Hindawi
Complexity
Volume 2020, Article ID 6643551, 12 pages
https://doi.org/10.1155/2020/6643551

Research Article

Driver Attribute Filling for Genes in Interaction Network via Modularity Subspace-Based Concept Learning from Small Samples

Fei Xie,[1] Jianing Xi ,[1,2] and Qun Duan[3]

[1]*School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China*
[2]*Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China*
[3]*School of Computer Science, Xianyang Normal University, Xi'an 712000, China*

Correspondence should be addressed to Jianing Xi; xjn@nwpu.edu.cn

Received 12 October 2020; Revised 9 November 2020; Accepted 11 November 2020; Published 23 November 2020

Academic Editor: Jianxin Li

The aberrations of a gene can influence it and the functions of its neighbour genes in gene interaction network, leading to the development of carcinogenesis of normal cells. In consideration of gene interaction network as a complex network, previous studies have made efforts on the driver attribute filling of genes via network properties of nodes and network propagation of mutations. However, there are still obstacles from problems of small size of cancer samples and the existence of drivers without property of network neighbours, limiting the discovery of cancer driver genes. To address these obstacles, we propose an efficient modularity subspace based concept learning model. Our model can overcome the curse of dimensionality due to small samples via dimension reduction in the task of attribute concept learning and explore the features of genes through modularity subspace beyond the network neighbours. The evaluation analysis also demonstrates the superiority of our model in the task of driver attribute filling on two gene interaction networks. Generally, our model shows a promising prospect in the application of interaction network analysis of tumorigenesis.

1. Introduction

Gene performs a function via the synthesis of its product protein encoded by the gene in human cells, and the interactions between proteins lead to the functional cooperation between different genes [1]. Subsequently, the aberrations of gene can not only alter the function of the gene itself but also influence the functions of other genes that interact with the aberrated gene, and both ways can lead to carcinogenesis of normal cells [2, 3]. Unlike the straightforward function abnormality of a gene caused by the aberration itself, the functional abnormality of genes caused by interaction among more than one gene is rather implicit to be understood [2]. To investigate the roles of genes among their interactions, the protein-protein interaction network has been established to systematically describe all the interactions between each of the gene pairs that have been collected so far [1]. When the network topology is investigated, it is observed that gene interaction network supports the property of being scale-free, which fairly satisfies the definition of complex network [4]. Consequently, through the network properties such as centrality and betweenness involved in complex network analytics [5, 6], it is an unprecedent opportunity to attributes of genes by investigating their roles in interaction network.

To unveil whether a certain gene has attributes of cancer drivers, that is, the capability of subverting a normal cell into malignant cancer cell [3], a direct way is to fill the attributes of nodes through their network properties. For a certain complex network, a number of numerical measurements have been established for describing the property of nodes in network, such as degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and Bonacich

centrality [7]. With the advancing of complex network based analytical methods [8, 9], the cancer driver attributes of genes can be discovered through the properties of their related nodes in the network [10, 11]. Nevertheless, recent studies have demonstrated that there are also a noticeable number of cancer driver genes that do not show explicit high centrality in gene interaction network [12, 13]. Therefore, despite the successful achievement obtained by the previously published complex network based methods, by directly using the previous network property based studies, these types of cancer drivers are very likely to be missed in the task of node attribute filling of genes in interaction network.

To identify cancer drivers without high centrality, an intuitive idea is to compensate the information beyond the gene interaction itself [2]. Fortunately, the existence of high-throughput sequencing data of cancer samples has provided another important source for the discovery of cancer driver attributes of genes [14]. Sequencing data can provide the mutations of each gene in the tested cancer samples [15, 16], containing the information of mutation profiles among cancer samples which is far beyond the network itself [3, 17]. Afterward, there have been a bunch of researches that incorporate information from both interaction network and mutation data via network propagation of mutation frequencies [2, 12, 13]. Notwithstanding, recent studies have shown that there exist a number of cancer driver genes that are neither frequently mutated nor close to other highly mutated genes in interaction network [18]. Furthermore, when the number of cancer samples is small, the mutation frequencies of the mutation data may not be representative for the identification of highly mutated cancer driver genes [19]. Since both the scale of gene interaction network and the dimension of mutation data are of the order of ten thousand, the small sample number that is usually less than one hundred due to the practical reason is also likely to result in the curse of dimensionality [18]. Generally, there is still a lack of a driver attribute filling method for genes in interaction network with capability of mutation compensation for small samples.

To discover cancer driver attributes of gene in interaction network with small samples of mutation data, we proposed a novel modularity subspace based concept learning method, which can efficiently identify the driver genes in interaction network. To circumvent the curse of dimensionality resulting from small size of samples, we introduce the dimension reduction paradigm to relieve the shortage of sample size. For detecting driver genes that are neither frequently mutated nor connect to other highly mutated genes, we introduce the network modularity of the genes as the subspace features to achieve dimension reduction of the gene interaction network, where the network modularity represents the membership of genes to different network modules [20, 21], independent of mutation frequencies and direct connections of genes. After obtaining the feature subspace learnt from the compensation of interaction network and mutation data, we further utilize supervised concept learning technique to learn the rules in modularity feature space which associate the concept of cancer drivers [22, 23]. A systematic assessment illustrates the superiority of our proposed model over the existing network based methods for driver attribute filling of gene in two interaction networks from two distinct sources. In summary, our proposed modularity subspace based concept learning model is efficient for driver attribute filling of genes in interaction network with small samples of mutation data.

2. Materials and Methods

2.1. Modularity of Interaction Network. Gene interaction network contains modules, defined as groups of genes as nodes, and the gene interactions are relatively denser within the same module in comparison to the interactions between different modules (the connections of nodes between modules are sparser relatively) [24, 25]. Considering the independence between the membership of genes to network modules and mutation frequencies or direct connections of genes, we utilize the network modules as features for discriminating cancer driver genes in our study. Furthermore, in most situations, the dimension of features defined by network modules is far smaller than that defined by network nodes or neighbours [24, 26], benefiting the alleviation of curse of dimensionality due to small sample problem. Technically, when we use the adjacency matrix to describe the gene interaction network, a module can be defined as a partition of the matrix, where the summation of elements for the submatrix of the selected nodes is expected to be large [20]. Here the partition of a module can be depicted through a vector whose dimension is equal to the number of gene nodes (suppose the total number of genes is p), and the coefficients larger than zero represent that the corresponding genes belong to the module.

When the total number of network modules is a predefined number k, we can use k independent vectors $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_k$ to delineate the partitions of the p genes to the k modules, denoted as matrix formation $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_k]$. According to the derivation by Newman [20], the density of gene nodes within the k modules can be approximately calculated through the modularity score, which is defined via a quadratic form of vectors $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_k$.

$$\text{modularity} = \sum_i \mathbf{h}_i^{\mathrm{T}} \left(\mathbf{A} - \frac{1}{2p} \mathbf{d}\mathbf{d}^{\mathrm{T}} \right) \mathbf{h}_i = \sum_i \mathbf{h}_i^{\mathrm{T}} \mathbf{B} \mathbf{h}_i = \mathrm{Tr}\left\{ \mathbf{H}^{\mathrm{T}} \mathbf{B} \mathbf{H} \right\}, \tag{1}$$

where matrix $\mathbf{A}$ is the normalized adjacency matrix of gene interaction network and vector $\mathbf{d}$ is the degree vector of the network, whose i-th coefficient is the summation of the i-th row of matrix $\mathbf{A}$(or i-th column summation, since matrix $\mathbf{A}$ is symmetric). For brevity, the modularity matrix $\mathbf{B}$ is used to represent the result of matrix subtraction $\mathbf{B} = \mathbf{A} - 1/2p\mathbf{d}\mathbf{d}^T$.

