



Predicting Drug-Target Interactions by Node2vec Node Embedding in Molecular Associations Network

Zhan-Heng Chen^{1,2}, Zhu-Hong You^{1,2(✉)}, Zhen-Hao Guo^{1,2},
Hai-Cheng Yi^{1,2}, Gong-Xu Luo^{1,2}, and Yan-Bin Wang^{3,4}

¹ The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy
of Sciences, Urumqi 830011, China

zhuhongyou@ms.xjb.ac.cn, zhuhongyou@gmail.com

² University of Chinese Academy of Sciences, Beijing 100049, China

³ School of Cyber Science and Technology, Zhejiang University,
Hangzhou 310058, China

⁴ College of Information Science and Engineering,
Zaozhuang, Shandong 277100, China

Abstract. Accurate identification of drug-target interactions (DTIs) is essential for drug development. It not only helps the researchers to understand the mechanism of drug action, but also contributes to the innovative drug discovery and repositioning. However, due to the limitation of the high cost and long time, the traditional experimental methods are difficult to be widely applied for DTIs prediction. In this study, we propose an *in silico* method for predicting drug-target interactions by Node2vec node embedding in molecular associations network (MAN). Specifically, the MAN is constructed by integrating the associations among drug, protein, disease, lncRNA and miRNA. Then, the node2vec embedding method is employed to obtain a behavior feature vector of each node in the network. The traditional attribute feature vector comes from the drug molecular fingerprint and protein sequences. Finally, a random forest (RF) classifier is performed on these features to predict potential drug-target pairs. The experimental results show that the behavior feature could obtain 87.37% accuracy, which is obviously better than the traditional attribute feature. This work is not only more robust and reliable for predicting DTIs, but also provides an alternative way for other biomolecules associations prediction.

Keywords: Drug-target interactions · Multi-molecular network · Node2vec

1 Introduction

Most drug molecules usually perform their functions through the interaction with target proteins in human body. The discovery for drug targets has become the significant focus of innovative drugs research [1, 2]. Hence, prediction of drug-target interactions (DTIs) is one of the most important steps in genomic drug discovery pipeline and drug repurposing [3–5], the purpose is to discover putative new drugs and new uses of existing drugs. Nevertheless, due to the limitation of throughput and cost, the

traditional experimental methods are difficult to be widely applied for DTIs prediction. It is of great significance to develop effective calculation methods to predict the interaction between drugs and targets.

Recently, there are already a variety of calculation methods used to identify molecular associations [6–16], especially the interaction between drugs and targets [17–21]. Luo *et al.* developed a computational pipeline to detect novel DTIs from a constructed heterogeneous network, which achieves substantial performance improvement over other state-of-the-art methods [22]. Van Laarhoven *et al.* proposed a simple machine learning method that uses Gaussian interaction profile kernel and regularized least squares for predicting drug-target interactions [23]. Chen *et al.* proposed a drug-target interaction prediction method by random walk on a large-scale heterogeneous network, which combines drug-drug similarity network, protein-protein similarity network and known drug-target interaction network [24]. Ezzat *et al.* provide a comprehensive overview and empirical evaluation on the computational DTIs prediction, which helps understanding the advantages and disadvantages of the state-of-the-art methods [25]. Based on these methods, we proposed a multi-molecular network, also called molecular associations network (MAN) [26] to detect the interactions between drug candidate and related target proteins.

2 Materials and Methods

2.1 Datasets Construction

In the multi-molecular network, the high quality data mainly from nine open source database, which obtained nine known relationships (shown in Table 1) and five types of molecules (shown in Table 2). MAN contained topological relationships and distributions among all the molecules in the heterogeneous network. The drug molecular data and target protein sequences can be collected from DrugBank database and STRING database.

