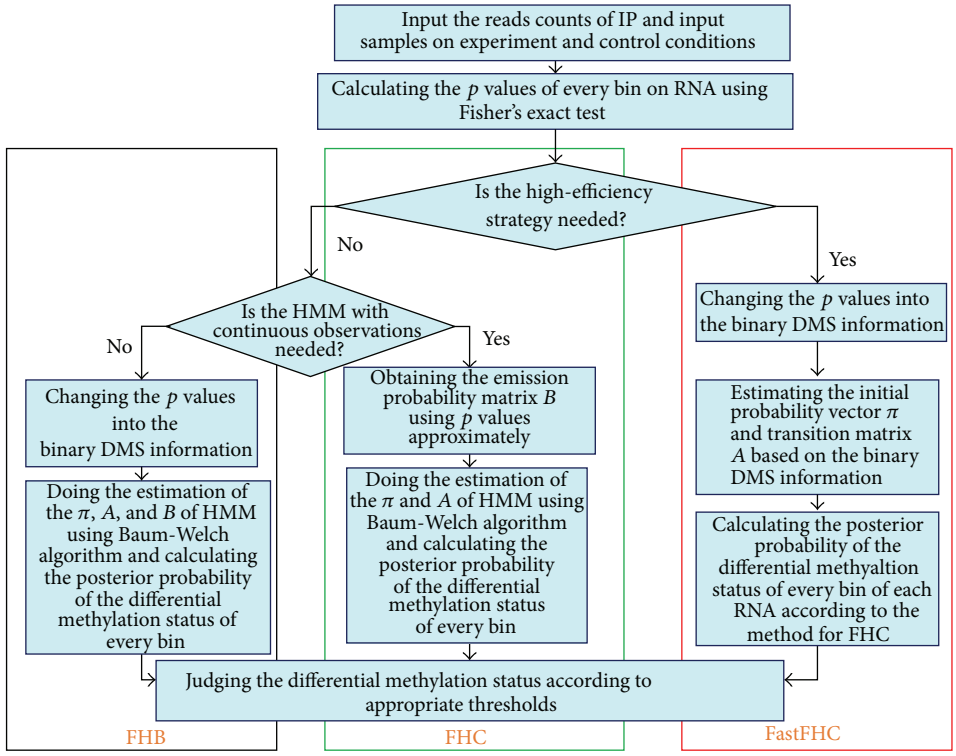


Figure 4: Comparison of different strategies. FHB strategy is the most naïve and straightforward; FHC is the most time consuming and performs better than FHB but is less robust. With FastFHC, the algorithm can now be applied to genome scale dataset in a timely and robust manner.

the performance of the proposed FET-HMM approach. The transition matrix A_{sim} was set as

$$A_{\text{sim}} = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \tag{17}$$

unless otherwise stated, and the initial probability $\pi = (0.5, 0.5)$ due to the lack of prior information. We considered three factors that may affect the performance of the algorithm, that is, the cut-off threshold applied to FET result for switching FDR (or p values) to the binary observed state (only for FHB), the transition matrix (degree of spatial dependency) used to generate the ground truth, and the sequencing depth (library size) of the data. The area under receiver operating characteristics curve (AUC) is calculated to evaluate the performance of the proposed algorithms under different settings of the 3 key factors to be tested.

In the first experiment, we tested the impact of cut-off threshold on the FHB strategy. As shown in Figure 5, although the choice of threshold does affect the performance of the algorithm, by incorporating spatial dependency, the proposed FHB strategy effectively improves the DMRs detection performance under all cut-off thresholds tested.

In the second experiment, we tested the impact of transition matrix, which indicates the degree of dependency between adjacent observations (bins). As shown in Figure 6, the performance of FHB and FHC strategies heavily relies