For matrix $\mathbf{H}$, its j-th column $\mathbf{h}_j$ represents the partition of gene nodes within or without the j-th network module, where the relative value of coefficients is able to represent the inclusion and exclusion of its corresponding genes. Since the trace of matrix $\mathbf{H}^{\mathrm{T}}\mathbf{B}\mathbf{H}$ is equivalent to the summation of the quadratic forms for the k partitions of modules, by finding the partitions that can maximize the modularity score, we

can obtain the feature vectors $\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_k$ that can best reflect the memberships of genes to different modules [21]. Meanwhile, the i-th row of matrix $\mathbf{H}$ can be regarded as the feature vector of the i-th gene, denoted as $\mathbf{H} = [\mathbf{h}_{(1)}^{\mathrm{T}}, \mathbf{h}_{(2)}^{\mathrm{T}}, \ldots, \mathbf{h}_{(i)}^{\mathrm{T}}, \ldots, \mathbf{h}_{(p)}^{\mathrm{T}}]^{\mathrm{T}}$. For the i-th gene, when we regard vector $\mathbf{h}_{(i)}$ as the feature vector of the gene instead of its neighbourhood relations, the feature dimension k is largely reduced in comparison to the primary dimension p. Therefore, we can denote the feature space of $\mathbf{h}_{(i)}$ as the modularity subspace for the i-th gene.

2.2. Dimension Reduction for Small Samples. In consideration of the information beyond gene interaction network, the mutation data collected from sequencing of cancer samples is also another crucial source for cancer driver attribute filling. Restricted to the costs of sequencing, the amount of sequenced cancer samples is still limited in a relatively small number [27]. Considering that the number of genes is over twenty thousand, the number of cancer samples is usually fifth to one hundred [27]. Subsequently, a direct effect resulting from the small sample problem is the curse of dimensionality during the task of driver attribute discovery. Hence, like the dimension reduction of interaction network via modularity subspace, we also introduce the matrix decomposition architecture to learn the low-dimension representation of the p genes:

$$\min_{\mathbf{W}\geq 0, \widetilde{H}\geq 0} \left\| \mathbf{D} - \mathbf{W}\widetilde{H}^{\mathrm{T}} \right\|_F^2, \tag{2}$$

where matrix $\mathbf{D}$ is the n by p matrix of mutation data (n is the number of samples, $n << p$). Matrix $\widetilde{H}$ is a p by k matrix whose rows are the representation vector of genes, and matrix $\mathbf{W}$ is a nonnegative n by k matrix.

A typical strategy to learn the low dimension representation from the data matrix $\mathbf{D}$ is to apply the optimization procedures of alternatively updating rules in nonnegative matrix factorization (NMF) [28]. When the iteration reaches convergence, the multiplication of the two learnt matrices $\mathbf{W}$ and $\widetilde{H}$ can effectively approximate the original data matrix $\mathbf{D}$. Here matrix $\mathbf{W}$ is the coefficient matrix of the linear transformation from low-dimension representation $\widetilde{H}$ to the original data $\mathbf{D}$. The rows of matrix $\widetilde{H}$ are the low-dimension representations learnt from mutation data, preserving the cancer mutation information beyond interaction network. Based on the matrix decomposition based architecture, we can also achieve the dimension reduction of the mutation data from small samples.

2.3. Modularity Subspace-Based Dimension Reduction. To compensate the information from the interaction network and the mutation data of cancer samples, we further proposed a joint dimension reduction framework that can efficiently fuse the feature vectors of both the modularity subspace and the low-dimension feature of mutation data. When there are p genes in the interaction network and n samples of mutation data, the objective function of the joint dimension reduction framework is

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0, \mathbf{H}_\varepsilon} -\frac{1}{p^2}\mathrm{Tr}\left\{\mathbf{H}^{\mathrm{T}}\mathbf{B}\mathbf{H}\right\} + \frac{\lambda_X}{np}\left\|\mathbf{D} - \mathbf{W}\left(\mathbf{H} + \mathbf{H}_\varepsilon\right)^T\right\|_F^2 + \frac{\lambda_H}{pk}\|\mathbf{H}\|_F^2 \frac{\lambda_\varepsilon}{pk}\left\|\mathbf{H}_\varepsilon\right\|_F^2, \tag{3}$$

where k is the predefined number of low dimensions, which is set to 64 empirically. Here the dimensions of matrices $\mathbf{B}$ and $\mathbf{D}$ are p by p and n by p, respectively. The p by k matrix $\mathbf{H}$ represents the modularity features of genes, and the p by k matrix $\widetilde{H}$ represents the low-dimension features from mutation data. To joint the two types of features, we also introduce the residual matrix $\mathbf{H}_\varepsilon$, which is defined by their difference $\mathbf{H}_\varepsilon = \widetilde{H} - \mathbf{H}$. When the elements of matrix $\mathbf{H}_\varepsilon$ are close to zero, matrix $\widetilde{H}$ is then approximately equal to $\mathbf{H}$. The parameters λ_X, λ_H, and λ_ε are tuning parameters to control the weights of different terms in objective function.

In detail, the joint dimension reduction framework is composed of form terms, denoted as modularity term, sample data term, feature regularization term, and approximate residual term, respectively. First, the modularity term is introduced to reduce the dimension of the gene interaction network $\mathbf{H}$ via the modularity subspace feature learning. Second, the sample data term with tuning parameter λ_X is incorporated to compensate the information of mutations from cancer samples, which can learn the low-dimension features of each investigated gene $\widetilde{H}$. Third, the feature regularization term with tuning parameter λ_H is used to avoid extreme values of the learnt feature vector of genes during the joint dimension reduction procedure. Fourth, the approximate residual term with tuning parameter λ_ε is aimed to bridge the modularity subspace features from network $\mathbf{H}$ and the low-dimension features from mutation data $\widetilde{H}$, via a small value residual matrix $\mathbf{H}_\varepsilon$. Here we set λ_X and λ_H to 1.0 and 100 empirically (see details in Supplementary Materials). Specifically, since matrix $\mathbf{H}_\varepsilon$ is served as residual in the fusion of the two types of features $\mathbf{H}$ and $\widetilde{H}$, we therefore set the parameter λ_ε being 100 times of λ_H to ensure the dominant of $\mathbf{H}$ and strong penalty of $\mathbf{H}_\varepsilon$ at scale.

Note that there are two inequality constraints $\mathbf{W} \geq 0$ and $\mathbf{H} \geq 0$; we further introduce two Lagrange multipliers, matrix Ψ and matrix Φ, respectively. Subsequently, we can derive the Lagrange function for the optimization of the joint dimension reduction framework:

$$L = -\frac{1}{p^2}\mathrm{Tr}\left\{\mathbf{H}^{\mathrm{T}}\mathbf{B}\mathbf{H}\right\} + \frac{\lambda_X}{np}\left\|\mathbf{D} - \mathbf{W}\left(\mathbf{H}+\mathbf{H}_\varepsilon\right)^{\mathrm{T}}\right\|_F^2 + \frac{\lambda_H}{pk}\|\mathbf{H}\|_F^2 + \frac{\lambda_\varepsilon}{pk}\|\mathbf{H}_\varepsilon\|_F^2 + \mathrm{Tr}\left\{\mathbf{\Psi}^{\mathrm{T}}\mathbf{W}\right\} + \mathrm{Tr}\left\{\mathbf{\Phi}^{\mathrm{T}}\mathbf{H}\right\}. \tag{4}$$

For the three variables $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{H}_\varepsilon$, we can obtain their partial derivatives, respectively:

