This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3027392, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Prediction of Essential Genes in Comparison States Using Machine Learning

Jiang Xie, Chang Zhao, Jiamin Sun, Jiaxin Li, Fuzhang Yang, Jiao Wang* and Qing Nie*

Abstract— Identifying essential genes in comparison states (EGS) is vital to understanding cell differentiation, performing drug discovery, and identifying disease causes. Here, we present a machine learning method termed Prediction of Essential Genes in Comparison States (PreEGS). To capture the alteration of the network in comparison states, PreEGS extracts topological and gene expression features of each gene in a five-dimensional vector. PreEGS also recruits a positive sample expansion method to address the problem of unbalanced positive and negative samples, which is often encountered in practical applications. Different classifiers are applied to the simulated datasets, and the PreEGS based on the random forests model (PreEGSRF) was chosen for optimal performance. PreEGSRF was then compared with six other methods, including three machine learning methods, to predict EGS in a specific state. On real datasets with four gene regulatory networks, PreEGSRF predicted five essential genes related to leukemia and five enriched KEGG pathways. Four of the predicted essential genes and all predicted pathways were consistent with previous studies and highly correlated with leukemia. With high prediction accuracy and generalization ability, PreEGSRF is broadly applicable for the discovery of disease-causing genes, driver genes for cell fate decisions, and complex biomarkers of biological systems.

Index Terms—Differential Network Analysis, Essential Genes in Comparison States, Machine Learning, Biomarker Discovery

1 Introduction

Essential genes account for a small fraction of genes in a genome; however, they are vital to the survival or development of organisms [1], [2]. They also play significant roles in cell state transformation during complex disease progression as well as in cell development and differentiation.

Essential genes in comparison states (EGS) are indispensable to the transformation of the two states, which can prevent, improve or adjust organisms to go from one state to another. Identifying EGS is significant not only for exploring the factors influential the survival and development of living organisms but also for finding pathogenic genes and potential drug targets for curing diseases [3]. For example, in terms of finding pathogenic genes, researchers identified the genes modules that were possibly responsible for STAT-mediated antiviral responses through gene co-expression networks for shrimp [4]. In terms of finding drug targets, the DrPOCS method was used to predict potential associations between drugs and diseases with matrix completion based on EGS [5]. Currently, there are three main types of experimental strategies for the genome-wide discovery of essential genes: gene knockout [6], gene knockdown [7] and transposon mutagenesis [8]. Although these experimental methods can accurately discover essential genes, they are expensive, time consuming and laborious. Importantly, these experiments cannot be performed in living organisms, especially in human beings. For complex organisms and diseases, the prediction performance is far from perfect.

Therefore, predicting EGS through computational methods may be a valuable tool for improved performance. Many computational methods and tools have been previously developed to identify the genes undergoing significant changes across biological comparison states. For example, the CSTEA webserver organized, analyzed and visualized the time-course gene expression data and essential genes during cell differentiation [9]; and the HISP hybrid intelligent method was used to determine the optimal topologies of signaling pathways in an accurate way to unveil a high-resolution signaling pathway [10]. These methods and tools enabled both experimental and computational biologists to better understand the mechanisms of cell fate determination.

In order to find EGS, various gene expression analysis methods have been developed, such as Student's t-test and the significance analysis of microarrays (SAM) [11]. However, because genes are strongly intertwined with each other in multiple pathways, those methods are strongly biased because gene interaction information is neglected. To address the potential bias introduced by conventional gene expression analysis methods, network-based methods were developed to provide better ways of exploring multiple interactions among genes. One popular approach, differential network analysis, has been frequently used to

- *J. Xie, C. Zhao, J.M. Sun, J.X. Li and F.Z. Yang are with the School of Computer Engineering and Science, Shanghai University, Nanchen Road 333, Shanghai, 200444, China. E-mail: jiangx@shu.edu.cn, changZ@shu.edu.cn, jiaminsun@i.shu.edu.cn, jiaxinlee@i.shu.edu.cn, crevenyoung@shu.edu.cn*
- *J. Wang is with the Laboratory of Molecular Neural Biology, School of Life Sciences, Shanghai University, Nanchen Road 333, Shanghai, 200444, China. E-mail: Jo717@shu.edu.cn*
- *Q. Nie is with the Department of Mathematics, the Center for Mathematical and Computational Biology, and the Center for Complex Biological Systems, University of California-Irvine, Irvine, CA 92697, USA. E-mail: qnie@math.uci.edu*

*Please note that all acknowledgments should be placed at the end of the paper, before the bibliography (**note that corresponding authorship is not noted in affiliation box, but in acknowledgment section**).*

1545-5963 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
xxx-xxx/xx/$xx.00 © 200x IEEE Published by the IEEE Computer Society
Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 13,2020 at 07:01:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3027392, IEEE/ACM Transactions on Computational Biology and Bioinformatics

measure the changes in networks in comparison states [12],[13]. Such an approach could also be utilized to identify essential biomolecules between two states. Many structure/function analyses of protein sequences have been performed to improve the prediction accuracy in different states [14], [15], [16]. Diffk [17] used differential degree centrality (DDC) [12] to score all nodes in networks, and the essential genes were then obtained by ranking the scores. Another method, DiffRank [18], ranked genes based on their contributions to the differences between two networks by using two new structural scoring measures. In DCloc [19], differential correlation patterns were identified by comparing the local or global topology of correlation networks.

The methods mentioned above have contributed to identifying EGS through network analyses. However, the methods based on statistics or graph theory were insufficient to discover the potential rules and could not be generalized across biological processes.

Recently, machine learning combined with bionetwork analysis has been found to be a powerful approach in classification problems, including tumors, protein sequences and essential biomolecules, and has already been proven useful in predicting essential biomolecules. Support vector machines and neural networks on protein-protein networks have been used to estimate whether each protein in the network is essential or inessential [20]. The NC method [21] found essential proteins by considering the modular nature of protein essentiality through calculating the edge clustering coefficient. The support vector machine (SVM) classifier has been utilized to classify the protein sequences, protein-protein interactions and hot spot residues in protein interfaces [22],[23],[24]. D.S. Huang combined ICA and regularized regression models to classify tumors by gene expression data [25]. HED is a machine learning method used to predict potential associations between drugs and diseases based on a drug-disease heterogeneous network [26]. PREvaIL [27] is an integrative machine learning approach that combines sequence, structural, and network features to determine the catalytic residues in an enzyme. A method based on extreme gradient boosting in a specific network (XGBoost_EG) [28] can constructively project essential genes by integrating homology, gene-intrinsic characteristics, and network topology features.

The current machine learning methods focus on finding the essential genes in specific networks in a static way. The critical network features associated with the changes between networks in different states, which are likely strongly associated with alterations in biological functions or pathways, are not considered. Here, we present a novel method termed Prediction of Essential Genes in Comparison States (PreEGS), which is used to extract the differential information of each gene between two networks. The method considers not only topological structure but also gene expression, and we transform this information into vectors to construct the learning model. Then, the well-trained model is used to find the essential genes that cause the differences between two networks of comparison states.

2 METHODS

PreEGS is based on the machine learning model and aims to predict the essential genes in differential networks. A novel feature extraction method is built using both network topology and gene expression. Moreover, PreEGS introduces a sample expansion method to address the problem of unbalanced positive and negative samples.