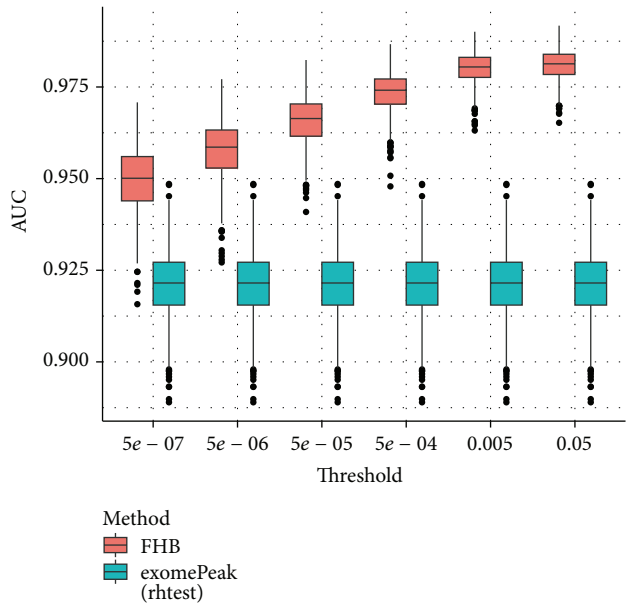


Figure 5: Boxplot of AUCs for different thresholds applied to switch FDR to the binary state. This figure shows that with the variation of thresholds, the performance of FHB outperforms exomePeak in AUC on 100 datasets. exomePeak does not use the cut-off threshold so its performance remains the same. The performance is evaluated at bin level rather than peak level in all experiments.

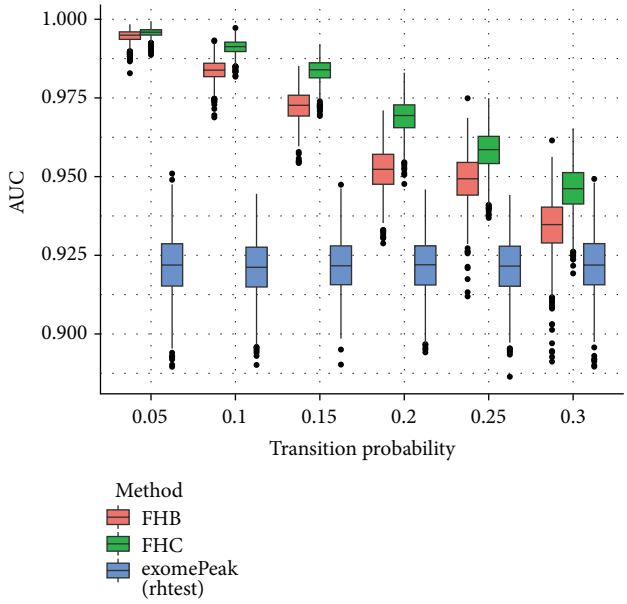


Figure 6: Boxplot of AUCs for different transition matrices used to generate the ground truth. The performance of FHB and FHC strategies heavily relies on the transition matrix setting, which reflects the degree of dependence between adjacent bins; and FHC strategy outperforms FHB and exomePeak under different settings tested.

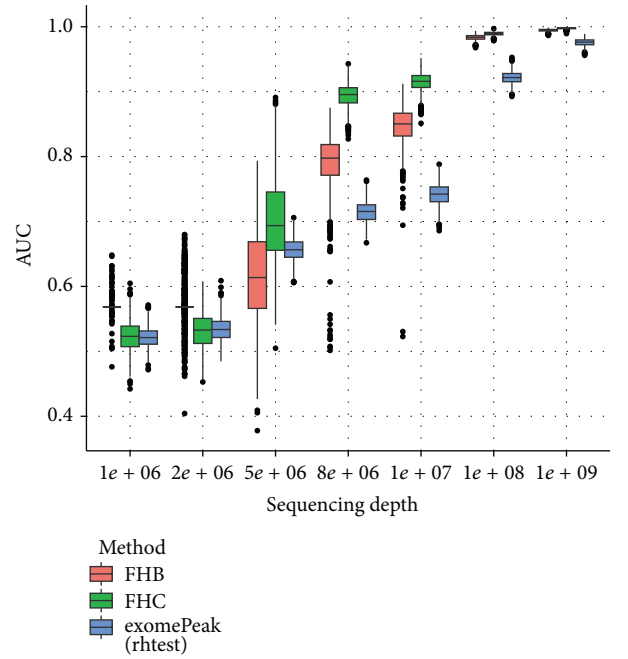


Figure 7: Boxplot of AUCs for different sequencing depths. The performance of all 3 approaches decreases together with the sequencing depth. FHC strategy gives the best performance and the advantage of FET-HMM over exomePeak is the most prominent when the data is of mediocre sequencing depth.

on the transition matrix setting, which reflects the degree of dependence between adjacent bins; and FHC strategy outperforms FHB and exomePeak under different settings tested.