Table 1. Nine known relationships in the molecular associations network

Relationship	Database	Number
Drug-target	DrugBank [27]	11107
Drug-disease	CTD [28]	18416
Protein-disease	DisGeNET [29]	25087
lncRNA-target	lncRNA2Target [30]	690
lncRNA-disease	lncRNADisease [31]	1264
	lncRNASNP2 [32]	
miRNA-target	miRTarBase [33]	4944
miRNA-disease	HMDD [34]	16427
miRNA-lncRNA	lncRNASNP2 [32]	8374
Protein-protein	STRING [35]	19237
Total	N/A	105546

Table 2. The number of 5 types of biomolecules from the nine known relationships

Biomolecule	Number
Drug	1025
Target/Protein	1649
MiRNA	1023
LncRNA	769
Disease	2062
Total	6528

2.2 Molecular Association Network

From the collection of nine known relationships between five types of biomolecules annotated in many famous databases which mentioned above, we constructed a multi-molecular network, also called Molecular Associations Network (MAN) by linking arbitrary two association nodes. The complex molecular associations network is shown in Fig. 1. Based on the known associations, some biomolecules are suggested to interact with each other. In the network graph, the heterogeneous nodes correspond to five types of biomolecules (drug, protein, disease, miRNA and lncRNA), and edges correspond to associations among them. The construction of systematic and MAN network provides a new perspective for predicting interactions between drug and target.

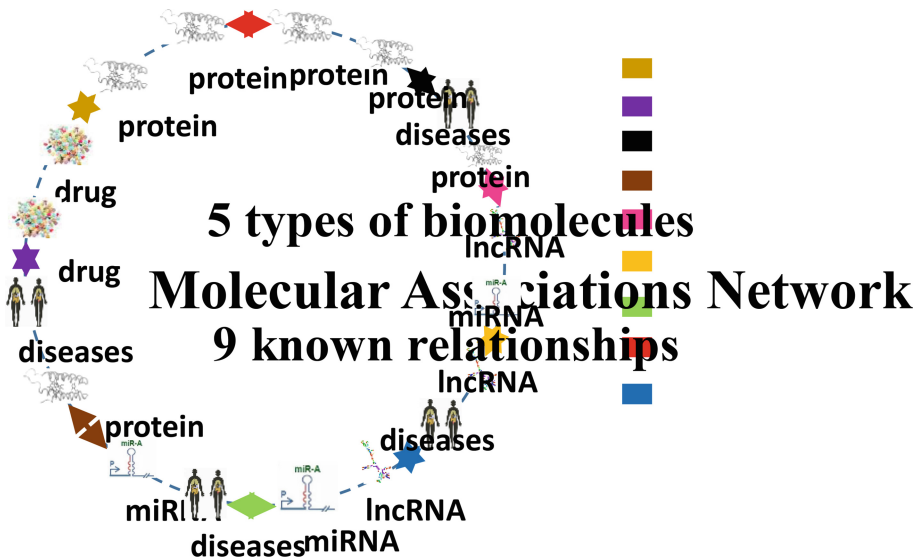


Fig. 1. Construction of multi-molecular network

2.3 Network Embedding: Node2Vec

Grover Aditya and Jure Leskovec proposed a graph embedding method, called Node2Vec, an algorithmic framework for learning continuous feature representations for nodes in graphs [36]. Different from the traditional graph embedding model, it can be seen as an extension of DeepWalk [37]. On the basis of DeepWalk, Node2Vec introduces two biased random walk methods: breadth-first search (BFS) [38] and depth-first search (DFS) [39], to characterize the structural equivalence and homophily of the network. Compared with random walk without any guidance, this method achieves the purpose of biased random walk by introducing Return Parameter and In-out Parameter, that is, the whole random walk process moves between BFS and DFS by setting different offsets. Take node V11 as an example, two search strategies as shown in Fig. 2.

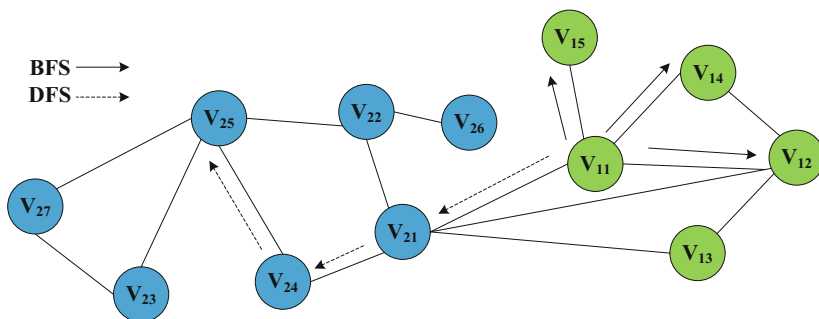


Fig. 2. Two search strategies from node V11 (step = 3)