2.1 Dataset Sources

2.1.1 Training and Test Dataset Sources

Both the training and test datasets include pairs of contrasted biomolecular networks along with the gene expression data for each gene. Each network is constructed by the gene expression in the specific state, and a pair of contrasted networks indicate two comparison states (e.g., the pathological disease state and the control state). Here, the expression data of each gene were collected from various databases, such as the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih. gov/geo/).

2.1.2 Marking the Training Data

Every piece of data should be marked in the training dataset. This is a binary classification problem. Each gene is regarded as essential (marked as 1) or unessential (marked as 0). Currently, the essential biomolecules in specific states can be integrated from databases, such as DEG [29]. However, there are few databases that contain marked essential genes in comparison states. Therefore, the genes supported by literature reports are defined as the essential genes (marked as 1) in this paper.

2.2 Generating 5-Dimensional Feature Vector

2.2.1 Topological Features of Nodes in Networks

It is essential to describe each node in the network using a feature vector. While many types of network topological features have been developed [30], the most-frequently used features [31] for the machine learning-based prediction of essential biomolecules are degree centrality (DC) [32], betweenness centrality (BC) [33], closeness centrality (CC) [34] and clustering coefficient (CCo) [35].

1) DC: DC indicates how many nodes are connected to one node, and it can measure one node's apparent 'centrality'. It is formalized by (1):

$$C_D(v) = deg(v) \tag{1}$$

2) BC: For a node v, BC is defined as the average length of the shortest path through it. If more shortest paths pass through v, then v has a higher centrality. BC's equation is as follows:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2}$$

where σ_{st} means the number of shortest paths through node s and node t, and $\sigma_{st}(v)$ means the number of paths that go through node v.

3) CC: CC indicates the degree of a node v that communicates with other nodes in the network with n nodes. It is calculated by the sum of the shortest distance between v and the others:

1545-5963 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 13,2020 at 07:01:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3027392, IEEE/ACM Transactions on Computational Biology and Bioinformatics

$$C_c(v) = \sum_{j=1}^{n-1} x_{vj} \ (j \neq v) \tag{3}$$

where x_{vj} is the shortest path form node v to node j . CC relates to the scale of the network. A larger network may lead to a higher CC. To eliminate the impact of network size on CC, (3) is converted into (4):

$$C_c'(v) = \frac{n-1}{C_c(v)} \tag{4}$$

4) CCo: CCo is a coefficient indicating the level of aggregation around a node in the network.

$$C_{CCo}(v) = \frac{m}{C_k^2} = \frac{2m}{k(k-1)} \tag{5}$$

As shown above, k represents the total number of neighbors of node v, and m represents the number of edges that link all adjacent nodes of node v.

2.2.2 Feature Extraction of PreEGS

A novel feature extraction method was proposed to illustrate the comprehensive features of differential networks, including topological structure and gene expression. A five-dimensional vector $x_v = \{x_v^1, x_v^2, x_v^3, x_v^4, x_v^5\}$ was constructed for node v to quantify the differences in v in comparison states.

1) $x_v^1 = \Delta DC$. The degree change in node v is the most intuitive way to describe the differences in v in comparison states. The equation of DC is shown in (1); therefore, the calculation method of x_v^1 is as follows:

$$x_v^1 = |C_D(v) - C_D'(v)| \tag{6}$$

where $C_D(v)$ and $C_D'(v)$ denote the degree of node v in the two states.

2) $x_v^2 = \Delta BC$. As shown in (2), BC means a kind of 'centrality' of one node. The more shortest-paths pass through a node, the more critical the node might be in the network. Therefore, ΔBC can reflect the change in centrality in comparison states.

$$x_v^2 = |C_B(v) - C_B'(v)| \tag{7}$$

$C_B(v)$ and $C_B'(v)$ denote BC of v in comparison states.

3) $x_v^3 = \Delta CC$. CC is shown in (4). ΔCC shows the difference in closeness centrality in comparison states. x_v^3 is calculated as follows:

$$x_v^3 = |C_C'(v) - C_C''(v)| \tag{8}$$

$C_C'(v)$ and $C_C''(v)$ denote the CC of v in comparison states.

4) $x_v^4 = \Delta CCo$. A node v with high CCo has high power to influence the network. ΔCCo can reflect the change in the power of v.

$$x_v^4 = |C_{CCO}(v) - C_{CCo}'(v)| \tag{9}$$

5) $x_v^5 = p - value(E, E')$. E and E' represent the expression values of each gene in two comparison states. The differences between E and E' are direct representations of genes changes. x_v^5 represents the statistical significance of the difference in gene expression.

2.3 Expanding Positive Samples

Usually, the essential genes marked by the known literature studies are not abundant, causing positive samples and negative samples to be out of balance. Under such circumstance, the predictions of the model have no significance and the algorithm lacks extensiveness and predictability [36].

The genes supported by literature studies are essential, although genes that have not been reported in existing work may also be essential. Therefore, we extend the positive samples to address the problem of unbalanced data. In this paper, a positive sample extension method is presented by using the Pearson correlation coefficient (PCC) and setting up a threshold ε.

$$PCC(x,y) = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^2 - (\sum x_i)^2}\sqrt{N\sum y_i^2 - (\sum y_i)^2}} \tag{10}$$

where N indicates the number of all genes. x_i, y_i indicate the five-dimensional vectors of two genes (x, y). The value of i is from 1 to 5 in this paper.

All genes are divided into the essential gene set $R = \{r_0, r_1, r_2, \dots, r_n\}$ supported by the literature research and the unmarked gene set $U = \{u_0, u_1, u_2, \dots, u_m\}$, where $n \ll m$. After calculating all $PCC(r_x, u_y)$ between $r_x \in R$ and $u_y \in U$, u_y is marked as an essential gene if $PCC(r_x, u_y)$ is larger than the threshold ε.

The principle for setting ε is that the number of positive and negative samples should be as balanced as possible after extension. The greater the absolute value of the PCC is, the stronger the correlation is. Generally, a PCC over 0.8 means that there is a strong correlation between the two genes [37], indicating that the unmarked genes are more likely to be essential genes. To ensure the biological sense, we set the threshold ε to no less than 0.8 in this paper.

2.4 Training the Model

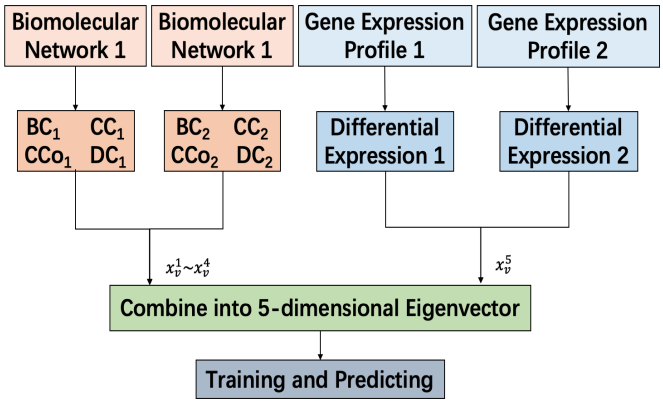

Fig. 1. Flow Chart of PreEGS.

The flow chart of PreEGS is shown in Figure. 1. The parameters in x_v^1, x_v^2, x_v^3 and x_v^4 are constructed by four topological structures in (6), (7), (8) and (9), and x_v^5 is constructed by differential expression from the gene expression profile.

The 5-dimension feature vector $x_v = \{x_v^1, x_v^2, x_v^3, x_v^4, x_v^5\}$ is built by both network topology and gene expression to train the model.

