2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669479

Predicting Hepatoma-Related Genes Based on Representation Learning of PPI network and Gene Ontology Annotations

Tao Wang
School of Computer Science
Northwestern Polytechnical University
Xi'an, China
twang@nwpu.edu.cn

Zhiyuan Shao
Honor College
Northwestern Polytechnical University
Xi'an, China
2018300081@mail.nwpu.edu.cn

Yifu Xiao
School of Computer Science
Northwestern Polytechnical University
Xi'an, China
1529641817@qq.com

Xuchao Zhang
Honor College
Northwestern Polytechnical University
Xi'an, China
2019300046@mail.nwpu.edu.cn

Yitian Chen
Honor College
Northwestern Polytechnical University
Xi'an, China
oscarchen@mail.nwpu.edu.cn

Binze Shi
Honor College
Northwestern Polytechnical University
Xi'an, China
chenqi0124@gmail.com

Siyu Chen
Honor College
Northwestern Polytechnical University
Xi'an, China
2019300081@mail.nwpu.edu.cn

Yuxian Wang
School of Computer Science
Northwestern Polytechnical University
Xi'an, China
yxwang@mail.nwpu.edu.cn

Jiajie Peng*
School of Computer Science
Northwestern Polytechnical University
Xi'an, China
jiajiepeng@nwpu.edu.cn

Xuequn Shang*
School of Computer Science
Northwestern Polytechnical University
Xi'an, China
Shang@nwpu.edu.cn

***Abstract*—Hepatoma is the most common type of primary liver cancer with a high mortality rate in the world. The genetic causes of the disease pathology remain largely unknown. Effective discovery of the genes associated with hepatoma has become important in disease prevention, early diagnosis, and therapeutic treatments. With the developments of molecular networks, graph-based methods have been tremendously successful in predicting disease genes based on the hypothesis of guilt-by-association. Network representation learning (NRL) techniques have accelerated disease gene discovery in recent years because of their powerful network feature extraction ability. However, the current network representation learning-based methods for disease gene discovery did not consider the gene features derived from gene ontology annotations, which apriori group genes with similar functions. To fill this gap, here we propose a novel framework to predict hepatoma-related genes based on representation learning from both protein-protein interactions (PPI) network and gene ontology annotations. Our framework has three steps: learning features from PPI network and gene ontologies using NRL techniques, integrating different features based on autoencoder, predicting hepatoma-related genes using machine learning classifiers. Experiments have demonstrated that our framework could accurately predict hepatoma-related genes with AUROC and AUPRC reaching 0.93 and 0.94, respectively. Compared with other methods using only single representation features, our framework also shows superior performance on hepatoma gene prediction.**

***Index Terms*—Hepatoma, disease gene prediction, network representation learning, gene ontology, PPI network**

I. Introduction

Hepatoma is also known as hepatocellular carcinoma (HCC) and hepatocarcinoma, which is a widespread and destructive disease with a high mortality rate around the world. Although the viral and environmental risk factors for the development of HCC have been elucidated, the genetic causes of the disease pathology and malignant transformation of liver cells are still unclear [1]. It has been of high value to reveal the underlying molecular mechanisms of the disease in terms of prevention, early diagnosis, and treatment of hepatoma [2]. In the past decade, the advances of high-throughput sequencing technology have enabled the discovery of many genetic variants,

* to whom correspondence should be addressed: Jiajie Peng (jiajiepeng@nwpu.edu.cn) or Xuequn Shang (Shang@nwpu.edu.cn)

This research work was funded by the National Natural Science Foundation of China (Nos. 6210071334, 62072376,U1811262), the China National Key R&D Program (No. 2020AAA0108500), the National Natural Science Foundation of China (No. 6207071829), the Fundamental Research Funds for the Central Universities (No. G2021KY05112).

genes, and dysregulated signaling pathways associated with hepatoma [3], [4]. However, there is still a poor survival rate associated with HCC due to a lack of efficient therapeutic targets. The discovery of novel genes could contribute to a better understanding of the disease and bolster the therapeutic pipelines.