The last factor that may affect the simulation results is the sequencing depth (the total number of reads). In our simulation, the sequencing depths (SD) of the four samples varied from 10^9 to 10^6. From Figure 7, we can see that the performances of FHB, FHC, and exomePeak are all satisfactory when sequencing depth is high enough (SD = 10^9); their performance all decreases together with the sequencing depth. Among the 3 methods tested, FHC gives the best performance and the advantage of FET-HMM over exomePeak is the most prominent when the sequencing depth is low. When the sequencing depth is very low, none of the 3 approaches can identify DMRs effectively.

We also consider here another scenario of unbalanced sequencing depth; that is, only one of the 4 samples has very large or small sequencing depth, and the results are highly consistent with previous result. As shown in Figure 8, the performance of all 3 approaches decreases as the sequencing depth decreases and FHC strategy outperforms FHB and exomePeak on most settings.

In general, the computational complexity of the proposed approaches increases together with the number of the genes, the length of the genes, and the resolution of the analysis (the size of the bin); and since FHB and FHC require iterative refinement, their computational complexity is also proportional to the number of iterations required to research convergence. To further evaluate the computational complexity of the 3 strategies, we conducted one additional

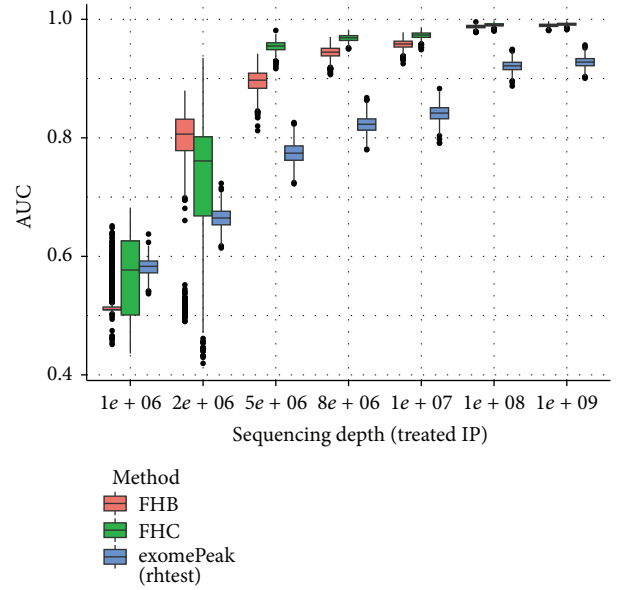


Figure 8: Boxplot of AUCs for different unbalanced sequencing depths. The performance of all 3 approaches decreases as the sequencing depth decreases and FHC strategy outperforms FHB and exomePeak on most settings. In this test, the sequencing depth of IP sample under treated condition varies with that of the other 3 samples unchanged.

Table 1: Comparison of different approaches.

Method	AUC	Time
FHB	0.960	4.39 s
FHC	0.987	0.85 s
FastFHC	0.962	0.12 s
exomePeak (rhtest)	0.924	0.02 s

Table 2: MeRIP-Seq data used.

Dataset	Cell	Treatment	Replicates (IP/input)	Reference
1	Hela	Control	4 & 4	[40]
2	Hela	METTL3 K/O	2 & 2	[40]
3	Hela	METTL14 K/O	2 & 2	[40]

experiment. In this experiment, we simulated a dataset of 7 genes, each with a different length (50, 100, 150, 200, 250, 300, and 350) and the methylation state transition probability is set to be 0.95. A total of 10 datasets are generated for evaluation purposes and the average performance and time consumption are calculated. As it can be seen from Table 1, on the simulated setting, FastFHC is comparable to FHB and FHC in performance, but much faster, making it a reasonable choice for genome-scale data with more than a few thousands of genes.

3.2. Test on MeRIP-Seq Data. In order to test our proposed method in real applications, we chose the human MeRIP-Seq data from Hela cells and from METTL3/METTL14 knockout conditions [40] as shown in Table 2. Previous study shows that METTL3 and METTL14 are components of RNA methyltransferase complex [40, 41], and we would like to identify their respective targeted RNA methylation sites from the following analysis. The original raw data in SRA format was downloaded directly from Gene Expression Omnibus (GEO) GSE46705, which consists of 8 IP and 8 Input MeRIP-Seq replicates obtained under wild type condition and after METTL3 or METTL14 knockout, respectively (a total of 16 libraries). The short sequencing reads are firstly aligned to human genome assembly hg19 with Tophat2 [42], and then the same types of samples obtained under the same condition are merged together for differential RNA methylation analysis.