2.3.1 Biased Random Walk

Suppose that a random walk started with node M and end with node N . Here, due to the use of two different search strategies (BFS and DFS), the selection of strategy will directly affect the result of random walk. The unnormalized transition probability algorithm is introduced to solve this problem. The transfer probability between the two nodes can be described as follow:

$$\pi_{NX} = \alpha_{pq}(M, X) \bullet w_{NX} \tag{1}$$

where, X represents the next position. w is the weight of the edge of the two nodes, which is based on the scenario. α is search bias.

$$\alpha_{pq}(M, X) = \begin{cases} \frac{1}{p} & \text{if } d_{MX} = 0 \\ 1 & \text{if } d_{MX} = 1 \\ \frac{1}{q} & \text{if } d_{MX} = 2 \end{cases} \tag{2}$$

where, d_{MX} is the shortest distance between M and N ; p is return parameter, which controls the probability of returning to the original node; q represents in-out parameter,

which controls the relationship between BFS and DFS. The setting of different p and q determine the priority of node sequence. When training the model, to find the best p and q by according to the needs of the scene and grid search.

- (1) When $d = 0$, it means to return to node M from N . At this time, the search bias is $1/p$, which can be understood as returning to the previous step with a probability of $1/p$;
- (2) When $d = 1$, X is the direct neighbor of M , which is equivalent to BFS, then the bias is 1;
- (3) When $d = 2$, X is the neighbor's neighbor of M , which is equivalent to DFS, then the bias is $1/q$.

2.3.2 Feature Learning

Now, suppose there is a graph $G = (V, E)$, where V is the set of nodes and E is the set of edges. The objective function for maximizing log-property could be described as follows:

$$\max_f \sum_{v \in V} \log Pr(N_s(v)|f(v)) \quad (3)$$

where, function $f: V \rightarrow R^d$ to represent the mapping from vertex to feature representation, where d is a pre-set hyper-parameter that represents the dimension of feature representation of each vertex. As a result, f is a matrix whose size is $|V| \times d$. $v \in V$, and $N_s(v) \subset V$ represents the neighbor vertex of vertex v under the sampling strategy s .

Assuming that the possibility of observing neighborhood nodes is independent of the feature representation of observing any other neighborhood nodes of a given vertex, so as to decompose this conditional probability, then

$$Pr(N_s(v)|f(v)) = \prod_{n_i \in N_s(v)} Pr(n_i|f(v)) \quad (4)$$

Assuming that the influence between two vertices in the feature space is symmetrical, then

$$Pr(n_i|f(v)) = \frac{\exp(f(n_i) \bullet f(v))}{\sum_{x \in V} \exp(f(x) \bullet f(v))} \quad (5)$$

The purpose of these two assumptions is to better handle the optimization problem. Based on the above assumptions, the objective function can be simplified as follow:

$$\max_f \sum_{v \in V} \left[-\log Z_v + \sum_{v_i \in N_s(v)} f(v_i) \bullet f(v) \right] \quad (6)$$

For each node,

$$Z_v = \sum_{x \in V} \exp(f(v) \bullet f(x)) \tag{7}$$

2.4 Random Forest

Random forest (RF) is an ensemble algorithm that includes a number of decision trees [40]. Focus on the problem of classification, each decision tree is treated as a classifier. Each sample is input into each tree for classification, and the category with the largest number of votes is designated as the final output.

In the process of feature importance assessment using random forest, it depends on the contribution of each feature to each tree in the RF. The contribution usually measured by *Gini* index or error rate of out-of-bag (OOB) data. Assuming that there is n features $f_1, f_2, f_3, \dots, f_n$, the Gini variable importance measures (VIM) of each feature f_i can be described as follow:

$$Gini_n = \sum_{m=1}^{|M|} \sum_{m' \neq m} p_{nm} p_{nm'} = 1 - \sum_{m=1}^{|M|} p_{nm}^2 \tag{8}$$

where, m represents m classes. p_{nm} is the proportion of class k in node n .