There have been intense efforts to discover disease-related targets, including genes [5], [6], mutations [7]–[9], microRNAs [10], modules [11], [12],and risk exposures [13], [14]. Traditional wet-lab methods relying on hypothesis and experiments have made great contributions in paving the way for disease therapeutics. However, multiple limitations have been slowing down the process, such as cost and low efficiency. In recent years, with the development of high-throughput sequencing techniques, a huge volume of omics data has been accumulated and analyzed [15]–[18], which enables the new paradigm of disease gene discovery based on the integration of big data, computational methods, and biological evaluation [19], [20]. Compared with traditional experiments, the computation-guided methods are more economical and time-saving [21].

A common approach towards disease gene prediction is through molecular networks [22]. The network-based methods are mainly based on the hypothesis of guilt-by-association (GBA) [23]. Based on the hypothesis, the genes closely connected with disease genes in a biological network, such as protein-protein interaction network (PPI), have a higher likelihood of disease-associated disease than those far from known disease genes. Many methods have been proposed based on the topological structure of a biological network. One of the fundamental tasks in those methods is to measure the similarity between two genes in the network. Oti et al. proposed a method based on neighbor counting, which directly calculated the number of disease-related genes in the neighborhood of a gene in the PPI network [24]. Krauthammer et al. adapted the shortest path-based method to evaluate the gene similarities and predicted novel disease genes based on their shortest distances to known disease genes [25]. However, this method did not consider the number of common neighbors that two genes share. In addition, both of the above methods only consider the local structural features while ignoring global structures. In order to capture the global structural information, methods based on random walks have been proposed, which define the node similarity incorporating both local and higher-order neighborhood information. For example, random walk with restart (RWR) [26] allows us to measure the proximity between nodes in a network very efficiently. It starts from the known disease genes, then randomly walks on the network multiple times. Each random walk will have a certain probability of jumping back to the starting point (disease gene). After random walks, the candidate genes with more visits (from disease genes) are likely to be disease genes. The similarities measured in this way will consider richer topological information, such as (1) multiple connections and paths between a pair of nodes, (2) the directions of these connections, and (3) degrees of the nodes on the path.

With the accumulation of big data, deep learning-based methods have been widely applied in bioinformatics and medical applications [27]–[33]. In particular, multiple network representation learning methods have been proposed and applied to disease gene prediction, such as DeepWalk [34], Node2vec [35] and LINE [36], etc. Most of these cutting-edge network representation methods use the random walk technique to measure node similarities and can effectively obtain the local and global characteristics of the network. Due to the superior graph representation ability, these network embedding-based disease gene prediction methods have achieved excellent performance in multiple scenarios. However, the current disease gene prediction methods based on network representation learning have not incorporated gene ontology (GO) annotations [37]. Inspired by the recent advances in the representation learning of GO terms and related genes, in this paper, we aim to propose a novel framework by integrating molecular networks and GO annotations to learn gene features and predict HCC associated genes.

In this paper, we propose a novel framework to predict hepatoma-related genes. The advantage of our framework is to integrate gene features from both the PPI network and the gene ontology annotations. Our method can be divided into three steps: feature extraction, feature fusion, and disease gene prediction. In the first step, we use the cutting-edge network representation learning methods to extract gene features from the PPI network and gene ontology annotations. Next, we use a stacked autoencoder [38] to project the two features into the same feature space, namely feature fusion. Finally, we apply widely used machine learning classifiers to predict hepatoma-related genes. Experiments have demonstrated that our framework could accurately predict hepatoma-related genes (mean AUROC = 0.93, mean AUPRC = 0.94 in five-fold cross-validation). Our framework has better performance than traditional methods using only single representation features. Furthermore, our framework has strong potential to be applied to other diseases.

II. Methods

A. Overview

The framework has three steps: (1) feature extraction, where the cutting-edge network representation learning methods are applied to extract gene features from both PPI network and gene ontology annotations. (2) feature fusion, where stacked autoencoder [38] is used to project the two features from different space into the same feature space. And (3) disease gene prediction, where widely-used machine learning classifiers are adapted to predict hepatoma-related genes. The workflow has been illustrated in Figure 1.