1545-5963 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 13,2020 at 07:01:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3027392, IEEE/ACM Transactions on Computational Biology and Bioinformatics

To validate the prediction, various well-known data mining algorithms, such as support vector machine (SVM), K-nearest neighbor (KNN), random forest (RF), Gaussian naive Bayes (GNB), and logistic regression (LG), were applied to the datasets. All the machine learning algorithms above are implemented by the scikit-learn package [38]. After model training, the EGS of the life process are predicted and can provide new inspiration for bioscience or medical science research.

2.5 Performance Evaluation

To evaluate the performance of a prediction model, true negative (TN), false positive (FP), false negative (FN) and true positive (TP) [39] are used to calculate evaluation indicators, including accuracy, precision, recall and F1-Score, as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Sensitivity = Recall = \frac{TP}{TP + FN} \quad (13)$$

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (14)$$

Moreover, the F1-Score is a useful indicator for measuring the accuracy of a binary classification model. The F1-Score considers precision and recall, which ranges from 0 to 1. The model is better if the F1-Score is closer to 1.

3 Results

In this section, PreEGS was first verified by the application of different classifiers to simulated datasets. Next, the classifier with the optimal PreEGS performance was compared with three classical methods (including DCloc, DiffRank and DEC [40]). Furthermore, to compare with machine learning methods, NC, FVM [31] and XGboost_EG are used to identify essential genes on the simulated datasets. Finally, we applied the proposed method to neurocytoma and leukemia datasets to identify the EGS and modules.

3.1 Simulated Datasets Sources

According to the scale-free property of biomolecular networks [41], the algorithm [42] would output a pair of networks along with the gene expression and a list of essential genes. In the algorithm, the parameters n_1 and n_2 represent the number of genes, and m represents the number of essential genes in the two networks. The parameter ρ is the proportion of differential edges driven by perturbed genes. The smaller ρ is, the more difficult it is to find EGS. We used this algorithm to generate 200 pairs of networks, each containing 100 nodes whose degree distribution followed a power-law distribution. These simulated networks, among which 100 pairs, named 100A_dataset, were used to train the PreEGS model, and the remaining 100 pairs, named the 100B_dataset, were used for testing. To simulate the EGS proportions of the real dataset and minimize the problem of unbalanced data [28], $n_1 = n_2 = 100$, $m = 30$, $\rho = 0.05$.

3.2 Performance Comparison on Simulated Datasets

The PreEGS method was tested in several different ways, and this section is organized as follows.

First, the PreEGS method was compared with different classifiers by 10-fold cross-validation on the 100A_dataset. We found that the PreEGS based on random forests (PreEGSRF) had the optimal performance.

Second, to demonstrate the superiority of PreEGSRF over classical methods, we compared the prediction performance with that of other methods, including DCloc, DiffRank and DEC on the 100B_dataset.

Third, PreEGSRF was compared with machine learning-based methods including NC, FVM and XGboost_EG on the 100A_dataset. The results indicated that PreEGSRF had higher performance than the state-of-the-art methods.

3.2.1 Performance Evaluation of the PreEGS Method for Different Classifiers

We adopted five machine learning methods for the EGS prediction task: SVM, KNN, RF, GNB and LG.

The *K*-fold cross-validation technique was used to verify the reliability of the experiments on the 100A_dataset. The cross-validation process was repeated *k* times, with each of the *k* subsamples used exactly once as the validation data. In this paper, the number of folds was set to be 10 for the approximation of error.

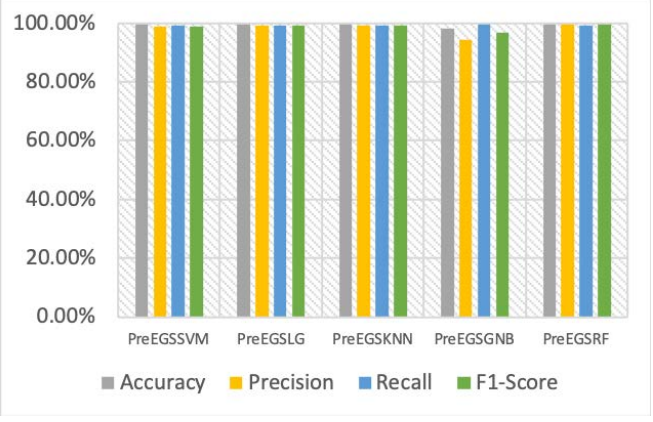

Fig. 2. Evaluation Indicators of PreEGS with Different Classifiers.

Figure 2 shows the evaluation indicators obtained by PreEGS on different original machine learning classifiers.

The results showed that the PreEGS method based on RF performed best among the different machine learning classifiers. In further experiments, PreEGSRF was the representative whose performance was compared with that of other methods.

3.2.2 Performance Evaluation of PreEGSRF with Classical Methods

In simulated experiments, PreEGSRF was compared with three other classical methods that could predict EGS: DCloc, DiffRank and DEC.

These three methods were based on the traditional numerical calculation method, which could score all nodes but not classify whether the gene was essential or not in the differential networks.

In this paper, 100B_dataset was calculated by the three methods. Each of the top 30 score genes was compared with 30 marked essential nodes, and the indicators of each method were the mean values of 100 experiments as shown in Figure 3.

1545-5963 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 13,2020 at 07:01:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3027392, IEEE/ACM Transactions on Computational Biology and Bioinformatics

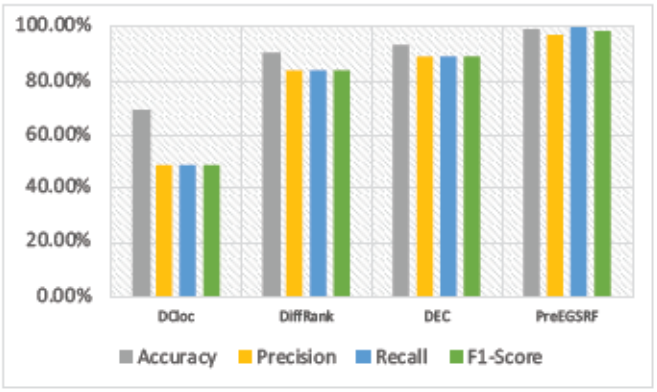

Fig. 3. Evaluation Indicators of Three Classical Methods Compared with PreEGS.

Based on the trained PreEGSRF model in section 3.2.1, the 100B_dataset was used for testing. Figure 3 lists the mean values of the 100 tests.

As shown in Figure 3, PreEGSRF is superior to the other three methods in terms of these evaluation indicators. In particular, the F1-Score of PreEGSRF is 0.982, which is closest to 1, thus proving that PreEGSRF has a good performance for essential node prediction.

The results showed that traditional numerical calculation methods lack generalization ability. In most cases, they had less ability to detect potential differential indicators. PreEGSRF learned multiple kinds of features and could take various kinds of differences into consideration to achieve a better prediction.

3.2.3 Performance Evaluation of PreEGSRF with Machine Learning Methods

EGS imply an underlying mechanism of transformation from one state to another. However, the prediction of essential genes based on machine learning is currently mostly focused on specific networks. The state-of-the-art methods included NC, FVM and XGBoost_EG. When working on networks in comparison states by 10-fold cross-validation on the 100A_dataset, these methods will respectively obtain two sets of essential genes corresponding to one of the specific states. To evaluate their performance on comparison states, the evaluation indicators are the average value of the two states. The evaluation indicators of PreEGSRF in Figure 2 were compared with these three methods in Figure 4.