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{W}} &= -\frac{2\lambda_X}{np}\mathbf{D}\left(\mathbf{H}+\mathbf{H}_\varepsilon\right) + \frac{2\lambda_X}{np}\mathbf{W}\left(\mathbf{H}+\mathbf{H}_\varepsilon\right)^{\mathrm{T}}\left(\mathbf{H}+\mathbf{H}_\varepsilon\right) + \mathbf{\Psi};\\
\frac{\partial L}{\partial \mathbf{H}} &= -\frac{2}{p^2}\mathbf{B}\mathbf{H} - \frac{2\lambda_X}{np}\mathbf{D}^{\mathrm{T}}\mathbf{W} + \frac{2\lambda_X}{np}\left(\mathbf{H}+\mathbf{H}_\varepsilon\right)\mathbf{W}^{\mathrm{T}}\mathbf{W} + \frac{2\lambda_H}{pk}\mathbf{H} + \mathbf{\Phi};\\
\frac{\partial L}{\partial \mathbf{H}_\varepsilon} &= -\frac{2\lambda_X}{np}\mathbf{D}^{\mathrm{T}}\mathbf{W} + \frac{2\lambda_X}{np}\left(\mathbf{H}+\mathbf{H}_\varepsilon\right)\mathbf{W}^{\mathrm{T}}\mathbf{W} + \frac{2\lambda_\varepsilon}{pk}\mathbf{H}_\varepsilon.
\end{aligned} \tag{5}$$

We further employ the Karush-Kuhn-Tucker (KKT) conditions [28], where the three partial derivatives are all equal to zero: $\partial L/\partial \mathbf{H} = 0$, $\partial L/\partial \mathbf{W} = 0$, and $\partial L/\partial \mathbf{H} = 0$; and the complementary slackness conditions are also equal to zero: $\Psi_{ij}W_{ij} = 0$ and $\Phi_{ij}H_{ij} = 0$. Through these conditions, we can derive the three following equations:

$$\begin{aligned}
&\frac{\lambda_X}{np}\left[\mathbf{D}\left(\mathbf{H}+\mathbf{H}_\varepsilon\right)\right]_{ij}W_{ij} = \frac{\lambda_X}{np}\left[\mathbf{W}\left(\mathbf{H}+\mathbf{H}_\varepsilon\right)^{\mathrm{T}}\left(\mathbf{H}+\mathbf{H}_\varepsilon\right)\right]_{ij}W_{ij} = 0;\\
&\left[\frac{1}{p^2}\mathbf{A}\mathbf{H} + \frac{\lambda_X}{np}\mathbf{D}^{\mathrm{T}}\mathbf{W}\right]_{ij}H_{ij} = \left[\frac{1}{2p^3}\mathbf{d}\mathbf{d}^{\mathrm{T}}\mathbf{H} + \frac{\lambda_X}{np}\left(\mathbf{H}+\mathbf{H}_\varepsilon\right)\mathbf{W}^{\mathrm{T}}\mathbf{W} + \frac{\lambda_H}{pk}\mathbf{H}\right]_{ij}H_{ij};\\
&\mathbf{H}_\varepsilon\left(\frac{\lambda_X}{np}\mathbf{W}^{\mathrm{T}}\mathbf{W} + \frac{\lambda_\varepsilon}{pk}\mathrm{I}\right) = \frac{\lambda_X}{np}\left(\mathbf{D}^{\mathrm{T}}\mathbf{W} + \mathbf{H}\mathbf{W}^{\mathrm{T}}\mathbf{W}\right).
\end{aligned} \tag{6}$$

Through the above equations, we can reach the updating rules of the three variables:

$$\begin{aligned}
&W_{ij} \leftarrow W_{ij}\frac{\left[\mathbf{D}\left(\mathbf{H}+\mathbf{H}_\varepsilon\right)\right]_{ij}}{\left[\mathbf{W}\left(\mathbf{H}+\mathbf{H}_\varepsilon\right)^{\mathrm{T}}\left(\mathbf{H}+\mathbf{H}_\varepsilon\right)\right]_{ij}},\\
&H_{ij} \leftarrow H_{ij}/\left[npk\mathbf{A}\mathbf{H} + \lambda_X p^2 k\mathbf{D}^{\mathrm{T}}\mathbf{W}\right]_{ij}\left[\frac{nk}{2}\mathbf{d}\mathbf{d}^{\mathrm{T}}\mathbf{H} + \lambda_X p^2 k\left(\mathbf{H}+\mathbf{H}_\varepsilon\right)\mathbf{W}^{\mathrm{T}}\mathbf{W} + \lambda_H np^2\mathbf{H}\right]_{ij},\\
&\mathbf{H}_\varepsilon = \lambda_X k\left(\mathbf{D}^{\mathrm{T}}\mathbf{W} + \mathbf{H}\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)\left(\lambda_X k\mathbf{W}^{\mathrm{T}}\mathbf{W} + \lambda_\varepsilon n\mathrm{I}\right)^{-1}.
\end{aligned} \tag{7}$$

To learn the three variables, we alternatively apply the three updating rules until the iteration reaches its convergence. Finally, we can obtain the final results of the three matrices $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{H}_\varepsilon$ after convergence. For the i-th gene, the row vector $\mathbf{h}_{(i)}^{\mathrm{T}}$ in matrix $\mathbf{H}$ is the feature vector in modularity subspace compensated with mutation samples.

2.4. Modularity Subspace Concept Learning. To inference the cancer driver attributes of genes through the feature vectors in modularity subspace compensated with mutation samples, we further incorporate the idea of concept learning instead of the network property based strategy. Through the learnt rules, concept learning can efficiently identify the associated attributes of nodes if its features matching the rules [22, 23], which demonstrates more advantages in recognition task compared to the network property based strategy. Through the feature vectors with information from both modularity subspace and mutation samples and the assumption of independence of the k module dimension, we can adopt Bayesian based concept learning [23, 29] to establish the probabilistic rules that are associated with cancer driver attributes of genes:

$$y_i = \arg\max_c P(c)P\left(\mathbf{h}_{(i)}|c\right) = \arg\max_c P(c)\prod_{j=1}^{k} P\left(h_{ij}|c\right), \tag{8}$$

where the attribute index is denoted as c, whose value equaling 1 represents driver attribute and the value of 0 denotes nondriver attribute. The probability $P(c)$ is the prior probability of the attributes estimated from the distribution of the known driver attributes among genes. At the same time, the conditional probability $P(h_{ij}|tc)$ is estimated from the distribution between feature matrix $\mathbf{H}$ and the known driver attributes among genes. Consequently, after the estimation of these aforementioned probabilities, for the i-th tested gene, we can adopt its feature vector $\mathbf{h}_{(i)}$ into the probabilistic rules to infer the attribute y_i of the gene. The overall pipeline of the proposed modularity subspace based concept learning model is also drawn as schematic plot (see Figure 1).

3. Results

To evaluate the performance of our proposed driver attribute filling model, we apply our model on two gene interaction networks from independent sources. The first network is STRING network [30], which provides an integration of critical assessed interactions including both direct physical interactions and indirect functional associations between the proteins of their related genes. The second network is iRefIndex network [31], which curates protein interaction data of their related genes from a variety of sources and carefully consolidates the redundancy of interaction. In addition to the network data, we also incorporate the mutation data of sequencing samples from two distinct types of cancer, prostate cancer [32] and thyroid cancer [33]. The mutation data of both types of cancers are accessed from the cBioPortal database [34], which offers a web resource for cancer genomics data. The cancer driver attribute annotations of genes are collected from the COSMIC Cancer Gene Census database [35], which provides well-curated cancer driver genes that have been widely acceptable in tumorigenesis field.