B. Representation Learning From PPI Network

In this work, we try to used four cutting-edge methods for the purpose of network representation learning: DeepWalk [34], LINE [36], Node2vec [35], and RWR [26]. The method with the best prediction performance will be integrated

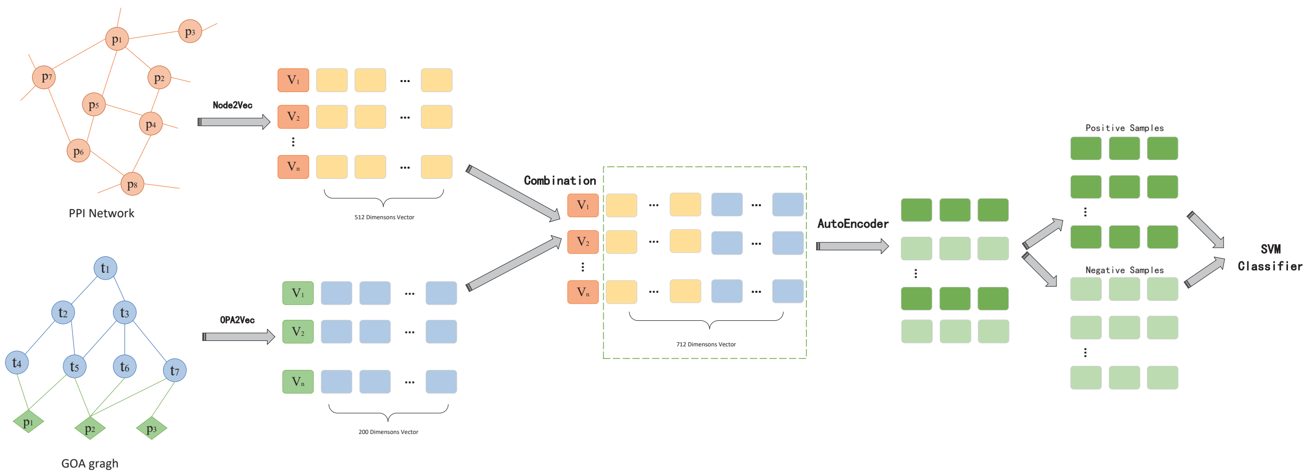

Fig. 1. Workflow of predicting genes associated with hepatoma. t_n denotes a go term, p_n denotes a protein and v_n denotes its k-dimensional vector.

into our framework. We will briefly introduce these methods in the following sections.

1) DeepWalk: DeepWalk [34] mainly has two steps. First, it employs random walk multiple times for each node to obtain a corpus of random walk paths with fixed length. Second, DeepWalk applies the Skip-gram neuron network to learn the node embedding, which is similar to Word2vec. The intuition of the network representation is that two nodes that are closer to each other in the network should have similar feature vectors in the feature space.

2) Node2vec: Node2vec [35] borrows the idea of DeepWalk, except that Node2vec proposes a novel biased random walk process to generate the corpus of vertices sequences [35]. In particular, Node2vec introduces two parameters p and q to control the random walk process. Assuming that the last vertex random walk visited is u, the current visiting vertex is v, v is connected with u, x_1, x_2, x_3, and x_1 is connected with u, the probability of v moving to u (back to the previous vertex is $\frac{1}{p}$, the probability of moving to x_1 (those three vertices are interconnected) is 1, and the probability of moving to x_2, x_3 (other vertices) is $\frac{1}{q}$. The transport probability is shown in equation 1. Using these two parameters to control the direction of random walk affects whether the random walk goes broader or deeper in the network. When $p = 1$ and $q = 1$, the walk mode is equivalent to the random walk in DeepWalk.

$$\alpha_{pq}(u,x) = \begin{cases} \dfrac{1}{p}, & d_{ux} = 0; \\ 1, & d_{ux} = 1; \\ \dfrac{1}{q}, & d_{ux} = 2. \end{cases} \tag{1}$$

3) LINE: LINE [36] considers two types of similarities: first-order similarity and second-order similarity. The first-order similarity describes the local similarity between paired vertices in the graph. If there is a direct edge between u and v, the edge's weight is the first-order similarity of two vertices. If there is no direct edge between two vertices, the first-order similarity is 0. The joint probability between vertices v_i and v_j is defined as:

$$p_1(v_i, v_j) = \frac{1}{1 + exp(-\vec{u}_i^T * \vec{u}_j)}$$

The empirical probability is defined as:

$$\hat{p}_1(i,j) = \frac{w_{ij}}{\sum_{(i,j)\in E} w_{ij}}$$

The objective function is:

$$O_1 = d(\hat{p}_1(.,.), p_1(.,.))$$

After using KL divergence and ignoring the constant term:

$$O_1 = -\sum\nolimits_{(i,j)\in E} w_{ij} log p_1(v_i, v_j)$$

The second-order similarity describes that two vertices have multiple common adjacent nodes, reflecting the global similarity. The second-order similarity is the number of identical neighbors of two vertices. If there is no same neighbor vertex between u and v, the second-order similarity is 0.