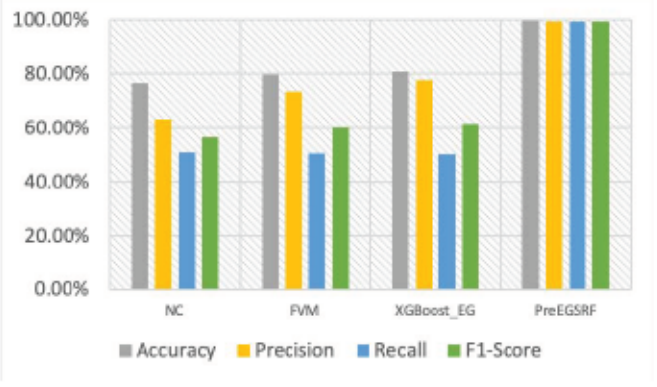

Fig. 4. Evaluation Indicators of Three Machine Learning Methods Compared with PreEGSRF.

As shown in Figure 4, PreEGSRF is superior to the other machine learning methods. It detects the differences between two comparison states beyond the features of a specific network so that the evaluation performance is improved for EGS prediction.

3.3 Application to Leukemia

3.3.1 Data Sources

In the real data experiments, two sets of networks were found in the Interactome dataset [43], [44] (http://www.regulatorynetworks.org/, January, 2019). The astrocyte gene regulatory network (NHA) and the neuroblastoma gene regulatory network (SKNSH) were used as the training set. The microvascular endothelium, adult, blood gene regulatory network (HMVEC_dBlAd) and promyelocytic leukemia gene regulatory network (NB4) were selected as the test set. Moreover, this paper referred to the GEO datasets for gene expression profile data. All the data sources are shown in Table 1.

TABLE 1
Training and Test Sets

	Network	Network-Scale	GEO ID
Training Set	NHA	516 genes 9296 edges	GSE99051 [45]
	SKNSH	508 genes 12761 edges	GSE112384 [46]
Test Set	HMVEC_dBlAd	520 genes 13510 edges	GSE12679 [47]
	NB4	525 genes 18960 edges	GSE73157 [48]

For the training set, 23 essential genes for neuroblastoma were supported by 14 literature studies [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62]: TP53, BRCA1, MYCN, E2F1, FOXA1, ZFX, PRDM1, BCL6, XBP1, ASCL1, TP73, ESR1, ZBTB33, PPARA, E2F2, BACH1, BACH2, PBX1, MEIS1, GATA3, HIF1A, ZNF148 and BPTF; these genes were marked as essential(1), while the remaining genes were marked tentatively as unessential(0).

3.3.2 Features Extraction and Sample Balancing

In the differential network analysis, significant differential changes in the common genes are the focus of concern. Thus, the common genes between two differential networks should be found first. There were 486 and 247 common genes in the two sets of networks, respectively. In the training set, the SKNSH network had 12149 edges, while the NHA network had 8649 edges among the common 486 genes. In the test set, the HMVEC_dBlAd network had 480 edges, while the NB4 network had 663 edges among the common 247 genes.

Next, using the PreEGSRF method, each common node in the training set was vectorized as 5-dimensional feature vectors $x_v = \{x_v^1, x_v^2, x_v^3, x_v^4, x_v^5\}$.

There were only 23 essential genes out of 486 genes. The positive samples and negative samples were out of balance. Therefore, the extension method was used to balance the two kinds of samples. After setting ε=0.9, the number of essential genes was 198 and the number of unmarked genes was 288 after extension.

3.3.3 Validation of EGS in Leukemia Datasets

1545-5963 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 13,2020 at 07:01:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3027392, IEEE/ACM Transactions on Computational Biology and Bioinformatics

To further evaluate the performance, PreEGSRF was compared with three other machine learning methods focused on specific states.

To avoid the deviation caused by randomness, the train-test procedure was repeated 100 times. Every predicted gene was selected as a candidate EGS.

When working on networks in comparison states on the leukemia datasets, three machine learning methods (NC, FVM and XGBoost_EG) would respectively obtain two sets of essential genes, each corresponding to one of the specific states. The predicted genes were the intersection of the two sets of essential genes.

In total, 30, 22, 38 and 25 genes were marked as candidate genes for leukemia by four methods (NC, FVM, XGBoost_EG and PreEGSRF, respectively). According to the predicted genes, the top five candidate genes were recognized as EGS to compare the performance in Table 2.

TABLE 2
The Top Five Candidate Genes Recognized by Four Methods.

Methods	Essential Genes
NC	GATA1[63], GATA2[64], PATZ1, SP1[65], SP3
FVM	MTF1, NR3C1[66], NFYA, GATA1[63], SP1[65]
XGBoost_EG	HMGA2, PATZ1, STAT6[67], GATA1[63], HES1[68]
PreEGSRF	HES1[68], STAT1[69], RFXANK, TAL1[70], SPI1[71]

In Table 2, the genes associated with the leukemia had been marked by the literature researches. For the three machine learning methods focused on specific states, the candidate essential genes were not enriched in any pathway of leukemia. However, these 25 candidate EGS predicted by PreEGSRF were enriched in the leukemia Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway.

The results show that PreEGSRF can detect more EGS and have better performance for predicting EGS.

For the PreEGSRF method, HES1, STAT1, RFXANK and TAL1 were marked as essential for 100 times, and SPI1 was marked for 97 times, while the others were marked less than 30 times. Afterward, these five most-frequently occurring genes were recognized as EGS and analyzed by literature research.

HES1: HES1 plays a critical role in the development of T cells and plays an essential role in the process of cancer. More importantly, HES1 can be treated as an essential drug target for leukemia [68].

STAT1: STAT1 is widely known as a suppressor during tumor growth. However, STAT1 can accelerate the process of leukemia [69].

TAL1: In the TAL1 gene, site-specific DNA recombination will occur in patient of leukemia. Therefore, the TAL1 gene is closely interrelated with leukemia [70].

SPI1: Friend viruses, such as the Rauscher virus, are a cause of leukemia [71]. SPI1 gene activation is a general feature of malignant proerythroblastic transformation that occurs in mice infected with Friend and Rauscher viruses.

In additional, SPI1 encodes transcription factor PU.1(another name for SPI1), whose knockout and overexpression are known to be associated with leukemia [72]. Recently, Ye et al. found that PU.1 plays a key role in early T cell differentiation through a core network topology [73]. Therefore, SPI1 plays an important role in leukemia.

RFXANK: No direct relation between RFXANK and leukemia has been found. However, as a kind of transcription factor, RFXB [74] (another name for RFXANK) positively regulates HLA genes, which are closely interrelated with leukemia. According to the prediction of PreEGSRF, it was hypothesized that RFXANK has a close relationship with leukemia. This hypothesis could be verified by later biological experiments.

In conclusion, the prediction of PreEGSRF has biological significance in that the five genes are essential in the process of leukemia.

3.3.4 Functional Enrichment Analyses

Gene Ontology (GO) and KEGG analyses were performed to understand the underlying biological mechanisms. GO analysis explored the biological significance of genes by using the R package 'clusterProfiler' [75]. The enriched GO terms were chosen based on *p-value* < 0.05 and count >5 [76]. The KEGG analyses were based on pathways with *p-value* < 0.05.

Based on the PreEGSRF method, 25 candidate EGS were used to perform the functional enrichment GO analyses. As shown in Table 3, the EGS are associated with transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific DNA binding, etc., which have a high correlation with leukemia.