Differential RNA methylation is predicted using exomePeak R/Bioconductor package [21] with UCSC gene annotation database [43] and with FastFHC strategy for comparison. Since METTL3 and METTL14 are methyltransferase, their target sites should exhibit hypomethylation under knockout condition. The hypomethylation sites under knockout condition (targeted RNA methylation sites) are then extracted and their sequences are submitted to MEME-ChIP for motif discovery. The identified motifs are summarized in Table 3. The enriched motifs are quite different in both datasets, indicating that there are multiple regulatory avenues to regulate the RNA methylome through sequence specificity.

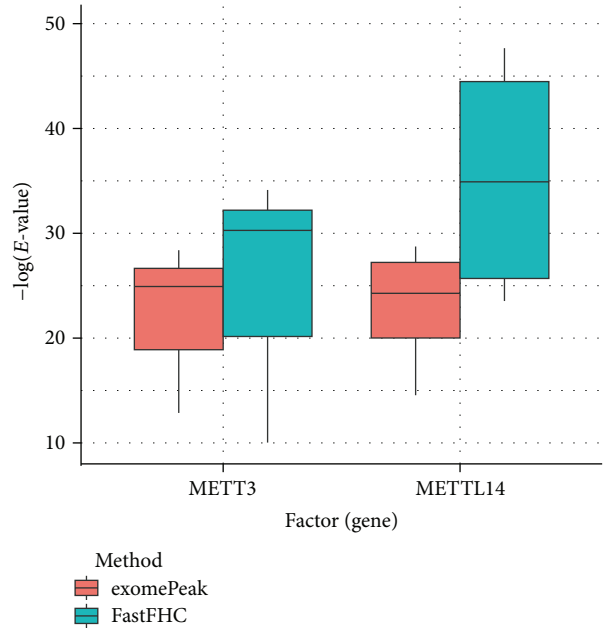


Figure 9: *E* values of motifs identified from differential methylation regions. The figure shows the motif *E* values from exomePeak and FastFHC strategy. With spatially enhanced differential methylation analysis, FastFHC identifies RNA methylation sites that are more biologically meaningful, indicating higher specificity compared with the exomePeak result.

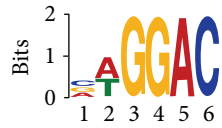
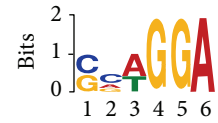
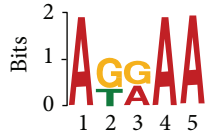
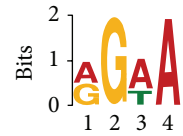
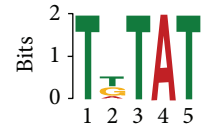
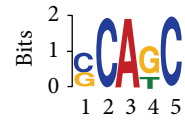
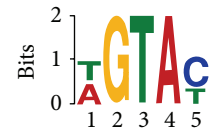
Despite the difference in sequences, as shown in Figure 9, the motifs identified by FastFHC results are more statistically significant than that from exomePeak, indicating higher sequence specificity, which is achieved by spatial enhancement with HMM in FET-HMM approach. The increased sequence specificity will be invaluable for decoding the structure of RNA methylation/demethylation enzymes.

We then checked the distribution of METTL3 and METTL14 targeted RNA methylation sites on mRNA and lncRNA. As shown in Figure 10, the targeted RNA methylation sites of METTL3 and METTL14 are relatively enriched near stop codon of mRNA. Interestingly, compared with METTL14 targets, METTL3 targets are relatively enriched on untranslated regions (5′ and 3′UTR), which is never reported before. Although existing studies suggest METTL3 and METTL14 function as an RNA methylation complex together with WTAP, our observation suggests that they may have their own respective functions as well. On lncRNA, their targets are almost uniformly distributed on the entire RNA with slight enrichment on 5′ end, whose reason is not yet clear.

4. Conclusion

In this paper, we developed an HMM-based method, FET-HMM, for spatially enhanced detection of differentially methylated region from MeRIP-Seq data. Compared with existing peak-based approaches which perform differential analysis on the entire methylation site, FET-HMM seeks to increase the resolution of detection to some extent by

Table 3: Motifs for target sites of METTL3 and METTL14.