Under the condition of current vertex v_i, the probability of generating neighbor vertex v_j is

$$p_2(v_j|v_i) = \frac{exp(\vec{u'}_i^T \vec{u}_i)}{\sum_{k=1}^{|V|} exp(\vec{u'}_k^T \vec{u}_i)}$$

And the objective function after optimization is:

$$O_2 = -\sum\nolimits_{(i,j)\in E} w_{ij} log p_2(v_i|v_j)$$

LINE algorithm adopts an objective function that retains the first-order similarity and the second-order similarity, and adopts negative sampling optimization. The probability of each edge being sampled is proportional to the weight of the edge, which reduces the amount of calculation when calculating the second-order similarity. LINE can be applied to various types of networks and large networks. However, some vertices have few adjacent vertices, which leads to insufficient learning of embedding vector and insufficient utilization of high-order information.

4) Random Walk with Restart (RWR): The framework of RWR [39] is also similar to DeepWalk, but it has a unique design on transportation probabilities. RWR has a restart probability r, that is, the probability of moving to the starting vertex when the current vertex walks randomly. And $(1-r)$ is the probability of accessing other neighbor vertices. Restarting random walk can capture various relationships between the two vertices and the overall structure information of the graph.

C. Representation Learning From Gene Ontologies

In addition to the feature vectors generated from PPI, we employed OPA2Vec (Ontologies Plus Annotations to Vectors) to obtain the other feature vectors representation from Gene Ontologies. Through OPA2Vec, GO (Gene Ontology) annotation and the content formalized in the Web Ontology Language (OWL) are combined. Elk is used as a semantic reasoner to generate the OPA2Vec corpus. According to this pattern, we selected Molecular Function (MF), Cellular Component (CC), and Biological Processes(BP) as input networks, which enables our method to better results. Prepared Word2Vec skip-gram model then is applied on the processed corpus to generate a set of feature vectors representation.

GO (Gene Ontology) annotation included in Gene Ontologies depicts the ontology classes, instances, and their relation in various aspects, which enriches the corpus to be trained, and the formal logic axioms that characterize classes and instances in Gene Ontologies can be applied on Onto2Vec algorithm. However, the words that possess linguistic meanings in the real world cannot easily be obtained through an ontology alone. Therefore, OPA2Vec pre-trained the Word2Vec skip-gram model on full-text biomedical articles to assign natural language words semantics based on their usages in the context of biomedical text. Elk is chosen as OWL reasoner due to its polynomial complexity, which enables our method to be applied to more complicated gene ontologies.

D. Combining Features from PPI Network and Gene Ontologies using Autoencoder

After nonlinear features are learned from PPI network and gene ontology annotations respectively, for each gene, we first combine the feature vectors by simple concatenation. Next, we use autoencoder to perform feature fusion, which can produce more compact and higher-level feature sets. At the same time, the autoencoder can be used to reduce feature dimension and remove noise.

An autoencoder is a neural network consisting of two parts: the encoder and the decoder. The encoder compresses the source features into hidden feature space, and the decoder tries to resemble the inputs as closely as possible. The entire network is trained to minimize the error between the input of the encoder and the output of the decoder. Our autoencoder neural network comprises three layers: an input layer in which the original feature vectors are placed; one hidden layer where the vectors encoded in lower dimensions are stored; an output layer where the reconstructed feature vectors are represented. The weights of the hidden layer after training this network will represent the final lower dimensional representations.

E. Disease Gene Prediction

After extracting low-dimensional gene feature vectors, we use three traditional classifiers to predict disease-related genes of hepatoma: Support Vector Machine (SVM), Random Forest (RF), and Logistics Regression (LR). These three classifiers are widely used in many classification tasks due to their stability, simplicity, and effectiveness. We will incorporate the one with the best performance in our task into our framework. We use five-fold cross-validation to evaluate the prediction performance, and use ROC-AUC, PR-AUC, and F1-score as the criteria to evaluate different methods.