TABLE 3
Enriched GO Terms in the Leukemia Dataset

Description	*p-value*	Count
GO:0001228~transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific DNA binding	2.96×10^{-14}	12
GO:0000978~RNA polymerase II proximal promoter sequence-specific DNA binding	5.03×10^{-14}	12
GO:0000987~proximal promoter sequence-specific DNA binding	7.52×10^{-14}	12
GO:0000982~transcription factor activity, RNA polymerase II proximal promoter sequence-specific DNA binding	1.19×10^{-12}	11
GO:0001077~transcriptional activator activity, RNA polymerase II proximal promoter sequence-specific DNA binding	1.47×10^{-9}	8
GO:0001227~transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific DNA binding	1.36×10^{-8}	7
GO:0001047~core promoter binding	2.28×10^{-8}	6
GO:0001078~transcriptional repressor activity, RNA polymerase II proximal promoter sequence-specific DNA binding	9.34×10^{-7}	5

Moreover, the 25 candidate EGS were enriched in five KEGG pathways, which were all related to leukemia. Specifically, four EGS reported by PreEGSRF were involved in the pathways as shown in Table 4.

Among the five pathways, the "Transcriptional

1545-5963 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 13,2020 at 07:01:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3027392, IEEE/ACM Transactions on Computational Biology and Bioinformatics

misregulation in cancer pathway" may lead to various kinds of cancers, such as leukemia [77] and small-cell lung cancer. The "Th17 cell differentiation pathway" is related to leukemia because the leukemia inhibitory factor inhibits T helper 17 cell differentiation and confers treatment effects of neural progenitor cell therapy in autoimmune disease [78]. The "Osteoclast differentiation pathway" demonstrates that adult T-cell leukemia cells overexpress Wnt5a and promote osteoclast differentiation [79]. The "Fanconi anemia pathway" is associated with leukemia [80]. The "Acute myeloid leukemia pathway" is enriched in leukemia.

TABLE 4
Enriched KEGG Pathways in the Leukemia Dataset; 4 EGS (bolded) are involved

Description	*p-value*	Genes
hsa05202~Transcriptional misregulation in cancer	4.72×10^{-6}	HMGA2/**RXRB**/RUNX1/MITF/**SPI1**
hsa04659~Th17 cell differentiation	4.84×10^{-4}	**STAT1**/**RXRB**/RUNX1
hsa04380~Osteoclast differentiation	8.18×10^{-4}	**STAT1**/MITF/**SPI1**
hsa03460~Fanconi anemia pathway	2.89×10^{-3}	**HES1**/ BRCA1
hsa05221~Acute myeloid leukemia	4.29×10^{-3}	RUNX1/ **SPI1**

4 Conclusions

Predicting the essential genes in differential network analyses is biologically significant. Here, we present a method of predicting EGS based on random forest model with two main features. First, the information of each node is vectorized to a 5-dimensional feature vector by extracting both topological structure and gene expression features in comparison states. Second, a positive sample expansion method based on PCC is introduced to address the problem of unbalanced positive and negative samples.

In the simulated data experiments, PreEGSRF has been compared with three classical methods and three machine learning-based methods. A series of indicators show the excellent performance of PreEGSRF in EGS prediction. This is partly because PreEGSRF has a strong ability to identify multiple features by comparing two biological states, namely, the topological structure of the network and gene expression, which may make a gene 'essential'.

In the real data experiments, PreEGSRF predicted five leukemia-related EGS, four of which were supported by literature researches. Moreover, the five enriched KEGG pathways involving these four EGS are closely interrelated with leukemia. The fifth predicted EGS RFXANK(RXRB) needs further study. While there is a lack of annotation information about the relationship between EGS RFXANK(RXRB) and leukemia, the EGS RFXANK(RXRB) has been found to be enriched in two KEGG pathways (hsa05202 and hsa04659), suggesting that RFXANK(RXRB) is closely interrelated with leukemia. New targeted biological experiments to examine our hypothesis would help test our predictions.

Single-cell RNA sequencing (scRNA-seq) has been increasingly used to study gene expression at the level of individual cells and graduated processes, thus adding another dimension to understand gene expression regulation and dynamics [81]. A network construction method has been developed in which a cell-specific network (CSN) [82] for each single cell from scRNA-seq data (i.e., one network for one cell) transforms the data from an 'unstable' gene expression form to 'stable' gene association form on a single-cell basis. In particular, CSN represents an excellent method of performing scRNA-seq data analyses and provides insights for expanding the application of PreEGSRF. In the future, PreEGSRF can be applied to scRNA-seq datasets based on CSN.

In general, the PreEGSRF method is a useful tool to identify essential genes in networks, and it has broad application prospects for the discovery of biomarkers of complex cellular systems, such as driver genes in cell fate decisions or diseases.

Acknowledgment.

This work was supported by the National Natural Science Foundation of China [No. 61873156], Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01), Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (LCNBI) and ZJLab. Q. Nie's research was partially supported by a NSF grant DMS1763272 and a grant from the Simons Foundation (594598).

*Corresponding author: Jiao Wang and Qing Nie.

References

[1] D. Yang, R. S. Parrish, and G. N. Brock, "Empirical evaluation of consistency and accuracy of methods to detect differentially expressed genes based on microarray data," *Computers in biology and medicine*, vol. 46, no. 1, pp. 1-10, 2014.

[2] E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, A. M. Chu, C. Connelly, K. Davis, F. Dietrich, S. W. Dow, M. El Bakkoury, F. Foury, S. H. Friend, E. Gentalen, G. Giaever, J. H. Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, D. J. Lockhart, A. Lucau-Danila, M. Lussier, N. M'Rabet, P. Menard, M. Mittmann, C. Pai, C. Rebischung, J. L. Revuelta, L. Riles, C. J. Roberts, P. Ross-MacDonald, B. Scherens, M. Snyder, S. Sookhai-Mahadeo, R. K. Storms, S. Veronneau, M. Voet, G. Volckaert, T. R. Ward, R. Wysocki, G. S. Yen, K. Yu, K. Zimmermann, P. Philippsen, M. Johnston, and R. W. Davis, "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis," *Science*, vol. 285, no. 5429, pp. 901-6, 1999.

[3] G. Lamichhane, M. Zignol, N. J. Blades, D. E. Geiman, A. Dougherty, J. Grosset, K. W. Broman, and W. R. Bishai, "A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: Application to Mycobacterium tuberculosis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 12, pp. 7213-7218, 2003.

[4] G. Zhu, S. Li, J. Wu, F. Li, and X. M. Zhao, "Identification of Functional Gene Modules Associated With STAT-Mediated Antiviral Responses to White Spot Syndrome Virus in Shrimp," *Front*

1545-5963 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 13,2020 at 07:01:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3027392, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Physiol, vol. 10, no. 212, pp. 1-9, 2019.

[5] Y. Y. Wang, C. Cui, L. Qi, H. Yan, and X. M. Zhao, "DrPOCS: Drug Repositioning Based on Projection Onto Convex Sets," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 16, no. 1, pp. 154-162, Jan-Feb, 2019.

[6] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K. D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Guldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kotter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C. Y. Wang, T. R. Ward, J. Wilhelmy, E. A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis, and M. Johnston, "Functional profiling of the Saccharomyces cerevisiae genome," *Nature*, vol. 418, no. 6896, pp. 387-91, 2002.