3.1. Result Evaluation for STRING Network. When we apply our model on the 12233 genes in the STRING network, our model firstly yields the joint dimension features of modularity subspace and reduced dimensions of data from the combination of network data and cancer mutation data. The learnt joint dimension features are then used in the training of the probabilistic concept learning for driver attributes. Here, we utilize tenfold cross validation that approximately evenly splits the driver annotation of genes into ten groups without any overlap [36]. Lastly, the average performance of driver attribute filling results in different folds that are used in the evaluation study. We also apply two sorts of previously published methods ReMIC [12] and MUFFINN [13] as compared methods in the performance evaluation, both of which are based on network propagation of mutation frequencies, and their input data are also interaction network and cancer mutation data. Detailly, we compare our model with the three versions of ReMIC, where the diffusion parameters are 0.01, 0.02, and 0.03 as suggested in their applications [12]. Meanwhile, the two versions of MUFFINN are also recommended as DNmax and DNsum [13], which calculate the effect from gene's direct neighbours by maximum and summation, respectively. Finally, we compare the results of our model and the other two methods for performance evaluation.

Here we use the Receiver Operating Characteristic (ROC) curve [37] to evaluate the performance achieved by our model and other compared methods. In the evaluation of ROC curves, the y-axis represents the sensitivities under different threshold, where sensitivity is the fraction of the identified drivers in all known drivers, while the x-axis denotes the 1 minus specificity under different threshold, where specificity is the fraction of the identified nondrivers in all known nondrivers. When we draw the ROC curves of the three methods in the application of STRING network and prostate cancer data, we can get the phenomenon that the curve of our model is located at the top left of those of the other methods (see Figure 2(a)), which indicates that our model achieves the best performance in the driver attribute filling task. For example, when we examine that the sensitivities of the compared methods under the condition of their specificities are fixed to 0.10, the sensitivities of MUFFINN-DNmax and MUFFINN-DNsum are 21.67% and 23.33%, respectively. Also, the values of sensitivities of the three versions of ReMIC range from 31.67% to 36.67%. In contrast, the sensitivity yielded by our proposed model is 76.67%, higher than those of the other competing methods. Generally, through the assessment of ROC curves, we can observe a distinct advantage of our methods over the others in driver attribute filling of genes of prostate cancer.

For the results of STRING network for thyroid cancer, we can also observe similar results that our proposed model yields better results compared to the other compared methods, shown in the assessment of ROC curves (see Figure 2(b)). For example, when we investigate the sensitivities of the compared methods with their specificities fixed to 0.10, the sensitivities of MUFFINN-DNmax and MUFFINN-DNsum are 26.67% and 28.33%, respectively, and those of the three versions of ReMIC range from 46.67% to 55.00%. In comparison, the sensitivity achieved by our proposed model is 56.67%, demonstrating that our model also outperforms the others for data of thyroid cancer samples. Moreover, we also examine the specificities of these methods when their sensitivities are fixed to 0.10. In this case, the specificities yielded by ReMIC (Beta = 0.01), ReMIC (Beta = 0.02), and ReMIC (Beta = 0.03) are 53.12%, 50.75%, and 50.09%, respectively. MUFFINN-DNmax achieves a specificity of 48.57%, and MUFFINN-DNsum obtains a value of 29.57%. Here, our proposed modularity subspace based concept learning model accesses a specificity of 70.38%, which is also greater than those of the other compared methods. Thus, we can see a distinct advantage of

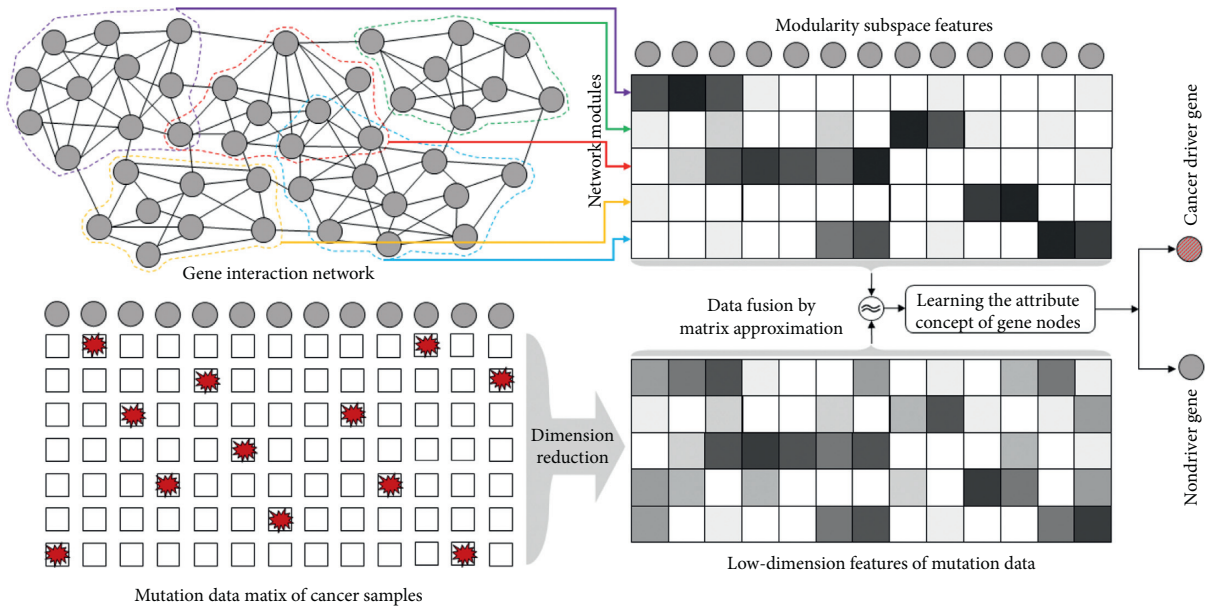

Figure 1: Schematic plot of pipeline of the proposed modularity subspace based concept learning model. The network modules and reduced dimensions are fused as the gene features in driver attribute concept learning.

our model in driver attribute filling when compared with the existing methods.

Since the Area Under the Curve (AUC) of ROC curve is a comprehensive performance measurement, we also compare the values of AUCs of our proposed modularity subspace based concept learning model and the other methods, for both prostate cancer (see Figure 2(c)) and thyroid cancer (see Figure 2(d)). Specifically, in the evaluation of attribute filling for prostate cancer, the AUC of MUFFINN-DNmax is 71.07% and that of MUFFINN-DNsum is 61.96%, and the AUCs of ReMIC with Beta = 0.01, 0.02, and 0.03 are 69.30%, 71.42%, and 72.31%, respectively. As for our proposed model, the value of AUC is 87.43% and is at least 20.9% higher than those of the other methods. Meanwhile, when we investigate the AUCs of attribute filling results for thyroid cancer, we can find that the AUCs of MUFFINN-DNmax, MUFFINN-DNsum, ReMIC (Beta = 0.01), ReMIC (Beta = 0.02), and ReMIC (Beta = 0.03) are 75.03%, 64.84%, 80.10%, 80.39%, and 81.63%, respectively. In comparison, the AUC of our proposed model reaches 86.10%, which is higher than AUCs of all the other methods. In addition, we also conduct ablation study on different modules in our model for the two cancer types (see Figure S1 in the Supplementary Materials). Consequently, through the application on STRING network with mutation data of two distinct types of cancers, we can observe the superiority of our model compared to the existing network based methods.