III. Results and Discussion

A. Datasets

In the experiment, gene features mainly come from two data sets: the Protein-protein interaction Network(PPI), which can be downloaded at https://string-db.org/cgi/input.pl and Gene Ontology in OWL format, which can be downloaded at http://geneontology.org/docs/download-ontology/.

PPI Network has 13460 protein nodes and 141296 interactions. The Go protein annotations we used have 5360 protein nodes with 116701 go terms. While we combined the features extracted from the two datasets, we selected 4,497 pairs of common protein genes as a population sample. In order to verify the effectiveness of our method for Hepatom prediction, 43 genes associated with hepatoma were selected from DisGeNet databased [40].

B. Prediction Performance using Different Features

To evaluate the effect of these features from different datasets on hepatoma gene prediction and the validity of the combined features, we compared three combinations extracting features: using Features from PPI Network only; using Features from Gene Ontologies only; using Combined Features from both PPI network and Gene Ontologies. In addition to the different ways of obtaining feature vectors, the remaining steps of the experiment are consistent with the workflow in Figure 1.

We use Node2vec to extract features from the PPI network and Opa2vec to extract features from Gene Ontologies. To get the combined features, we copy the 200-dimension feature

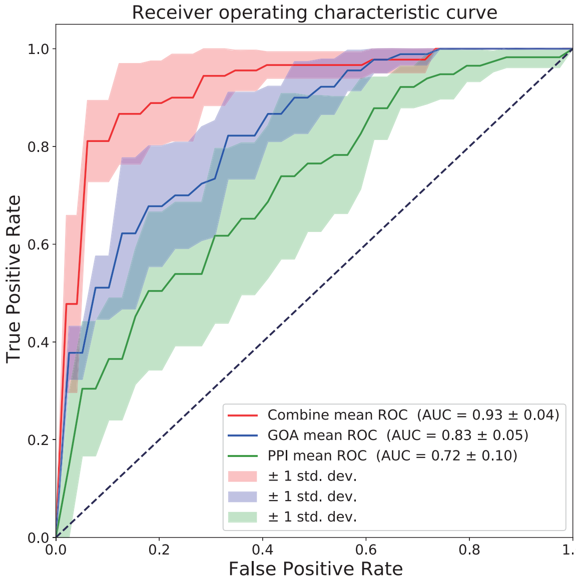

Fig. 2. Receiver operating characteristic (ROC) curve with different features. The Red line represents the ROC curve of combined-feature, the green line represents the ROC curve of features from PPI network only, the blue line represents the ROC curve of features from gene ontologies only, shadow means the standard deviation of each curve.

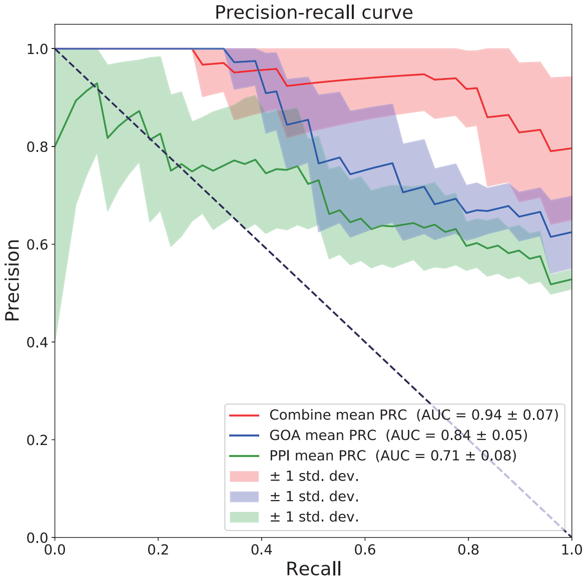

Fig. 3. Precision-recall curve with different features. The Red line represents the Precision-recall curve of combined-feature, the green line represents the Precision-recall curve of features from PPI network only, the blue line represents the Precision-recall curve of features from gene ontologies only, shadow means the standard deviation of each curve.

vectors extracted from Gene Ontologies and paste them after the 512-dimension feature vectors with the same GeneID extracted from the PPI network.

We plotted receiver operating characteristic (ROC) curve and Precision-Recall Curve based on the experimental results, which are shown in Figure 1, 2. We trained our model with the method of 5-fold cross-validation, averaged the results of five training sessions, and added standard deviation to indicate the fluctuation range of results. According to Figure 2, the training results obtained by using splined features are significantly better than those obtained by using a feature alone. Its mean AUROC value can reach 0.94, and its average AUPRC value can reach 0.94. It shows that the combination of features extracted from these two datasets is of great help in disease prediction.