	Rank	exomePeak		FET-HMM	
		Motif	*E*-value	Motif	*E*-value
METTL3 K/O	1		2.3×10^{-27}		2.4×10^{-33}
	2		4.7×10^{-13}		7.1×10^{-24}
	3		1.5×10^{-11}		1.5×10^{-15}
	4		2.6×10^{-6}		4.4×10^{-5}
METTL14 K/O	1		3.3×10^{-13}		1.4×10^{-19}
	2		2.5×10^{-12}		2.0×10^{-21}
	3		3.3×10^{-10}		3.4×10^{-12}
	4		4.8×10^{-7}		6.0×10^{-11}

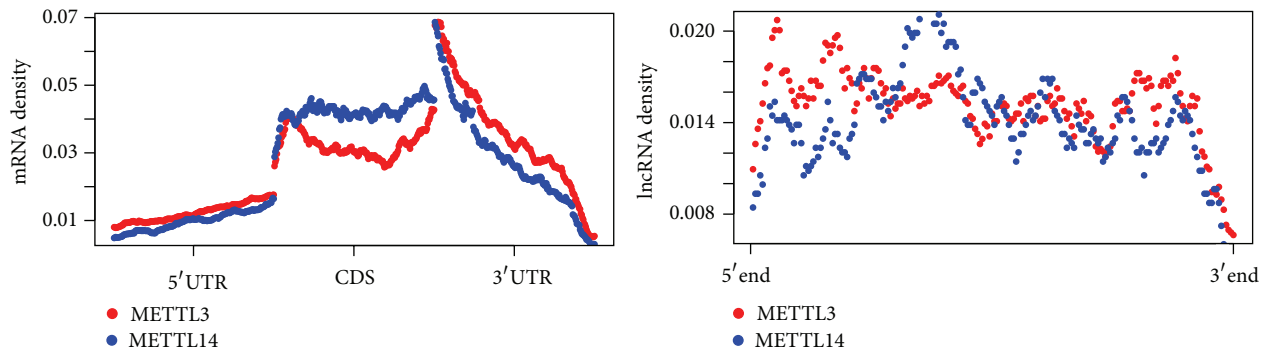


Figure 10: Distribution of METTL3 and METTL14 targeted RNA methylation sites. For both METTL3 and METTL14, their targeted RNA methylation sites are relatively enriched near stop codon of mRNA; however, compared with METTL14 targets, METTL3 targets are relatively enriched on untranslated regions (5′ and 3′UTR). On lncRNA, their targets are unfirmly distributed with slightly enriched on 5′ end.

dividing the single RNA methylation site into multiple adjacent bins (as shown in Figure 1), resulting in the improved detection performance. We developed 3 different strategies for this purpose, each with different advantage and disadvantages, and the FastFHC strategy can be directly applied to genome scale dataset. We show on the simulated and real datasets that the proposed approaches outperform original approach in detection performance and report more statistically significant DMRs on real MeRIP-Seq data.

It is important to note that exomePeak, which adopts a hypothesis testing scheme, relies on a cut-off threshold to report differential methylation sites, while FET-HMM, which assumes a hidden Markov model, needs a cut-off threshold for posterior probability. Although their performances can be compared under AUC, the two approaches are fundamentally different. It is suggested that both exomePeak and FET-HMM are used when analyzing specific datasets rather than using one approach only.

The proposed approach still has a number of limitations, many of which are shared by other existing MeRIP-Seq data analysis software. Firstly, the proposed approach could not model the within-group variation and thus cannot effectively take advantage of biological replicates. Currently, replicates are merged together which loses the biological variability. Secondly, the proposed approach cannot discriminate different isoforms of the same genes. MeRIP-Seq intrinsically poses very limited information regarding the methylation states of different isoform transcripts. Thirdly, even with the proposed approach, the spatial resolution is still not base-pair resolution. To obtain true base-pair solution, a more advanced computational approach needs to be developed to further combine the nucleotide sequence information (motif).

Disclosure

The open source R package implementing the proposed algorithm on MeRIP-Seq data is freely available from GitHub: https://github.com/lzcyzm/RHHMM.

Conflict of Interests

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

Acknowledgments

The authors thank the support from National Natural Science Foundation of China (61473232, 61401370, 91430111, 61170134, and 61201408) to Shao-Wu Zhang, Jia Meng, and Hui Liu; Jiangsu Science and Technology Program (BK20140403) to Jia Meng; Fundamental Research Funds for the Central Universities (2014QNB47, 2014QNA84) to Lin Zhang and Hui Liu. The authors also thank computational support from the UTSA Computational System Biology Core, funded by the National Institute on Minority Health and Health Disparities (G12MD007591) from the National Institutes of Health.