3.2. Result Evaluation for iRefIndex Network. At the same time, we also adopt the application of our model on the 12129 genes in the iRefIndex network. For mutation data of prostate cancer, when we assess the attribute filling results of our model and the other methods, the curve of our model is also closer to the northwest corner of the figures (see Figure 3(a)), showing a preferable performance over the other compared methods. In detail, when the specificities are fixed to 0.10, the corresponding sensitivities of different versions of ReMIC are rounded at 34.17%, and those of MUFFINN-DNmax and MUFFINN-DNsum are rounded at 22.50%. In contrast, the sensitivity acquired by our proposed model is 76.67%, which is distinctly larger than those of the other compared methods. When the sensitivities are locked at 0.10, compared to the fact that the specificities achieved by the other methods range from 30.49% to 45.78%, our model also outperforms the other compared methods, accessing the highest specificity of 55.85%. Generally, our model outperforms the other existing methods in the application of iRefIndex network with mutation data of prostate cancer.

For mutation data of thyroid cancer, when we evaluate the performance of the competing methods via ROC curve, we can acquire a similar phenomenon that our model exhibits a clear advantage over the other compared methods (see Figure 3(b)). Similar to the assessment above, when we fix the value of specificities to 0.10 and compare the sensitivities of these methods, our proposed method also obtains the largest sensitivity among those yielded by the others. Specifically, the sensitivity of our method is 56.67%, compared to 26.67% of MUFFINN-DNmax, 28.33% of MUFFINN-DNsum, 46.67% of ReMIC (Beta = 0.01), ReMIC (Beta = 0.02), and 55.00% of ReMIC (Beta = 0.03). Likewise, when the sensitivities of these competing methods are 0.10, the specificity of our proposed modularity subspace based concept learning model is 70.38%, which is also the highest among those of the compared methods. Considering that the values of the other compared method range from 29.57% to 53.12%, the value achieved by our model is at least 32.4% higher than those of the others. Consequently, when we use the mutation data from thyroid cancer samples, there is also

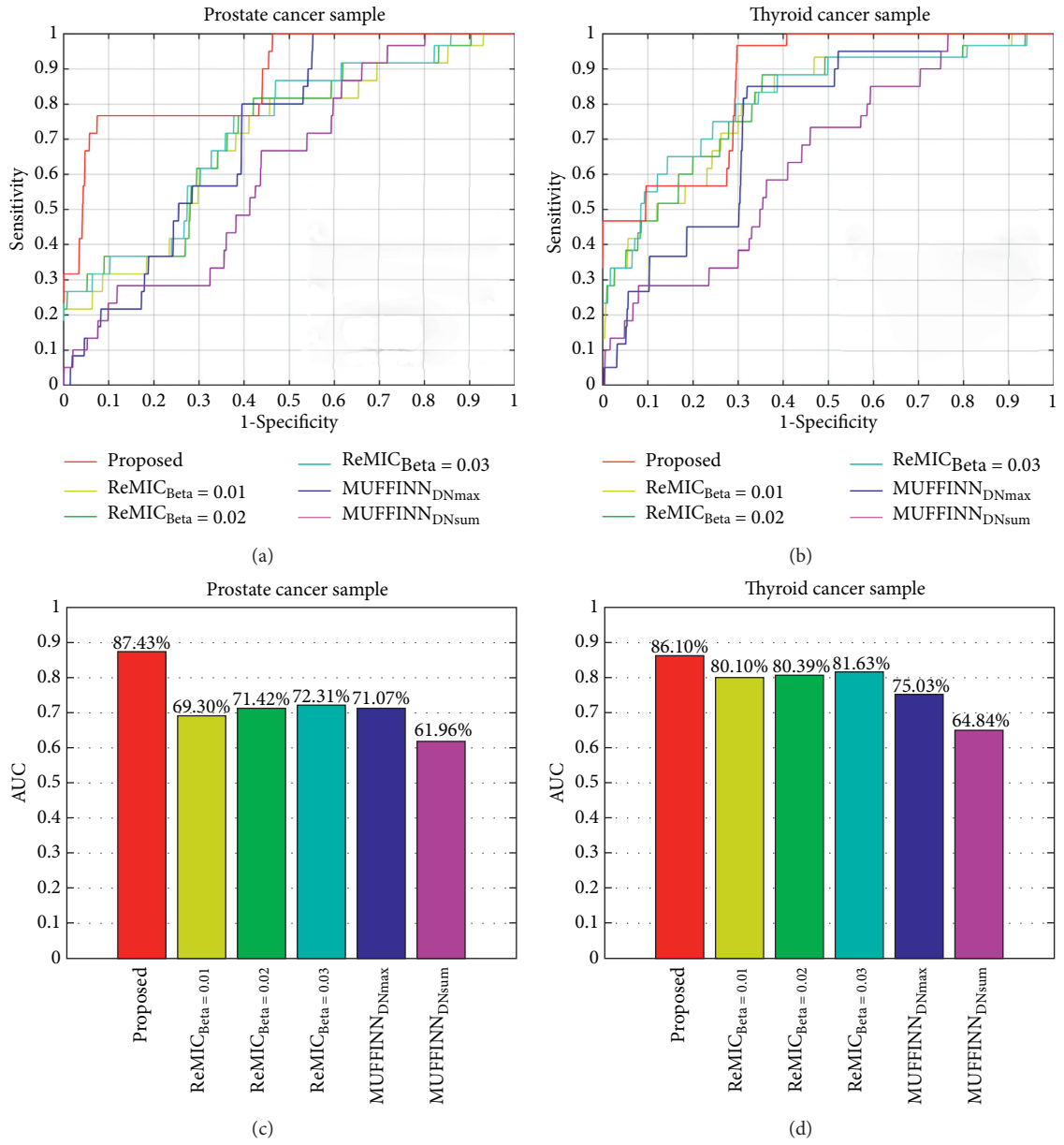

Figure 2: Performance evaluation of the compared methods on STRING network via ROC curves and AUC values. (a) ROC curve for prostate cancer. (b) ROC curve for thyroid cancer. (c) AUC for prostate cancer. (d) AUC values for thyroid cancer.

a distinct advantage of our method with input of iRefIndex network.

As for the comprehensive measurement AUC in the evaluation study, the assessment results for both prostate cancer (see Figure 3(c)) and thyroid cancer (see Figure 3(d)) also demonstrate that our proposed modularity subspace based concept learning model is superior to the other compared methods. For the results of prostate cancer, the AUCs obtained by MUFFINN-DNmax and MUFFINN-DNsum are 63.78% and 52.08%, respectively, and the AUCs yielded by different versions of ReMIC range in the interval from 65.79% to 67.57%. In contrast, our model achieves an AUC of 84.72%, and the value is the largest among those of the compared methods. Meanwhile, for the results of thyroid cancer, our model also surpasses the other competing methods with an AUC of 85.04%, where the AUCs of MUFFINN-DNmax, MUFFINN-DNsum, ReMIC (Beta = 0.01), ReMIC (Beta = 0.02), and ReMIC (Beta = 0.03) are 70.71%, 55.81%, 73.34%, 74.41%, and 75.15%, respectively. The ablation study for the two cancer types also shows the roles of different modules in our model (see Figure S2 in the Supplementary Materials). Therefore, we can conclude that, for iRefIndex network, our modularity subspace based concept learning model exhibits a superior performance compared to the other existing network based methods.