C. Prediction Performance using Different Representation Methods

Representation methods have a significant influence on the extracted features, so it is necessary to compare the performance of different representation methods on data sets. When extracting gene characteristics of the PPI network, we compared 4 representation methods: Random walk with restart (RWR), Node2vec, DeepWalk, LINE. Four groups of parallel experiments are carried out separately. The only difference between them was the method of feature vector extraction. We uniformly use the SVM classifier to verify the quality of different methods.

The experimental results are shown in Table 1. We find that Node2vec performs better than other algorithms. The AUROC and AUPRC are as high as 0.95 and 0.89, and the accuracy rate is 0.87. Next to Node2vec is DeepWalk, which also has an AUROC value of 0.90, and the LINE and RWR algorithms perform relatively poorly. So we ended up with the Node2vec method.

D. Prediction Performance using Different Classifiers

The last step of our method is to use classifiers for prediction. In this case, different classifiers will also have

TABLE I
PERFORMANCE USING DIFFERENT REPRESENTATION METHODS

Representation methods	AUPRC	AUROC	F1	Acc
Node2vec	**0.88969(0.01775)**	**0.94707(0.01904)**	**0.83206(0.03651)**	**0.86667(0.05666)**
DeepWalk	0.81461(0.05121)	0.90327(0.05503)	0.61787(0.13170)	0.81111(0.09027)
LINE	0.68665(0.04881)	0.78498(0.04669)	0.64248(0.04106)	0.66667(0.03514)
RWR	0.59678(0.02991)	0.65990(0.05099)	0.53513(0.10728)	0.54444(0.06479)

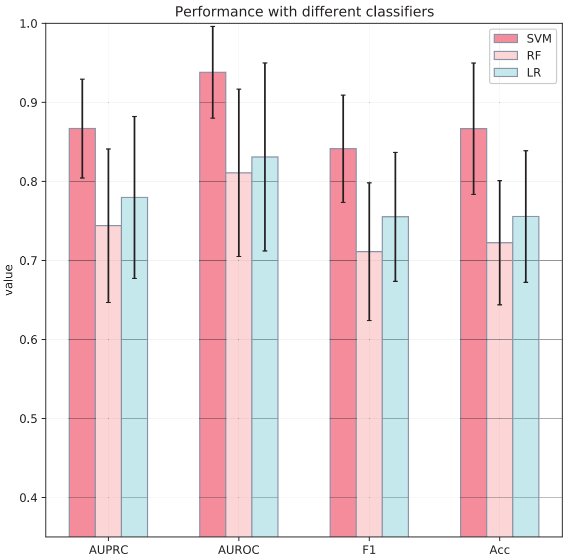

Fig. 4. The experimental results of using different classifiers. SVM:Support Vector Machines; LR: Logistic Regression; RF: Random Forest.

a great impact on the results. We compared three classical classifiers: SVM, Random Forest and Logistic Regression. Node2vec is used to extract PPI network features, and OPA2VEC is used to extract GO Ontologies features. Through Auto Encoder, a 712-dimension feature matrix is sent to the classifier. 43 genes related to hepatoma are selected as positive samples, and 43 from the remaining genes are selected randomly as negative samples.

The detailed results are shown in Figure 4. SVM classifier has the best prediction effect in our model, with the AUROC value of 0.94 and accuracy rate of 0.87, LR classifier has the second-best performance, and RF performs worst.

IV. Conclusion

In this work, we propose a novel disease gene prediction framework for hepatoma, which is a severe liver cancer with a high mortality rate but unclear genetic causes. Our framework first extracts gene features from both protein-protein interaction network and gene ontology annotations based on graph embedding methods. Then, we project the gene features from different feature spaces (PPI network and gene ontology) into the same feature space through an autoencoder. Finally, we predict the hepatoma-related genes through SVM. The advantage of our method is to integrate gene features learned from the molecular network with gene features learned from gene ontologies. Experiments have demonstrated that our framework could accurately predict hepatoma-related genes with AUROC and AUPRC reaching 0.93 and 0.94, respectively. And experiments also proved that our dual-feature-based method has superior prediction performance in the task of hepatoma-related gene discovery. In the future, we will try to use more cutting-edge deep learning-based network representation methods and end-to-end models to improve prediction accuracy further.