References

[1] G. G. Brownlee, F. Sanger, and B. G. Barrell, "Nucleotide sequence of 5S-ribosomal RNA from *Escherichia coli*," *Nature*, vol. 215, no. 5102, pp. 735–736, 1967.

[2] Y. Fu, D. Dominissini, G. Rechavi, and C. He, "Gene expression regulation mediated through reversible m^6A RNA methylation," *Nature Reviews Genetics*, vol. 15, no. 5, pp. 293–306, 2014.

[3] K. D. Meyer and S. R. Jaffrey, "The dynamic epitranscriptome: N6-methyladenosine and gene expression control," *Nature Reviews Molecular Cell Biology*, vol. 15, no. 5, pp. 313–326, 2014.

[4] J. König, K. Zarnack, N. M. Luscombe, and J. Ule, "Protein-RNA interactions: new genomic technologies and perspectives," *Nature Reviews Genetics*, vol. 13, no. 2, pp. 77–83, 2012.

[5] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz et al., "Topology of the human and mouse m^6A RNA methylomes revealed by m^6A-seq," *Nature*, vol. 484, no. 7397, pp. 201–206, 2012.

[6] K. D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C. E. Mason, and S. R. Jaffrey, "Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons," *Cell*, vol. 149, no. 7, pp. 1635–1646, 2012.

[7] D. Dominissini, S. Moshitch-Moshkovitz, M. Salmon-Divon, N. Amariglio, and G. Rechavi, "Transcriptome-wide mapping of N^6-methyladenosine by m^6A-seq based on immunocapturing and massively parallel sequencing," *Nature Protocols*, vol. 8, no. 1, pp. 176–189, 2013.

[8] J. M. Fustin, M. Doi, Y. Yamaguchi et al., "XRNA-methylation-dependent RNA processing controls the speed of the circadian clock," *Cell*, vol. 155, no. 4, pp. 793–806, 2013.

[9] M. E. Hess, S. Hess, K. D. Meyer et al., "The fat mass and obesity associated gene (Fto) regulates activity of the dopaminergic midbrain circuitry," *Nature Neuroscience*, vol. 16, no. 8, pp. 1042–1048, 2013.

[10] Y. Wang, Y. Li, J. I. Toth, M. D. Petroski, Z. Zhang, and J. C. Zhao, "N6 -methyladenosine modification destabilizes developmental regulators in embryonic stem cells," *Nature Cell Biology*, vol. 16, no. 2, pp. 191–198, 2014.

[11] M. Lee, B. Kim, and V. N. Kim, "Emerging roles of RNA modification: m(6)A and U-tail," *Cell*, vol. 158, no. 5, pp. 980–987, 2014.

[12] S. Schwartz, M. R. Mumbach, M. Jovanovic et al., "Perturbation of m6A writers reveals two distinct classes of mRNA methylation at internal and 5′ sites," *Cell Reports*, vol. 8, no. 1, pp. 284–296, 2014.

[13] J. Liu, Y. Yue, D. Han et al., "A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation," *Nature chemical biology*, vol. 10, no. 2, pp. 93–95, 2014.

[14] X. Wang, Z. Lu, A. Gomez et al., "N 6-methyladenosine-dependent regulation of messenger RNA stability," *Nature*, vol. 505, no. 7481, pp. 117–120, 2014.

[15] C. He, "Grand Challenge Commentary: RNA epigenetics?" *Nature Chemical Biology*, vol. 6, no. 12, pp. 863–865, 2010.

[16] D. Dominissini, "Roadmap to the epitranscriptome," *Science*, vol. 346, no. 6214, pp. 1192–1192, 2014.

[17] H. Liu, M. A. Flores, J. Meng et al., "MeT-DB: a database of transcriptome methylation in mammalian cells," *Nucleic Acids Research*, vol. 43, no. 1, pp. D197–D203, 2015.

[18] L. Liu, S. Zhang, Y. Zhang et al., "Decomposition of RNA methylome reveals co-methylation patterns induced by latent enzymatic regulators of the epitranscriptome," *Molecular BioSystems*, vol. 11, no. 1, pp. 262–274, 2015.

[19] J. Meng, X. Cui, M. K. Rao, Y. Chen, and Y. Huang, "Exome-based analysis for RNA epigenome sequencing data," *Bioinformatics*, vol. 29, no. 12, pp. 1565–1567, 2013.