[7] J. Harborth, S. M. Elbashir, K. Bechert, T. Tuschl, and K. Weber, "Identification of essential genes in cultured mammalian cells using small interfering RNAs," *Journal of Cell Science*, vol. 114, no. 24, pp. 4557-4565, 2001.

[8] L. A. Gallagher, E. Ramage, M. A. Jacobst, R. Kaul, M. Brittnacher, and C. Manoil, "A comprehensive transposon mutant library of Francisella novicida, a bioweapon surrogate," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 3, pp. 1009-1014, 2007.

[9] G. Zhu, H. Yang, X. Chen, J. Wu, Y. Zhang, and X. M. Zhao, "CSTEA: a webserver for the Cell State Transition Expression Atlas," *Nucleic Acids Res*, vol. 45, no. W1, pp. W103-W108, 2017.

[10] X. M. Zhao, and S. Li, "HISP: a hybrid intelligent approach for identifying directed signaling pathways," *J Mol Cell Biol*, vol. 9, no. 6, pp. 453-462, 2017.

[11] D. Yang, R. S. Parrish, and G. N. Brock, "Empirical evaluation of consistency and accuracy of methods to detect differentially expressed genes based on microarray data," *Computers in biology and medicine*, vol. 46, no. 1, pp. 1-10, 2014.

[12] A. de la Fuente, "From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases," *Trends Genet*, vol. 26, no. 7, pp. 326-33, 2010.

[13] S. P. Deng, L. Zhu, and D. S. Huang, "Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks," *BMC Genomics*, vol. 16, Suppl 3, pp. S4, 2015.

[14] D. S. Huang, and X. Huang, "Improved performance in protein secondary structure prediction by combining multiple predictions," *Protein Pept Lett*, vol. 13, no. 10, pp. 985-91, 2006.

[15] D. S. Huang, and H. J. Yu, "Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 10, no. 2, pp. 457-67, 2013.

[16] S. P. Deng, and D. S. Huang, "SFAPS: an R package for structure/function analysis of protein sequences based on informational spectrum method," *Methods*, vol. 69, no. 3, pp. 207-212, 2014.

[17] N. J. Hudson, A. Reverter, and B. P. Dalrymple, "A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation," *PLoS computational biology*, vol. 5, no. 5, pp. e1000382, 2009.

[18] O. Odibat, and C. K. Reddy, "Ranking differential hubs in gene co-expression networks," *Journal of bioinformatics computational biology*, vol. 10, no. 01, pp. 1240002, 2012.

[19] M. Bockmayr, F. Klauschen, C. Denkert, and J. Budczies, "New network topology approaches reveal differential correlation patterns; In breast cancer," *BMC Systems Biology*, vol. 7, no. 1, pp. 78-78, 2013.

[20] Y. Chen, and D. Xu, "Understanding protein dispensability through machine-learning analysis of high-throughput data," *Bioinformatics*, vol. 21, no. 5, pp. 575-581, 2004.

[21] J. Wang, M. Li, H. Wang, and Y. Pan, "Identification of essential proteins based on edge clustering coefficient," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 4, pp. 1070–1080, 2012

[22] D.S. Huang, X. M. Zhao, G. B. Huang, and Y. M. Cheung, "Classifying protein sequences using hydropathy blocks," *Pattern Recognition*, vol. 39, no. 12, pp. 2293–2300, 2006.

[23] J. F. Xia, X. M. Zhao, J. N. Song and D. S. Huang, "APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility," *BMC Bioinformatics*, vol. 11, no. 1, pp. 174, 2010.

[24] D. S. Huang, L. Zhang, K. Han, S. P. Deng, K. Yang, and H. B. Zhang, " Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," *Current Protein & Peptide Science*, vol. 15, no. 6, pp. 553-560, 2014.

[25] D. S. Huang, and C. H. Zheng, "Independent component analysis based penalized discriminant method for tumor classification using gene expression data," *Bioinformatics*, vol. 22, no. 15, pp. 1855-1862, 2006.

[26] K. Yang, X. Zhao, D. Waxman, and X. M. Zhao, "Predicting drug-disease associations with heterogeneous network embedding," *Chaos*, vol. 29, no. 12, pp. 123109, 2019.

[27] J. Song, F. Li, K. Takemoto, G. Haffari, T. Akutsu, K.-C. Chou, and G. I. Webb, "PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework," *Journal of theoretical biology*, vol. 443, pp. 125-137, 2018.

[28] A. Singhal, D. Roy, S. Mittal, J. Dhar, and A. Singh, "A New Computational Approach to Identify Essential Genes in Bacterial Organisms Using Machine Learning," *Theories, Applications and Future Directions-Volume I*, Singapore: Springer, vol. 798, pp. 67-79, 2019.

[29] R. Zhang, and Y. Lin, "DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes," *Nucleic Acids Research*, vol. 37, pp. D455-D458, 2009.

[30] C. Christensen, J. Thakar, and R. Albert, "Systems-level insights into cellular regulation: inferring, analysing, and modelling intracellular networks," *Iet Systems Biology*, vol. 1, no. 2, pp. 61-77, 2007.

[31] X. Zhang, M. L. Acencio, and N. Lemke, "Predicting Essential Genes and Proteins Based on Machine Learning and Network Topological Features: A Comprehensive Review," *Frontiers in Physiology*, vol. 7, nol. 1, pp. 75, 2016.

[32] S.B. Seidman, "Network structure and minimum degree." *Social Networks*, vol.5, no.3, pp. 269-287, 1983.

[33] L. C. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, vol. 40, no. 1, pp. 35-41, 1977.

[34] S. Wuchty, and P. F. Stadler, "Centers of complex networks," *Journal of Theoretical Biology*, vol. 223, no. 1, pp. 45-53, 2003.

[35] S. K. Jari, K. Mikko, O. Jukka-Pekka, K. Kimmo, and K. János, "Generalizations of the clustering coefficient to weighted complex networks," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 75, no. 2, pp. 027105, 2007.

[36] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. Bossuyt, "The diagnostic odds ratio: a single indicator of test performance," *Journal of clinical epidemiology*, vol.56, no.11, pp.1129-1135, 2003.

[37] A. Italiano, F. B. Vandenbos, J. Otto, J. Mouroux, D. Fontaine, P. Y. Marcy, N. Cardot, A. Thyss, and F. Pedeutour, "Comparison of the epidermal growth factor receptor gene and protein in

1545-5963 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 13,2020 at 07:01:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3027392, IEEE/ACM Transactions on Computational Biology and Bioinformatics

primary non-small-cell-lung cancer and metastatic sites: implications for treatment with EGFR-inhibitors," *Ann Oncol*, vol. 17, no. 6, pp. 981-985, 2006.

[38] Scikit-learn Machine Learning in Python. https://scikit-learn.org/stable/. 2019

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321-357, 2002.

[40] Y. Lichtblau, K. Zimmermann, B. Haldemann, D. Lenze, M. Hummel, and U. Leser, "Comparative assessment of differential network analysis methods," *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 837, 2016.

[41] B. Albert-László, "Scale-free networks: a decade and beyond," *Science*, vol. 325, no. 5939, pp. 412-413, 2009.

[42] X. F. Zhang, L. Ou-Yang, and H. Yan, "Incorporating prior information into differential network analysis using nonparanormal graphical models," *Bioinformatics*, vol. 33, no. 16, pp. 2436, 2017.

[43] N. Shane, A. B. Stergachis, R. Alex, S. Richard, B. Elhanan, and J. A. Stamatoyannopoulos, "Circuitry and dynamics of human transcription factor regulatory networks," *Cell*, vol. 150, no. 6, pp. 1274-1286, 2012.