3.3. Functional Enrichment Analysis. To exploit the capability of cancer driver filling of our proposed modularity

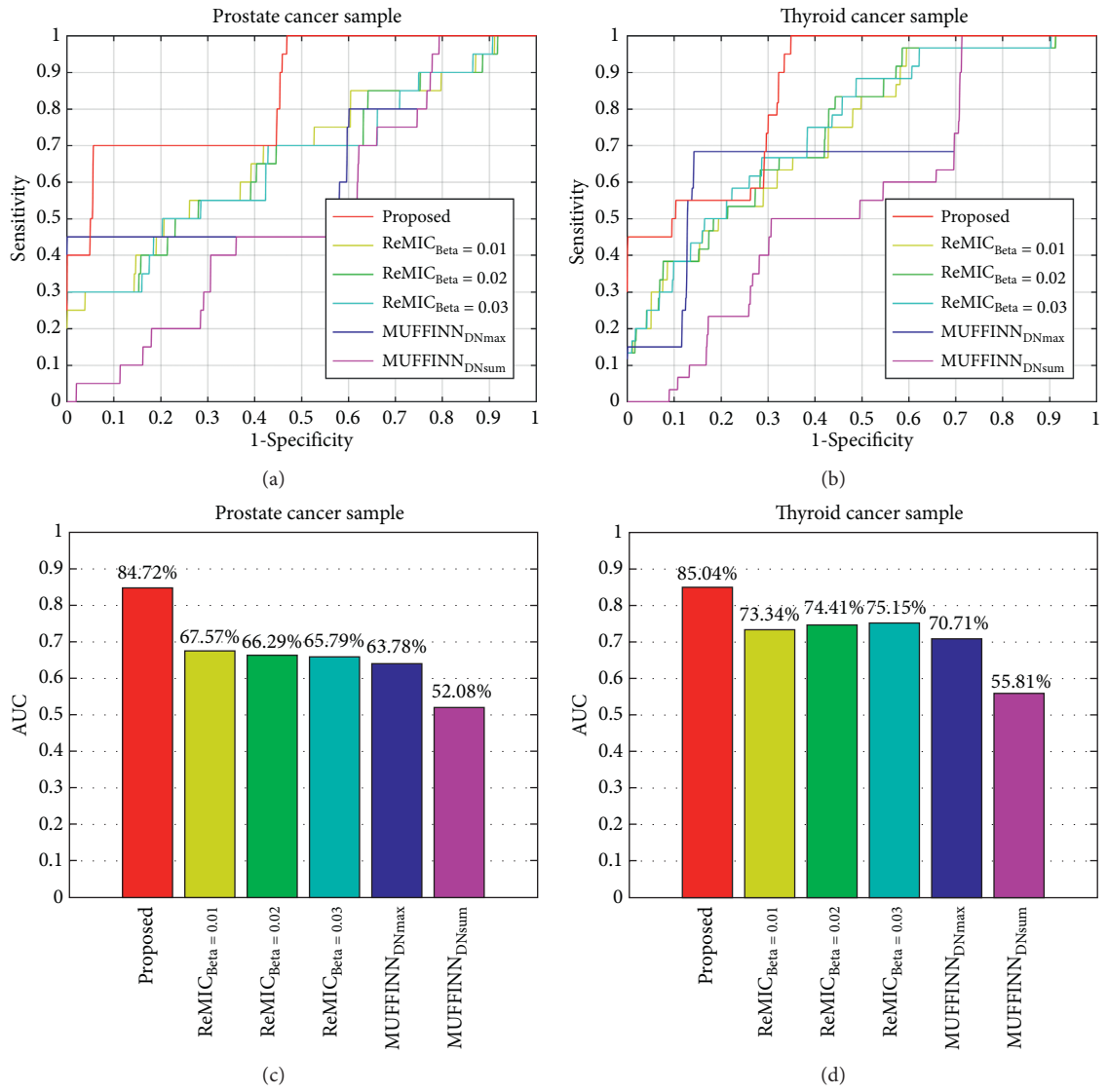

Figure 3: Performance evaluation of the compared methods on iRefIndex network via ROC curves and AUC values. (a) ROC curve for prostate cancer. (b) ROC curve for thyroid cancer. (c) AUC for prostate cancer. (d) AUC values for thyroid cancer.

subspace based concept learning model, we further examine the attribute filling results of our model on the data of the two types of cancers. When we apply the trained model on known drivers and mutation data of cancer samples, we can achieve a list of genes with the probabilities of driver attributes. Since the top ranked genes draw more attentions of researchers of tumorigenesis than the other genes, we further investigated the top twenty genes yielded by our model (see Table S1 in the Supplementary Materials for the predicted gene list). For the gene list of prostate cancer, we employ the functional enrichment analysis on STRING network to find the related specific biological functions [38]. Through the results of analysis, we can find that the set of genes obtained by our model are significantly enriched for a bunch of cancer related functions (see Table 1)). Detailly, the genes in the results for STRING network and prostate cancer data are highly involved in prostate cancer related functions, such as transcriptional misregulation in cancer, enzyme binding, sequence-specific DNA binding, and prostate cancer. Therefore, the gene list yielded by our model demonstrates significantly associations with known functions of prostate cancer.

At the same time, for the results of our model on STRING network with mutation data of thyroid cancer, we also adopt the functional enrichment analysis [38] on the predicted gene list (see Table S1 in the Supplementary Materials for the gene list) and exploit the highly related function terms (see Table 1). The results of enrichment analysis show that the top ranked function term is the pathway of thyroid cancer, indicating the affinity of the predicted genes to the corresponding cancer. The predicted genes are also significantly enriched for cancer related functions such as MAPK cascade, pathways in cancer, and central carbon metabolism in cancer, illustrating the strong relations between the gene list and cancer progresses. Meanwhile, when we employ functional enrichment analysis

Table 1: Functional enrichment analysis results of our proposed model on STRING network. (a) Enrichment results of prostate cancer. (b) Enrichment results of thyroid cancer.

a	Function term	p value
1	GO:0033148: positive regulation of intracellular estrogen receptor signaling pathway	$4.86E-05$
2	hsa05202: transcriptional misregulation in cancer	$1.42E-04$
3	GO:0019899: enzyme binding	$2.89E-04$
4	GO:0043565: sequence-specific DNA binding	$1.52E-03$
5	GO:0001077: transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding	$1.59E-03$
6	GO:0003700: transcription factor activity, sequence-specific DNA binding	$2.06E-03$
7	GO:0005634: nucleus	$5.58E-03$
8	GO:0060740: prostate gland epithelium morphogenesis	$7.48E-03$
9	GO:0060736: prostate gland growth	$8.55E-03$
10	hsa05215: prostate cancer	$9.83E-03$
b		
1	hsa05216: thyroid cancer	$4.21E-10$
2	GO:0000165: MAPK cascade	$1.49E-04$
3	hsa05200: pathways in cancer	$2.06E-04$
4	GO:0005801: cis-Golgi network	$7.75E-04$
5	hsa05219: bladder cancer	$1.84E-03$
6	hsa05213: endometrial cancer	$2.95E-03$
7	hsa05221: acute myeloid leukemia	$3.42E-03$
8	hsa05223: non-small cell lung cancer	$3.42E-03$
9	hsa04730: long-term depression	$3.91E-03$
10	hsa05230: central carbon metabolism in cancer	$4.44E-03$

on the result of iRefIndex network, we can also obtain similar phenomena for samples of both prostate cancer and thyroid cancer. The details of the gene lists (Table S1 in Supplementary Materials) and the enriched function terms (Table S2 in Supplementary Materials) on iRefIndex network also show high association with functions of the two types of cancers.