2.5 Performance Measurement Tools

In our study, in order to size up the effectiveness and steadiness of our constructed model, we counted the results of five parameters [41–43]: Accuracy (Acc), recall (sensitivity, hit rate, or true positive rate (TPR)), specificity(selectivity, or true negative rate (TNR)),precision (positive predictive value (PPV)) and Matthews’s Correlation Coefficient (MCC), respectively. These parameters can be represented as follows:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

$$TNR = \frac{TN}{FP + TN} \tag{11}$$

$$PPV = \frac{TP}{FP + TP} \tag{12}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{13}$$

3 Results and Discussion

In this paper, a deep learning method was derived from the idea of molecular association network, and proposed for predicting DTIs. Then, the node2vec embedding method was employed to obtain a behavior feature vector of each node from the multi-molecular network, which is constructed by integrating the associations among drug, protein, disease, lncRNA and miRNA. The good performance was obtained for this method based on behavior feature than traditional attribute feature on the collected datasets. Here, random forest classifier model was used to fulfill the experiment. During this experiment, we set the same parameters to compare the performances of the two different features on the model, the results as shown in Table 3 and 4. From the two tables, it is obvious that the accuracy on behavior features is 5% higher than the accuracy on attribute features under five-fold cross validation.

Table 3. Performance evaluation with RF on attribute features

5-folds	Acc (%)	TPR (%)	TNR (%)	PPV (%)	MCC
1	81.91	79.97	83.84	83.19	0.6386
2	83.12	82.04	84.2	83.85	0.6626
3	82.2	80.6	83.8	83.26	0.6443
4	81.77	80.02	83.53	82.93	0.6359
5	81.75	79.95	83.55	82.94	0.6354
Average	82.15 ± 0.57	80.52 ± 0.89	83.78 ± 0.27	83.23 ± 0.37	0.6434 ± 0.0113

Table 4. Performance evaluation with RF on behavior features

5-folds	Acc (%)	TPR (%)	TNR (%)	PPV (%)	MCC
1	87.38	82.49	92.26	91.42	0.7511
2	87.89	83.3	92.48	91.72	0.7611
3	87.4	83.44	91.36	90.62	0.7503
4	86.72	81.37	92.08	91.13	0.7387
5	87.47	82.47	92.47	91.64	0.7532
Average	87.37 ± 0.42	82.61 ± 0.83	92.13 ± 0.46	91.31 ± 0.45	0.7509 ± 0.0080

The ROC curve of random forest classifier based on attribute feature and behavior feature with 5-fold cross-validation that is shown in Fig. 3 and Fig. 4, respectively. It is obvious that the average of AUC is 0.8957 by using attribute information, the average of AUC is 0.9396 by using behavior information based on MAN network. So the behavior information of nodes play an important role in the DTIs predictions.

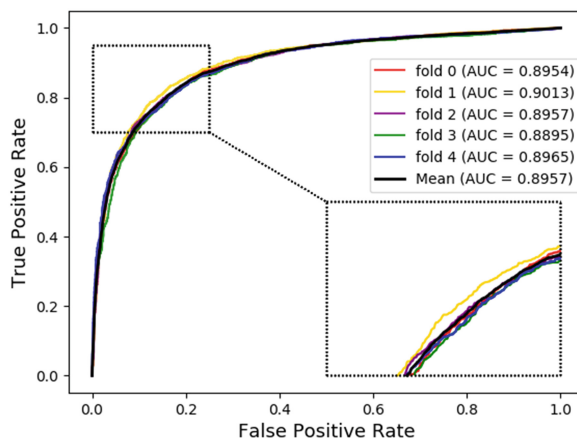


Fig. 3. The ROC curve of random forest on attribute feature

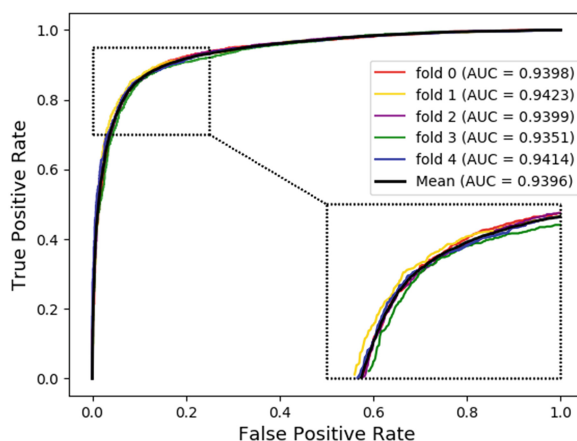


Fig. 4. The ROC curve of random forest on behavior feature

4 Conclusion

In this study, we developed a deep learning method to discover the potential interaction between drugs and target proteins on a large scale by investigating the relationship among five molecules (drug, protein, miRNA, lncRNA and disease). And we construct a novel scheme based on above five molecules and nine relationships between arbitrarily two molecules, which called MAN network. Focus on this network, each node can obtain a feature vector by using node behavior information (the relationship of each node with others could be described by node2vec graph embedding method). The traditional attribute feature vector comes from the drug molecular fin-gerprint and protein sequences on integrated dataset. Finally, a random forest (RF) classifier is