[44] A. B. Stergachis, N. Shane, S. Richard, H. Eric, A. P. Reynolds, Z. Miaohua, B. Rachel, C. Theresa, S. S. Sandra, and L. Kristen, "Conservation of trans-acting circuitry during mammalian regulatory evolution," *Nature*, vol. 515, no. 7527, pp. 365-70, 2014.

[45] S. A. Sloan, S. Darmanis, N. Huber, T. A. Khan, F. Birey, C. Caneda, R. Reimer, S. R. Quake, B. A. Barres, and S. P. PaåŸCa, "Human Astrocyte Maturation Captured in 3D Cerebral Cortical Spheroids Derived from Pluripotent Stem Cells," *Neuron*, vol. 95, no. 4, pp. 779-790, 2017.

[46] B. Hassannia, B. Wiernicki, I. Ingold, F. Qu, and T. V. Berghe, "Nano-targeted induction of dual ferroptotic mechanisms eradicates high-risk neuroblastoma," *Journal of Clinical Investigation*, vol. 128, no. 8, pp. 3341-3355, 2018.

[47] L. W. Harris, M. Wayland, M. Lan, M. Ryan, T. Giger, H. Lockstone, I. Wuethrich, M. Mimmack, L. Wang, and M. Kotter, "The cerebral microvasculature in schizophrenia: a laser capture microdissection study," *PLoS One*, vol. 3, no. 12, pp. e3964, 2008.

[48] S. Ganesan, A. Alex, E. Chendamarai, N. Balasundaram, H. Palani, S. David, U. Kulkarni, M. Aiyaz, R. Mugasimangalam, and A. J. L. Korula, "Rationale and efficacy of proteasome inhibitor combined with arsenic trioxide in the treatment of acute promyelocytic leukemia," *Leukemia*, vol. 30, no. 11, pp. 2169, 2016.

[49] M. Mahdavi, M. Nassiri, M. M. Kooshyar, M. Vakili-Azghandi, A. Avan, R. Sandry, S. Pillai, A. K. y. Lam, and V. Gopalan, "Hereditary breast cancer; Genetic penetrance and current status with BRCA," *Journal of cellular physiology*, vol. 234, no. 5, pp. 5741-5750, 2019.

[50] N. Aygun, and O. Altungoz, "MYCN is amplified during S phase, and c-myb is involved in controlling MYCN expression and amplification in MYCN-amplified neuroblastoma cell lines," *Molecular medicine reports*, vol. 19, no. 1, pp. 345-361, 2019.

[51] L. Cai, Y. H. Tsai, P. Wang, J. Wang, D. Li, H. Fan, Y. Zhao, R. Bareja, R. Lu, and E. M. Wilson, "ZFX Mediates Non-canonical Oncogenic Functions of the Androgen Receptor Splice Variant 7 in Castrate-Resistant Prostate Cancer," *Molecular Cell*, vol. 72, pp. 341-354, 2018.

[52] F. Desmots, M. Roussel, C. Pangault, F. Llamas-Gutierrez, C. Pastoret, E. Guiheneuf, J. Le Priol, V. Camara-Clayette, G. Caron, and C. Henry, "Pan-HDAC Inhibitors Restore PRDM1 Response to IL21 in CREBBP-Mutated Follicular Lymphoma," *Clinical Cancer Research*, vol. 25, no. 2, pp. 735-746, 2019.

[53] A. Narayanan, F. Gagliardi, A. L. Gallotti, S. Mazzoleni, M. Cominelli, L. Fagnocchi, M. Pala, I. S. Piras, P. Zordan, and N. Moretta, "The proneural gene ASCL1 governs the transcriptional subgroup affiliation in glioblastoma stem cells by directly repressing the mesenchymal gene NEGS1," *Cell Death Differentiation*, vol. 26, no. 9, pp. 1813-1831, 2018.

[54] A. García-Martínez, J. Sottile, L. Sánchez-Tejada, C. Fajardo, R. Cámara, C. Lamas, V. M. Barberá, and A. Picó, "DNA Methylation of Tumor Suppressor Genes in Pituitary Neuroendocrine Tumors," *The Journal of Clinical Endocrinology & Metabolism*, vol. 104, no. 4, pp. 1272-1282, 2018.

[55] L. Wang, J. Ma, X. Wang, F. Peng, X. Chen, B. Zheng, C. Wang, Z. Dai, J. Ai, and S. Zhao, "Kaiso (ZBTB33) Downregulation by Mirna-181a Inhibits Cell Proliferation, Invasion, and the Epithelial–Mesenchymal Transition in Glioma Cells," *Cellular Physiology and Biochemistry*, vol. 48, no. 3, pp. 947-958, 2018.

[56] Y. Gao, D. Han, L. Sun, Q. Huang, G. Gai, Z. Wu, W. Meng, and X. Chen, "PPARα Regulates the Proliferation of Human Glioma Cells through miR-214 and E2F2," *BioMed research international*, vol. 2018, no. 1, pp. 3842753, 2018.

[57] S. Davudian, B. Mansoori, N. Shajari, A. Mohammadi, and B. Baradaran, "BACH1, the master regulator gene: A novel candidate target for cancer therapy," *Gene*, vol. 588, no. 1, pp. 30-37, 2016.

[58] R. Roychoudhuri, R. L. Eil, D. Clever, C. A. Klebanoff, M. Sukumar, F. M. Grant, Z. Yu, G. Mehta, H. Liu, and P. Jin, "The transcription factor BACH2 promotes tumor immunosuppression," *The Journal of clinical investigation*, vol. 126, no. 2, pp. 599-604, 2016.

[59] F. Blasi, C. Bruckmann, D. Penkov, and L. Dardaei, "A tale of TALE, PREP1, PBX1, and MEIS1: Interconnections and competition in cancer," *Bioessays*, vol. 39, no. 5, pp. 1600245, 2017.

[60] M. Lin, J. Lin, C. Hsu, H. Juan, P. Lou, and M. Huang, "GATA3 interacts with and stabilizes HIF-1α to enhance cancer cell invasiveness," *Oncogene*, vol. 36, no. 30, pp. 4243, 2017.

[61] C. De Bustos, A. Smits, B. Strömberg, V. P. Collins, M. Nistér, and G. Afink, "A PDGFRA promoter polymorphism, which disrupts the binding of ZNF148, is associated with primitive neuroectodermal tumours and ependymomas," *Journal of medical genetics*, vol. 42, no. 1, pp. 31-37, 2005.

[62] L. Richart, F. X. Real, and V. J. Sanchez-Arevalo Lobo, "c-MYC partners with BPTF in human cancer," *Molecular & cellular oncology*, vol. 3, no. 3, pp. e1152346, 2016.

[63] J. T. Caldwell, H. Edwards, A. A. Dombkowski, S. A. Buck, L. H. Matherly, Y. Ge, and J. W. Taub, "Overexpression of GATA1 confers resistance to chemotherapy in acute megakaryocytic Leukemia," *PLoS One*, vol. 8, no. 7, pp. e68601, 2013.

[64] S. Saida, T. Zhen, E. Kim, K. Yu, G. Lopez, L. J. McReynolds, and P. P. J. L. Liu, "Gata2 deficiency delays leukemogenesis while contributing to aggressive leukemia phenotype in Cbfb-MYH11 knockin mice," vol. 34, no. 3, pp. 759-770, 2020.