4. Discussion and Conclusions

In gene interaction network, the aberrations of a gene can influence the functions of it and its interacting genes, both of which contribute to the development of cancer. Although there are many successes achieved by previous methods based on network properties of gene nodes and network propagation of mutation frequencies, there are still obstacles in the task of cancer driver attribute filling of genes. Specifically, the number of mutation samples of cancer patients is rather small when compared with the large scale of network, and this phenomenon causes the issue of curse of dimensionality. The existence of driver genes without distinct network property and high propagation influence of neighbours also leads to the missing attributes of driver genes in results of previous methods. To tackle these obstacles, we propose a novel modularity subspace based concept learning model, which can learn the modularity features of gene nodes beyond the network neighbours and reduce the feature dimensions to circumvent the curse of dimensionality. When we evaluate the performance of our model and those of the other compared methods on two gene interaction networks from independent sources, we can observe a distinct advantage of our proposed model in the driver attribute filling task. The enrichment analysis also shows the high correlation between the results of our model and the cancer related functions.

To seek the potential explanations of the improvement achieved by our proposed modularity subspace based concept learning, we can mainly conclude the following: information compensation, modularity feature, and dimension reduction. The first explanation is that we compensate the information from both gene interaction network and mutation data of cancer samples, by which we can incorporate the advantage of fusing the information from distinct independent sources. The second explanation is that we explore the features of genes from network neighbours and mutation frequencies to the modularity memberships of genes, by which we can break the limitation of the existing features of gene node in interaction network. The third explanation is that we employ the idea of dimension reduction on both the gene interaction network data and the mutation data of cancer samples, by which we can circumvent the negative effect of curse of dimensionality caused by the small sample problem. Based on the three concerns in this study, the results yielded by our proposed model demonstrate a superior performance compared to the existing compared methods, indicating its effectiveness in the task of driver attribute filling of genes.

In future work, there is still some room for improvement of our proposed model. One promising improvement is compensating more information from the data beyond interaction network and genomic mutation data, that is, integrated information from multiomics data such as transcriptome [39, 40], epigenome [41], and proteome [42]. Another point is the consideration of effect from deleterious synonymous variants into the framework of our model, that is, regarding the mutation types with more

precise resolution [43, 44]. In addition to genes in coding regions, our model is also potentially applicable for the analysis of attribute filling of noncoding regions in bioinformatics [45]. Furthermore, incorporating the technique of recent advanced artificial intelligence is also a promising direction for improvement [46, 47]. The framework of our model also illustrates the potential in application in various fields beyond gene interaction network analysis, such as hydrological models [48, 49], catalytic activity models [50, 51], and spectroscopy analysis [52, 53]. Last but not least, incorporating cancer samples with larger size into the analysis of gene's roles in interaction network is also a promising orientation for the task of driver attribute filling of gene nodes [54].

In conclusion, our proposed modularity subspace based concept learning model is capable of effectively compensating the information of gene interaction network and cancer mutation data and reducing the feature dimensions to circumvent the curse of dimensionality resulting from small sample problem in the attribute concept learning of cancer driver gene. The effectiveness of attribute filling of gene nodes of our model has been systematically evaluated through the application on two interaction networks. Considering the distinct performances, our model shows a promising potential in the analysis of cancer driver genes from interaction network, facilitating the comprehensive understanding of tumorigenesis.

Data Availability

The data repositories and deposition codes are freely available at https://github.com/JianingXi/ModularitySubspaceConceptLearning.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank Dr. Zhenhua Yu for his helpful suggestions. This work was supported in part by the National Natural Science Foundation of China under grant nos. 61876145, 61973249, 61973250, 62003279, and 61901322 and in part by Education Department of Shaanxi Provincial Government (project nos. 19JC041 and 19JC038).

Supplementary Materials

Table S1: predicted gene lists of our proposed model on STRING network and iRefIndex network for data of prostate cancer samples and thyroid cancer samples. Table S2: functional enrichment analysis results of our proposed model on iRefIndex network. (a) Enrichment results of prostate cancer. (b) Enrichment results of thyroid cancer. Figure S1: ablation study on modules in our model for STRING network. The red-dashed lines represent the AUCs for cases of all modules, and the four bars denote the AUCs for cases of random ablation of three-fourths of modules. (a) Prostate cancer samples. (b) Thyroid cancer samples. Figure S2: ablation study on modules in our model for iRefIndex network. The red-dashed lines represent the AUCs for cases of all modules, and the four bars denote the AUCs for cases of random ablation of three-fourths of modules. (a) Prostate cancer samples. (b) Thyroid cancer samples. Supplementary text: details of hyperparameter settings of our proposed model. (*Supplementary Materials*)

References

[1] G. Wu, X. Feng, and L. Stein, "A human functional protein interaction network and its application to cancer data analysis," *Genome Biology*, vol. 11, no. 5, p. R53, 2010.

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

[3] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[4] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási, "Functional and topological characterization of protein interaction networks," *Proteomics*, vol. 4, no. 4, pp. 928–942, 2004.

[5] J. Li, K. Deng, X. Huang, and J. Xu, "Analysis and applications of location-aware big complex network data," *Complexity*, vol. 2019, 2019.

[6] Q. He, Z. Sun, K. Nie, and W. Chen, "Sky images enhancement under complex weather conditions," in *Proceedings of the 2016 in IEEE International Conference on Information and Automation (ICIA)*, pp. 1490–1495, Ningbo, China, July 2016.

[7] X. Kong, F. Xia, K. Ma, J. Li, and Q. Yang, "Discovering transit-oriented development regions of megacities using heterogeneous urban data," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 5, pp. 943–955, 2019.

[8] T. Cai, J. Li, A. S. Mian, T. Sellis, and J. X. Yu, "Target-aware holistic influence maximization in spatial social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, 2020.

[9] H. Zhang, X. Wang, H. Fan et al., "Anchor vertex selection for enhanced reliability of traffic offloading service in edge-enabled mobile P2P social networks," *Journal of Communications and Information Networks*, vol. 5, no. 2, pp. 217–224, 2020.

[10] B. Bhattacharjee, R. M. Jayadeepa, U. Talambedu et al., "Complex network and gene ontology in pharmacology approaches:mapping natural compounds on potential drug target colon cancer network," *Current Bioinformatics*, vol. 6, no. 1, pp. 44–52, 2011.

[11] E. Ramsahai, K. Walkins, V. Tripathi, and M. John, "The use of gene interaction networks to improve the identification of cancer driver genes," *Peer Journal*, vol. 5, Article ID e2568, 2017.

[12] S. Babaei, M. Hulsman, M. Reinders, and J. D. Ridder, "Detecting recurrent gene mutation in interaction network context using multi-scale graph diffusion," *BMC Bioinformatics*, vol. 14, no. 1, p. 29, 2013.

[13] A. Cho, J. E. Shim, E. Kim et al., "MUFFINN: cancer gene discovery via network analysis of somatic mutation data," *Genome Biology*, vol. 17, no. 1, pp. 1–16, 2016.

[14] M. S. Lawrence, P. Stojanov, P. Polak et al., "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, 2013.

[15] N. Cheng, M. Li, L. Zhao et al., "Comparison and integration of computational methods for deleterious synonymous mutation prediction," *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 970–981, 2020.

[16] Z. Yu, F. Du, R. Ban, and Y. Zhang, "SimuSCoP: Reliably simulate Illumina sequencing data based on position and context dependent profiles," *BMC Bioinformatics*, vol. 21, no. 1, pp. 1–18, 2020.

[17] Z. Yue, X. Chu, and J. Xia, "PredCID: prediction of driver frameshift indels in human cancer," *Briefings in Bioinformatics*, vol. 45, 2020.

[18] J. Xi, M. Wang, and A. Li, "Discovering mutated driver genes through a robust and sparse co-regularized matrix factorization framework with prior information from mRNA expression patterns and interaction network," *BMC Bioinformatics*, vol. 19, no. 1, p. 214, 2018.

[19] J. Xi, X. Yuan, M. Wang, A. Li, X. Li, and Q. Huang, "Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication," *Bioinformatics (Oxford, England)*, vol. 36, no. 6, pp. 1855–1863, 2020.