performed on these features to predict potential drug-target pairs. Experimental results indicated that the behavior feature could be performed better on random forest classifier. It is also demonstrated that the use of behavior information is very helpful for addressing the problem of drug molecules and target proteins. This work is a new attempt to predict DTIs and would have potential applications for drug discovery and repositioning.

Acknowledgments. This work is supported by the Xinjiang Natural Science Foundation under Grant 2017D01A78. This work is also supported in part by the National Natural Science Foundation of China, under Grants 61902342.

Author Contributions. ZH Chen and ZH You conceived the algorithm, carried out analyses, prepared the data sets, carried out experiments, and wrote the manuscript; ZH Guo and HC Yi designed, performed; GX Luo and YB W analyzed experiments and checked the manuscript; All authors read and approved the final manuscript.

Conflicts of Interest. The authors declare no conflict of interest.

References

1. Rask-Andersen, M., Almén, M.S., Schiöth, H.B.: Trends in the exploitation of novel drug targets. *Nat. Rev. Drug Discov.* **10**, 579–590 (2011)
2. Li, Y.H., et al.: Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. *Brief. Bioinform.* **21**, 649–662 (2020)
3. Liu, Y., Wu, M., Miao, C., Zhao, P., Li, X.-L.: Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput. Biol.* **12**, e1004760 (2016)
4. Tang, J., et al.: Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions. *Cell Chem. Biol.* **25**, 224–229 (2018). e222
5. Wang, L., You, Z.-H., Li, L.-P., Yan, X., Zhang, W.: Incorporating chemical sub-structures and protein evolutionary information for inferring drug-target interactions. *Sci. Rep.* **10**, 1–11 (2020)
6. Ji, B.-Y., You, Z.-H., Cheng, L., Zhou, J.-R., Alghazzawi, D., Li, L.-P.: Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model. *Sci. Rep.* **10**, 1–12 (2020)
7. You, Z.-H., et al.: PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput. Biol.* **13**, e1005455 (2017)
8. Chen, X., Yan, C.C., Zhang, X., You, Z.-H.: Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* **18**, 558–576 (2017)
9. Wang, M.-N., You, Z.-H., Wang, L., Li, L.-P., Zheng, K.: LDGRNMF: LncRNA-disease associations prediction based on graph regularized non-negative matrix factorization. *Neurocomputing* (2020)
10. Wang, M.-N., You, Z.-H., Li, L.-P., Wong, L., Chen, Z.-H., Gan, C.-Z.: GNMFLMI: graph regularized nonnegative matrix factorization for predicting LncRNA-MiRNA interactions. *IEEE Access* **8**, 37578–37588 (2020)
11. Wong, L., Huang, Y.A., You, Z.H., Chen, Z.H., Cao, M.Y.: LNRLMI: linear neighbour representation for predicting lncRNA-miRNA interactions. *J. Cell Mol. Med.* **24**, 79–87 (2020)