[65] B. Liu, H. Ma, Q. Liu, Y. Xiao, S. Pan, H. Zhou, and L. Jia, "MiR-29b/Sp1/FUT4 axis modulates the malignancy of leukemia stem cells by regulating fucosylation via Wnt/beta-catenin pathway in acute myeloid leukemia," *J Exp Clin Cancer Res*, vol. 38, no. 1, pp. 200, 2019.

[66] H. Xiao, Y. Ding, Y. Gao, L. M. Wang, H. Wang, L. Ding, X. Li, X. Yu, and H. Huang, "Haploinsufficiency of NR3C1 drives glucocorticoid resistance in adult acute lymphoblastic leukemia cells by down-regulating the mitochondrial apoptosis axis, and is sensitive to Bcl-2 blockage," *Cancer Cell Int*, vol. 19, pp. 218, 2019.

[67] V. J. Weston, W. Wei, T. Stankovic, and P. Kearns, "Synergistic action of dual IGF1/R and MEK inhibition sensitizes childhood acute lymphoblastic leukemia (ALL) cells to cytotoxic agents and involves downregulation of STAT6 and PDAP1," *Exp Hematol*, vol. 63, pp. 52-63 e5, 2018.

[68] A. Rani, R. Greenlaw, R. A. Smith, and C. Galustian, "HES1 in immunity and cancer," *Cytokine & growth factor reviews*, vol. 30, no. 100, pp. 113-117,

1545-5963 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 13,2020 at 07:01:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3027392, IEEE/ACM Transactions on Computational Biology and Bioinformatics

2016.

[69] B. Kovacic, D. Stoiber, R. Moriggl, E. Weisz, R. G. Ott, R. Kreibich, D. E. Levy, H. Beug, M. Freissmuth, and V. Sexl, "STAT1 acts as a tumor promoter for leukemia development," *Cancer cell*, vol. 10, no. 1, pp. 77-87, 2006.

[70] L. Brown, J.-T. Cheng, Q. Chen, M. Siciliano, W. Crist, G. Buchanan, and R. Baer, "Site-specific recombination of the tal-1 gene is a common occurrence in human T cell leukemia," *The EMBO journal*, vol. 9, no. 10, pp. 3343-3351, 1990.

[71] F. Moreau-Gachelin, D. Ray, P. Tambourin, and A. Tavitian, "Spi-1 oncogene activation in Rauscher and Friend murine virus-induced acute erythroleukemias," *Leukemia*, vol. 4, no. 1, pp. 20-23, 1990.

[72] M. A. Yui, and E. V. Rothenberg, "Developmental gene networks: a triathlon on the course to T cell identity," *Nat Rev Immunol*, vol. 14, no. 8, pp. 529-45, 2014.

[73] Y. Ye, X. Kang, J. Bailey, C. Li, and T. Hong, "An enriched network motif family regulates multistep cell fate transitions with restricted reversibility," *PLoS Comput Biol*, vol. 15, no. 3, pp. e1006855, 2019.

[74] B. Das, and D. Majumder, "Information theory based analysis for understanding the regulation of HLA gene expression in human leukemia," Int. J. Infor. Sci. Techq, vol. 5, no. 2, pp. 39-55, 2012.

[75] G. C. Yu, L. G. Wang, Y. Y. Han, and Q. Y. He, "clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters," *Omics-a Journal of Integrative Biology*, vol. 16, no. 5, pp. 284-287, 2012.

[76] W. Yuan, X. Li, L. Liu, C. Wei, D. Sun, S. Peng, L. J. B. Jiang, and b. r. communications, "Comprehensive analysis of lncRNA-associated ceRNA network in colorectal cancer," vol. 508, no. 2, pp. 374-379, 2019.

[77] N. Zhang, Y. Chen, Y. Shen, S. Lou, and J. J. B. C. Deng, "Comprehensive analysis the potential biomarkers for the high-risk of childhood acute myeloid leukemia based on a competing endogenous RNA network," *Blood Cell, Molecules, and Diseases*, vol. 79, pp. 102352, 2019.

[78] W. Cao, Y. Q. Yang, Z. Y. Wang, A. L. Liu, L. Fang, F. L. Wu, J. Hong, Y. F. Shi, S. Leung, C. Dong, and J. W. Z. Zhang, "Leukemia Inhibitory Factor Inhibits T Helper 17 Cell Differentiation and Confers Treatment Effects of Neural Progenitor Cell Therapy in Autoimmune Disease," *Immunity*, vol. 35, no. 2, pp. 273-284, 2011.

[79] S. Quentin, W. Cuccuini, R. Ceccaldi, O. Nibourel, C. Pondarre, M. P. Pages, N. Vasquez, C. D. d'Enghien, J. Larghero, R. P. de Latour, V. Rocha, J. H. Dalle, P. Schneider, M. Michallet, G. Michel, A. Baruchel, F. Sigaux, E. Gluckman, T. Leblanc, D. Stoppa-Lyonnet, C. Preudhomme, G. Socie, and J. Soulier, "Myelodysplasia and leukemia of Fanconi anemia are associated with a specific pattern of genomic abnormalities that includes cryptic RUNX1/AML1 lesions," *Blood*, vol. 117, no. 15, pp. E161-E170, 2011.

[80] M. Bellon, N. L. Ko, M. J. Lee, Y. Yao, T. A. Waldmann, J. B. Trepel, and C. Nicot, "Adult T-cell leukemia cells overexpress Wnt5a and promote osteoclast differentiation," *Blood*, vol. 121, no. 25, pp. 5045-5054, 2013.

[81] J. Xie, J. Sun, J. Feng, F. Yang, J. Wang, T. Wen, and Q. Nie, "Kernel Differential Subgraph Analysis to Reveal the Key Period Affecting Glioblastoma," *Biomolecules*, vol. 10, no. 2, pp. 318, 2020.

[82] H. Dai, L. Li, T. Zeng, and L. Chen, "Cell-specific network constructed by single-cell RNA sequencing data," *Nucleic Acids Res*, vol. 47, no. 11, pp. e62, 2019.

Jiang Xie received her Ph.D. in computer science from Shanghai University, China in 2008. She is currently an associate professor in the School of Computer Engineering and Science in Shanghai University. Her main research topics include computational biology, bioinformatics, and high-performance computing.

Chang Zhao received her B.S. degree from Henan Polytechnic University in 2019. She is now studying towards a master's degree in the School of Computer Engineering and Science of Shanghai University. Her main research topics are bioinformatics and high-performance computing.

Jiamin Sun received her M.S. degree from Shanghai Univerisity in 2020. Her main research topics are bioinformatics and high-performance computing.

Jiaxin Li received his M.S. degree from Shanghai Univerisity in 2019. His main research topics are bioinformatics and high-performance computing.

Fuzhang Yang received his M.S. degree from Shanghai Univerisity in 2020. His main research topics are bioinformatics and high-performance computing.

Jiao Wang received her Ph.D. in neural science from Shanghai University, China, in 2012. She is currently an associate professor at the Life Science Centre. Her main research topics include brain research, bioinformatics and neural degeneration disease.

1545-5963 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 13,2020 at 07:01:04 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3027392, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Qing Nie is a Chancellor's professor in the Department of Mathematics and Department Developmental and Cell Biology at the University of California, Irvine. His main research areas include computational systems biology, cell signaling, stem cell, morphogenesis, bioinformatics, and scientific computing.

1545-5963 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 13,2020 at 07:01:04 UTC from IEEE Xplore. Restrictions apply.