References

[1] M. A. Buendia, "Genetics of hepatocellular carcinoma," in *Seminars in cancer biology*, vol. 10, no. 3. Elsevier, 2000, pp. 185–200.

[2] J. Balogh, D. Victor III, E. H. Asham, S. G. Burroughs, M. Boktour, A. Saharia, X. Li, R. M. Ghobrial, and H. P. Monsour Jr, "Hepatocellular carcinoma: a review," *Journal of hepatocellular carcinoma*, vol. 3, p. 41, 2016.

[3] T. Chiba, O. Yokosuka, K. Fukai, Y. Hirasawa, M. Tada, R. Mikata, F. Imazeki, H. Taniguchi, A. Iwama, M. Miyazaki *et al.*, "Identification and investigation of methylated genes in hepatoma," *European journal of cancer*, vol. 41, no. 8, pp. 1185–1194, 2005.

[4] Z.-S. Niu, X.-J. Niu, and W.-H. Wang, "Genetic alterations in hepatocellular carcinoma: An update," *World journal of gastroenterology*, vol. 22, no. 41, p. 9069, 2016.

[5] J. Peng, Y. Wang, J. Guan, J. Li, R. Han, J. Hao, Z. Wei, and X. Shang, "An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction," *Briefings in Bioinformatics*, 2021.

[6] J. Peng, H. Xue, Z. Wei, I. Tuncali, J. Hao, and X. Shang, "Integrating multi-network topology for gene function prediction using deep neural networks," *Briefings in bioinformatics*, vol. 22, no. 2, pp. 2096–2105, 2021.

[7] L. Cheng, X. Han, Z. Zhu, C. Qi, P. Wang, and X. Zhang, "Functional alterations caused by mutations reflect evolutionary trends of SARS-CoV-2," *Brief Bioinform*, vol. 22, no. 2, pp. 1442–1450, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/33580783

[8] T. Wang, Y. Liu, J. Ruan, X. Dong, Y. Wang, and J. Peng, "A pipeline for RNA-seq based eQTL analysis with automated quality control procedures," *BMC bioinformatics*, vol. 22, no. 9, pp. 1–18, 2021.

[9] T. Wang, Q. Peng, B. Liu, X. Liu, Y. Liu, J. Peng, and Y. Wang, "eQTLMAPT: fast and accurate eQTL mediation analysis with efficient permutation testing approaches," *Frontiers in Genetics*, vol. 10, p. 1309, 2019.

[10] J. Peng, W. Hui, Q. Li, B. Chen, J. Hao, Q. Jiang, X. Shang, and Z. Wei, "A learning-based framework for miRNA-disease association identification using neural networks," *Bioinformatics*, vol. 35, no. 21, pp. 4364–4371, 2019.

[11] J. Peng, J. Lu, X. Shang, and J. Chen, "Identifying consistent disease subnetworks using dnet," *Methods*, vol. 131, pp. 104–110, 2017.

[12] T. Wang, Q. Peng, B. Liu, Y. Liu, and Y. Wang, "Disease Module Identification Based on Representation Learning of Complex Networks Integrated From GWAS, eQTL Summaries, and Human Interactome," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 418, 2020.

[13] L. Cheng, H. Zhuang, H. Ju, S. Yang, J. Han, R. Tan, and Y. Hu, "Exposing the Causal Effect of Body Mass Index on the Risk of Type 2 Diabetes Mellitus: A Mendelian Randomization Study," *Front Genet*, vol. 10, p. 94, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30891058

[14] L. Cheng, H. Zhao, P. Wang, W. Zhou, M. Luo, T. Li, J. Han, S. Liu, and Q. Jiang, "Computational methods for identifying similar diseases," *Molecular Therapy-Nucleic Acids*, 2019.

[15] Y. Chu, M. Teng, and Y. Wang, "Simulating genetically heterozygous genomes in the tumour tissue according to its clonal evolution history," *International Journal of Computational Biology and Drug Design*, vol. 12, no. 2, pp. 143–152, 2019.

[16] Y. Chu, C. Nie, and Y. Wang, "A pipeline for reconstructing somatic copy number alternation's subclonal population-based next-generation sequencing data," *Frontiers in genetics*, vol. 10, p. 1374, 2020.