[20] Y. Li, S. Song, C. Li, and J. Yu, "MeRIP-PF: an Easy-to-use Pipeline for High-resolution Peak-finding in MeRIP-Seq Data," *Genomics, Proteomics & Bioinformatics*, vol. 11, no. 1, pp. 72–75, 2013.

[21] J. Meng, Z. Lu, H. Liu et al., "A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package," *Methods*, vol. 69, pp. 274–281, 2014.

[22] C. Bock, "Analysing and interpreting DNA methylation data," *Nature Reviews Genetics*, vol. 13, no. 10, pp. 705–719, 2012.

[23] M. D. Robinson, A. Kahraman, C. W. Law et al., "Statistical methods for detecting differentially methylated loci and regions," *Frontiers in Genetics*, vol. 5, article 324, 2014.

[24] D. Sun, Y. Xi, B. Rodriguez et al., "MOABS: model based analysis of bisulfite sequencing data," *Genome Biology*, vol. 15, article R38, 2014.

[25] P. A. Stockwell, A. Chatterjee, E. J. Rodger, and I. M. Morison, "DMAP: differential methylation analysis package for RRBS and WGBS data," *Bioinformatics*, vol. 30, no. 13, pp. 1814–1822, 2014.

[26] J. Meng, X. Cui, H. Liu et al., "Unveiling the dynamics in RNA epigenetic regulations," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '13)*, pp. 139–144, Shanghai, China, December 2013.

[27] Y. Zhang, T. Liu, C. A. Meyer et al., "Model-based analysis of ChIP-Seq (MACS)," *Genome Biology*, vol. 9, article R137, 2008.

[28] M. Seifert, S. Cortijo, M. Colomé-Tatché, F. Johannes, F. Roudier, and V. Colot, "MeDIP-HMM: genome-wide identification of distinct DNA methylation states from high-density tiling arrays," *Bioinformatics*, vol. 28, no. 22, pp. 2930–2939, 2012.

[29] M. Seifert, J. Keilwagen, M. Strickert, and I. Grosse, "Utilizing gene pair orientations for HMM-based analysis of promoter array ChIP-chip data," *Bioinformatics*, vol. 25, no. 16, pp. 2118–2125, 2009.

[30] H. Xu, C.-L. Wei, F. Lin, and W.-K. Sung, "An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data," *Bioinformatics*, vol. 24, no. 20, pp. 2344–2349, 2008.

[31] K. Krishnamoorthy and J. Thomson, "A more powerful test for comparing two Poisson means," *Journal of Statistical Planning and Inference*, vol. 119, no. 1, pp. 23–35, 2004.

[32] J. Przyborowski and H. Wilenski, "Homogeneity of results in testing samples from Poisson series with an application to testing clover seed for dodder," *Biometrika*, vol. 31, pp. 313–323, 1940.

[33] C. Becker, J. Hagmann, J. Müller et al., "Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome," *Nature*, vol. 480, no. 7376, pp. 245–249, 2011.

[34] Y. Li, J. Zhu, G. Tian et al., "The DNA methylome of human peripheral blood mononuclear cells," *PLoS Biology*, vol. 8, no. 11, Article ID e1000533, 2010.

[35] 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.

[36] R. Lister, M. Pelizzola, Y. S. Kida et al., "Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells," *Nature*, vol. 471, no. 7336, pp. 68–73, 2011.

[37] P. Baldi and Y. Chauvin, "Smooth on-line learning algorithms for hidden Markov models," *Neural Computation*, vol. 6, no. 2, pp. 307–318, 1994.

[38] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[39] A. Poritz, "Hidden Markov models: a guided tour," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '88)*, vol. 1, pp. 7–13, New York, NY, USA, April 1988.

[40] X. Wang, Z. Lu, A. Gomez et al., "N^6-methyladenosine-dependent regulation of messenger RNA stability," *Nature*, vol. 505, no. 7481, pp. 117–120, 2014.

[41] X.-L. Ping, B.-F. Sun, L. Wang et al., "Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase," *Cell Research*, vol. 24, no. 2, pp. 177–189, 2014.

[42] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biology*, vol. 14, article R36, 2013.

[43] D. Karolchik, G. P. Barber, J. Casper et al., "The UCSC genome browser database: 2014 update," *Nucleic Acids Research*, vol. 42, no. 1, pp. D764–D770, 2014.