12. Guo, Z.-H., You, Z.-H., Wang, Y.-B., Huang, D.-S., Yi, H.-C., Chen, Z.-H.: Bioentity2vec: attribute-and behavior-driven representation for predicting multi-type relationships between bioentities. *GigaScience* **9**, g1aa032 (2020)
13. Guo, Z.-H., et al.: MeSHHeading2vec: a new method for representing MeSH headings as vectors based on graph embedding algorithm. *Brief. Bioinf.* (2020)
14. You, Z.-H., Zhou, M., Luo, X., Li, S.: Highly efficient framework for predicting interactions between proteins. *IEEE Trans. Cybern.* **47**, 731–743 (2016)
15. Wang, Y., You, Z.-H., Yang, S., Li, X., Jiang, T.-H., Zhou, X.: A high efficient biological language model for predicting protein–protein interactions. *Cells* **8**, 122 (2019)
16. Chen, Z.-H., You, Z.-H., Li, L.-P., Wang, Y.-B., Wong, L., Yi, H.-C.: Prediction of self-interacting proteins from protein sequence information based on random projection model and fast Fourier transform. *Int. J. Mol. Sci.* **20**, 930 (2019)
17. Chen, Z.-H., You, Z.-H., Guo, Z.-H., Yi, H.-C., Luo, G.-X., Wang, Y.-B.: Prediction of drug-target interactions from multi-molecular network based on deep walk embedding model. *Front. Bioeng. Biotechnol.* **8**, 338 (2020)
18. Wang, Y.-B., You, Z.-H., Yang, S., Yi, H.-C., Chen, Z.-H., Zheng, K.: A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. *BMC Med. Inform. Decis. Mak.* **20**, 1–9 (2020)
19. Wang, L., et al.: Identification of potential drug–targets by combining evolutionary information extracted from frequency profiles and molecular topological structures. *Chem. Biol. Drug Design* (2019)
20. Huang, Y.-A., You, Z.-H., Chen, X.: A systematic prediction of drug-target interactions using molecular fingerprints and protein sequences. *Curr. Protein Pept. Sci.* **19**, 468–478 (2018)
21. Li, Z., et al.: In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences. *Sci. Rep.* **7**, 1–13 (2017)
22. Luo, Y., et al.: A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* **8**, 1–13 (2017)
23. Van Laarhoven, T., Nabuurs, S.B., Marchiori, E.: Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* **27**, 3036–3043 (2011)
24. Chen, X., Liu, M.-X., Yan, G.-Y.: Drug–target interaction prediction by random walk on the heterogeneous network. *Mol. BioSyst.* **8**, 1970–1978 (2012)
25. Ezzat, A., Wu, M., Li, X.-L., Kwok, C.-K.: Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Brief. Bioinform.* **20**, 1337–1357 (2019)
26. Guo, Z.-H., Yi, H.-C., You, Z.-H.: Construction and comprehensive analysis of a molecular association network via lncRNA–miRNA–disease–drug–protein graph. *Cells* **8**, 866 (2019)
27. Wishart, D.S., et al.: DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018)
28. Davis, A.P., et al.: The comparative toxicogenomics database: update 2019. *Nucleic Acids Res.* **47**, D948–D954 (2019)
29. Piñero, J., et al.: DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* gkw943 (2016)
30. Cheng, L., et al.: LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* **47**, D140–D144 (2019)
31. Chen, G., et al.: LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* **41**, D983–D986 (2012)

32. Miao, Y.-R., Liu, W., Zhang, Q., Guo, A.-Y.: lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. *Nucleic Acids Res.* **46**, D276–D280 (2018)
33. Chou, C.-H., et al.: miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* **46**, D296–D302 (2018)
34. Huang, Z., et al.: HMDD v3.0: a database for experimentally supported human microRNA–disease associations. *Nucleic acids research* **47**, 1013–1017 (2019)
35. Szklarczyk, D., et al.: The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* gkw937 (2016)
36. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864 (2016)
37. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–710 (2014)
38. Kurant, M., Markopoulou, A., Thiran, P.: On the bias of BFS (breadth first search). In: *2010 22nd International Teletraffic Congress (ITC 22)*, pp. 1–8. IEEE (2010)
39. Cui, L., et al.: A novel artificial bee colony algorithm with depth-first search framework and elite-guided search equation. *Inf. Sci.* **367**, 1012–1044 (2016)
40. Liaw, A., Wiener, M.: Classification and regression by randomForest. *R News* **2**, 18–22 (2002)
41. Wong, L., You, Z.-H., Guo, Z.-H., Yi, H.-C., Chen, Z.-H., Cao, M.-Y.: MIPDH: a novel computational model for predicting microRNA–mRNA interactions by DeepWalk on a heterogeneous network. *ACS Omega* **5**, 17022–17032 (2020)
42. Chen, Z.-H., Li, L.-P., He, Z., Zhou, J.-R., Li, Y., Wong, L.: An improved deep forest model for predicting self-interacting proteins from protein sequence using wavelet transformation. *Front. Genet.* **10**, 90 (2019)
43. Wang, Y., You, Z., Li, L., Chen, Z.: A survey of current trends in computational predictions of protein-protein interactions. *Front. Comput. Sci.* **14**, 144901 (2020)