[17] Y. Chu, Z. Wang, R. Wang, N. Zhang, J. Li, Y. Hu, M. Teng, and Y. Wang, "Wdnfinder: a method for minimum driver node set detection and analysis in directed and weighted biological network," *Journal of bioinformatics and computational biology*, vol. 15, no. 05, p. 1750021, 2017.

[18] Y. Chu, M. Teng, and Y. Wang, "Modeling and correct the gc bias of tumor and normal wgs data for scna based tumor subclonal population inferring," *BMC bioinformatics*, vol. 19, no. 5, pp. 79–87, 2018.

[19] T. Wang, Y. Liu, Q. Yin, J. Geng, J. Chen, X. Yin, Y. Wang, X. Shang, C. Tian, Y. Wang, and Others, "Enhancing discoveries of molecular QTL studies with small sample size using summary statistic imputation," *Briefings in Bioinformatics*, 2021.

[20] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay *et al.*, "Impact of high-throughput screening in biomedical research," *Nature reviews Drug discovery*, vol. 10, no. 3, pp. 188–195, 2011.

[21] J. Peng, K. Bai, X. Shang, G. Wang, H. Xue, S. Jin, L. Cheng, Y. Wang, and J. Chen, "Predicting disease-related genes using integrated biomedical networks," *BMC genomics*, vol. 18, no. 1, pp. 1–11, 2017.

[22] M. Li, X. Wu, J. Wang, and Y. Pan, "Towards the identification of protein complexes and functional modules by integrating ppi network and gene expression data," *BMC bioinformatics*, vol. 13, no. 1, pp. 1–15, 2012.

[23] J. Peng, J. Guan, and X. Shang, "Predicting parkinson's disease genes based on node2vec and autoencoder. front genet. 2019; 10: 226," 2019.

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

[25] M. Krauthammer, C. A. Kaufmann, T. C. Gilliam, and A. Rzhetsky, "Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in alzheimer's disease," *Proceedings of the National Academy of Sciences*, vol. 101, no. 42, pp. 15 148–15 153, 2004.

[26] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in *Sixth international conference on data mining (ICDM'06)*. IEEE, 2006, pp. 613–622.

[27] B. Hosseini, R. Montagne, and B. Hammer, "Deep-aligned convolutional neural network for skeleton-based action recognition and segmentation," *Data Science and Engineering*, vol. 5, no. 2, pp. 126–139, 2020.

[28] S. Siuly and Y. Zhang, "Medical big data: neurological diseases diagnosis through medical data analysis," *Data Science and Engineering*, vol. 1, no. 2, pp. 54–64, 2016.

[29] B. Fjukstad and L. A. Bongo, "A review of scalable bioinformatics pipelines," *Data Science and Engineering*, vol. 2, no. 3, pp. 245–251, 2017.

[30] W. Pan, Z. Li, Y. Zhang, and C. Weng, "The new hardware development trend and the challenges in data management and analysis," *Data Science and Engineering*, vol. 3, no. 3, pp. 263–276, 2018.

[31] C. Yang, B. He, C. Li, and J. Xu, "A feedback-based approach to utilizing embeddings for clinical decision support," *Data Science and Engineering*, vol. 2, no. 4, pp. 316–327, 2017.

[32] T. Wang, J. Ruan, Q. Yin, X. Dong, and Y. Wang, "An automated quality control pipeline for eQTL analysis with RNA-seq data," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 1780–1786.

[33] T. Wang, J. Peng, Q. Peng, Y. Wang, and J. Chen, "Fsm: Fast and scalable network motif discovery for exploring higher-order network organizations," *Methods*, vol. 173, pp. 83–93, 2020.

[34] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 701–710.

[35] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864.

[36] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 1067–1077.

[37] R. P. Huntley, T. Sawford, P. Mutowo-Meullenet, A. Shypitsyna, C. Bonilla, M. J. Martin, and C. O'Donovan, "The goa database: gene ontology annotation updates for 2015," *Nucleic acids research*, vol. 43, no. D1, pp. D1057–D1063, 2015.

[38] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in neural information processing systems*, 2007, pp. 153–160.

[39] J. Qiu, J. Tang, H. Ma, Y. Dong, K. Wang, and J. Tang, "Deepinf: Social influence prediction with deep learning," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2110–2119.

[40] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, "Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic acids research*, p. gkw943, 2016.