[20] M. E. Newman, "Spectral methods for community detection and graph partitioning," *Physical Review E*, vol. 88, no. 4, Article ID 042822, 2013.

[21] X. Wang, P. Cui, J. Wang et al., *Community Preserving Network Embedding*, AAAI, Menlo Park, CL, USA, 2017.

[22] J. B. Tenenbaum, "Bayesian modeling of human concept learning," *Advances in Neural Information Processing Systems*, vol. 14, pp. 59–68, 1999.

[23] Y. Jia, J. T. Abbott, J. L. Austerweil, T. Griffiths, and T. Darrell, "Visual concept learning: combining machine vision and bayesian generalization on concept hierarchies," *Advances in Neural Information Processing Systems*, vol. 21, pp. 1842–1850, 2013.

[24] Y.-A. Kim, D.-Y. Cho, P. Dao, and T. M. Przytycka, "MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types," *Bioinformatics*, vol. 31, no. 12, pp. i284–i292, 2015.

[25] F. Li, L. Gao, and B. Wang, "Detection of driver modules with rarely mutated genes in cancers," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 24, 2018.

[26] F. Li, L. Gao, P. Wang, and Y. Hu, "Identifying cancer specific driver modules using a network-based method," *Molecules*, vol. 23, no. 5, p. 1114, 2018.

[27] J. Xi, M. Wang, and A. Li, "Discovering potential driver genes through an integrated model of somatic mutation profiles and gene functional information," *Molecular BioSystems*, vol. 13, no. 10, pp. 2135–2144, 2017.

[28] J. Xi, A. Li, and M. Wang, "A novel unsupervised learning model for detecting driver genes from pan-cancer data through matrix tri-factorization framework with pairwise similarities constraints," *Neurocomputing*, vol. 296, pp. 64–73, 2018.

[29] A. Li, Q. Zang, D. Sun, and M. Wang, "A text feature-based approach for literature mining of lncRNA-protein interactions," *Neurocomputing*, vol. 20, pp. 73–80, 2016.

[30] D. Szklarczyk, A. Franceschini, S. Wyder et al., "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, no. 1, pp. D447–D452, 2015.

[31] S. Razick, G. Magklaras, and I. M. Donaldson, "iRefIndex: a consolidated protein interaction database with provenance," *BMC Bioinformatics*, vol. 9, no. 1, p. 405, 2008.

[32] A. Abeshouse, J. Ahn, R. Akbani et al., "The molecular taxonomy of primary prostate cancer," *Cell*, vol. 163, no. 4, pp. 1011–1025, 2015.

[33] N. Agrawal, R. Akbani, B. A. Aksoy et al., "Integrated genomic characterization of papillary thyroid carcinoma," *Cell*, vol. 159, no. 3, pp. 676–690, 2014.

[34] J. Gao, B. A. Aksoy, U. Dogrusoz et al., "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Science Signaling*, vol. 6, no. 269, p. pl1, 2013.

[35] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, "The COSMIC cancer gene census: describing genetic dysfunction across all human cancers," *Nature Reviews Cancer*, vol. 18, no. 11, pp. 696–705, 2018.

[36] F. Luo, M. Wang, Y. Liu, X.-M. Zhao, and A. Li, "Deepphos: prediction of protein phosphorylation sites with deep learning," *Bioinformatics*, vol. 35, no. 16, pp. 2766–2773, 2019.

[37] J. Yang, A. Li, Y. Li, X. Guo, and M. Wang, "A novel approach for drug response prediction in cancer cell lines via network representation learning," *Bioinformatics*, vol. 35, no. 9, pp. 1527–1535, 2019.

[38] D. W. Huang, B. T. Sherman, Q. Tan et al., "DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists," *Nucleic Acids Research*, vol. 35, no. suppl_2, pp. W169–W175, 2007.

[39] J. Shang, Q. Ding, S. Yuan, J.-X. Liu, F. Li, and H. Zhang, "Network analyses of integrated differentially expressed genes in papillary thyroid carcinoma to identify characteristic genes," *Genes*, vol. 10, no. 1, p. 45, 2019.

[40] L. Zhang, Q. Wang, L. Wang et al., "OSskcm: an online survival analysis webserver for skin cutaneous melanoma based on 1085 transcriptomic profiles," *Cancer Cell International*, vol. 20, pp. 1–8, 2020.

[41] W. Wei, X. Ji, X. Guo, and S. Ji, "Regulatory role of N6-methyladenosine (m6A) methylation in RNA processing and human diseases," *Journal of Cellular Biochemistry*, vol. 118, no. 9, pp. 2534–2543, 2017.

[42] Y. Guo, W. Ning, P. Jiang et al., "GPS-PBS: a deep learning framework to predict phosphorylation sites that specifically interact with phosphoprotein-binding domains," *Cells*, vol. 9, no. 5, p. 1266, 2020.

[43] P. Wen, P. Xiao, and J. Xia, "dbDSM: a manually curated database for deleterious synonymous mutations," *Bioinformatics*, vol. 32, no. 12, pp. 1914–1916, 2016.

[44] F. Shi, Y. Yao, Y. Bin, C.-H. Zheng, and J. Xia, "Computational identification of deleterious synonymous variants in human genomes using a feature-based approach," *BMC Medical Genomics*, vol. 12, no. 1, pp. 81–88, 2019.

[45] W. Fan, J. Shang, F. Li et al., "IDSSIM: an lncRNA functional similarity calculation model based on an improved disease semantic similarity method," *BMC Bioinformatics*, vol. 21, no. 1, pp. 1–14, 2020.

[46] X. Li, M. Chen, F. Nie, and Q. Wang, *Locality Adaptive Discriminant Analysis*, IJCAI, Montreal, Canada, 2017.

[47] X. Li, M. Chen, F. Nie, and Q. Wang, *A Multiview-Based Parameter Free Framework for Group Detection*, Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CL, USA, 2017.

[48] R. Sun, H. Yuan, and X. Liu, "Effect of heteroscedasticity treatment in residual error models on model calibration and prediction uncertainty estimation," *Journal of Hydrology*, vol. 554, pp. 680–692, 2017.

[49] R. Sun, F. Hernández, X. Liang, and H. Yuan, ""A calibration framework for high-resolution hydrological models using a

multiresolution and heterogeneous strategy," *Water Resources Research*, vol. 56, no. 8, 2020.

[50] W. Gao, C. Wang, F. Ma, and D. Wen, "Highly active electrocatalysts of CeO2 modified NiMoO4 nanosheet arrays towards water and urea oxidation reactions," *Electrochimica Acta*, vol. 320, Article ID 134608, 2019.

[51] W. Gao, F. Ma, C. Wang, and D. Wen, "Ce dopant significantly promotes the catalytic activity of Ni foam-supported Ni3S2 electrocatalyst for alkaline oxygen evolution reaction," *Journal of Power Sources*, vol. 450, Article ID 227654, 2020.

[52] L. Zou, X. Yu, M. Li, M. Lei, and H. Yu, "Nondestructive identification of coal and gangue via near-infrared spectroscopy based on improved broad learning," *IEEE Transactions on Instrumentation and Measurement*, vol. 89, no. 9, 2020.

[53] L. Zou, X. Zhu, C. Wu, Y. Liu, and L. Qu, "Spectral-Spatial exploration for hyperspectral image classification via the fusion of fully convolutional networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 659–674, 2020.

[54] The ICGC/TCGA, "Pan-cancer analysis of whole genomes consortium. pan-cancer analysis of whole genomes," *Nature*, vol. 578, no. 7793, p. 82, 2020.